UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Reconhecimento de voz através de unidades menores do que a palavra, utilizando *Wavelet Packet* e SVM, em uma nova Estrutura Hierárquica de Decisão

Adriano de Andrade Bresolin

Orientador: Prof. Dr. ADRIÃO DUARTE DÓRIA NETO

Co-orientador: Prof. Dr. PABLO JAVIER ALSINA

Tese de Doutorado, apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da UFRN (área de concentração: Engenharia de Computação), como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica.

Natal, 02 de Dezembro de 2008.

Divisão de Serviços Técnicos

Catalogação da Publicação na Fonte. UFRN / Biblioteca Central Zila Mamede

Bresolin, Adriano de Andrade.
Reconhecimento de voz através de unidades menores do que a palavra, utilizando Wavelet Packet e SVM, em uma nova estrutura hierárquica de decisão / Adriano de Andrade Bresolin. – Natal, RN, 2008.
150 f. : il.

Orientador: Adrião Duarte Dória Neto.
Co-orientador: Pablo Javier Alsina.

Tese (Doutorado) – Universidade Federal do Rio Grande do Norte. Centro de Tecnologia. Programa de Pós-Graduação em Engenharia Elétrica.

1. Transformada de Wavelet – Tese. 2. Processamento de voz – Tese. 3. Wavelet packet – Tese. 4. Máquina de vetor de suporte – Tese. 5. Reconhecimento de voz – Tese. I. Dória Neto, Adrião Duarte. II. Alsina Pablo Javier. III. Título.

RN/UF/BCZM CDU 517.444(043.2)

Reconhecimento de voz através unidades menores do que a palavra, utilizando *Wavelet Packet* e SVM, em uma nova Estrutura Hierárquica de Decisão

Adriano de Andrade Bresolin

Tese de Doutorado aprovada em 02 de Dezembro de 2008 pela banca examinadora composta pelos seguintes membros:

_____

Prof. Dr. Adrião Duarte Dória Neto (Orientador) – DCA/ UFRN

_____

Prof. Dr. Pablo Javier Alsina (Co-orientador) – DCA/ UFRN

_____

Prof. Dr. Jorge Dantas de Melo – DCA/ UFRN (Examinador Interno)

_____

Prof. Dr. José Manoel de Seixas – COPPE/UFRJ (Examinador Externo)

_____

Prof. Dra. Joseana Macêdo Fechine – DSC /UFCG (Examinadora Externa)

Agradecimentos

"Ninguém alcança um objetivo sozinho", por isso:

Agradeço primeiramente a Deus pela vida.

Em segundo lugar, agradeço a minha família pelo apoio e incentivo à busca do conhecimento, em especial a minha esposa Elisabeti que sempre foi companheira, incentivadora e compreensiva pelos momentos que estive ausente em prol deste trabalho.

Agradeço também e com igual importância, ao professor e amigo Dr. Adrião Duarte Dória Neto (UFRN) que acreditou neste trabalho como orientador e sempre esteve presente incentivando, colaborando e cobrando em todas as etapas do desenvolvimento deste trabalho. Agradeço também ao co-orientador e professor Dr. Pablo Javier Alsina (UFRN) pelo incentivo, colaboração e apoio dispensados a este trabalho. Agradeço a todos os professores e funcionários do Departamento de Computação e Automação - DCA, que ministraram aulas ou colaboraram de alguma forma para a conclusão deste trabalho. Agradeço também prof. Dr. Diamantino Rui da Silva Freitas (FEUP-Porto) pela acolhida, dedicação e incentivo dispensados no decorrer do meu estágio de doutoramento no exterior realizado em Portugal na Universidade do Porto durante o ano de 2007.

Agradeço aos demais professores que participaram da banca de defesa: Prof. Dr. Jorge Dantas de Melo (UFRN), prof. Dr. José Manuel Seixas (UFRJ) e prof. Dra. Joseana Macêdo Fechine (UFCG) os quais colaboraram com correções, críticas e sugestões que representaram um aumento significativo na qualidade final deste trabalho.

Por fim e também com igual importância, agradeço aos amigos e colegas de curso e laboratório que estiveram juntos nesta caminhada. Em especial, agradeço de coração aos amigos Francisco de Chagas Lima Jr., Heliana Bezerra Soares, Rafael Marrocos Magalhães e Antonio de Pádua de Miranda Henriques pelo companheirismo, ajuda e amizade que demonstraram. Amizade que se iniciou, desenvolveu e permanecerá após a conclusão deste trabalho.

A Todos muito obrigado.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

A/D	Conversor Analógico Digital.
AMR	Análise de Múltipla Resolução (multiresolução).
ANN	*Artificial Neural Network*.
CMU	*Carnegie Mellon University.*
$coif_N$	Família de *Wavelets* Coiflet.
CWT	*Continuous Wavelet Transform.*
DARPA	*Defense Advance Research Projetcs Agency.*
dB	Decibéis.
db_N	Família de *Wavelets* Daubechies.
DCA	Departamento de Computação e Automação (UFRN).
DCT	*Discrete Cosine Transform.*
DFT	*Discrete Fourier Transform.*
DSP	*Digital Signal Processing* ou *Digital Signal Processor* (Equipamento).
DTFT	*Discrete-time Fourier transform.*
DWT	*Discrete Wavelet Transform*.
EUA	Estados Unidos da América.
FEUP	Faculdade de Engenharia da Universidade do Porto
FFT	*Fast Fourier Transform.*
FPGA	*Field-Programmable Gate Array.*
HMM	*Hidden Markov Models*.
IBM	*International Business Machines (Company).*
ICA	*Independent Component Analysis.*
IPA	*International Phonetic Alphabet.*
JPEG	*Joint Photographic Experts Group.*
LPC	*Linear Predictive Coding.*
ME	Mistura de Especialistas.
MFCC	*Mel-Frequency Cepstral Coefficient.*
MHE	Mistura Hierárquica de Especialistas.
MIT	*Massachusetts Institute of Technology.*
MLP	*Multi-Layer Perceptron.*
NBR 6023	Norma técnica Brasileira: referências bibliográficas.
NEC	*Nippon Electric Company.*
RBF	Função da Base Radial.
RCA	*Radio Corporation of America (Company).*
SNR	*Signal Noise Relation.*
SOM	*Self-Organizing Map.*
SRV	Sistemas de Reconhecimento de Voz.
STFT	*Short Time Fourier Transform.*
SVM	*Support Vector Machine.*
sym_N	Família de *Wavelets* Symlets.
TCZ	Taxa de Cruzamento por Zero.
TIMIT	*Texas Instruments / Massachusetts Institute of Technology.*
UFRN	Universidade Federal do Rio Grande do Norte.
VC	Dimensão VC (*Vapnik-Chervonenkis*).
WPT	*Wavelet Packet Transform.*
WT	*Wavelet Transform.*

Resumo

O reconhecimento automático da voz por máquinas inteligentes tem sido a meta de muitos pesquisadores nas últimas cinco décadas. Neste período, inúmeros avanços foram alcançados, como por exemplo no campo de reconhecimento de palavras isoladas (comandos), o qual atualmente apresenta taxas de reconhecimento muito altas. No entanto, ainda se está longe de desenvolver um sistema que possa ter um desempenho parecido com o ser humano, ou seja, reconhecimento automático de voz em modo contínuo.

Um dos grandes desafios das pesquisas de reconhecimento de voz contínuo é a grande quantidade de padrões existentes, pois as linguagens modernas tais como: Inglês, Francês, Espanhol e Português possuem aproximadamente 500.000 palavras ou padrões a serem identificados.

A proposta deste trabalho é utilizar unidades menores do que a palavra tais como: fonemas, difones e sílabas como unidades base para o reconhecimento da voz, visando o reconhecimento quaisquer palavras sem necessariamente utilizá-las. O objetivo principal deste trabalho é reduzir a restrição imposta pela quantidade excessiva de padrões existentes, ou seja, a quantidade excessiva de palavras. Com o objetivo de validar esta proposta, o sistema foi desenvolvido e testado para o reconhecimento de palavras isoladas no modo dependente do locutor.

O sistema apresentado neste trabalho foi desenvolvido com uma lógica de reconhecimento hierárquica baseada nas características de produção dos fonemas da língua Portuguesa do Brasil. Estas decisões são feitas através da utilização de redes neurais do tipo Máquinas de Vetor de Suporte agrupadas na forma de Máquinas de Cômite.

Os principais descritores do sinal de voz utilizados, foram obtidos através da Transformada *Wavelet Packet*. Os descritores MFCC (*Mel-Frequency Cepstral Coefficient)* também são utilizados neste trabalho.

Pode-se concluir que o método proposto apresentou bons resultados nas etapas de reconhecimento de vogais, consoantes (sílabas) e palavras se comparado com outros métodos existentes na literatura.

Palavras-chave: Reconhecimento de Voz, *Wavelet Packet* e Máquinas de Vetor de Suporte.

Abstract

The automatic speech recognition by machine has been the target of researchers in the past five decades. In this period have been numerous advances, such as in the field of recognition of isolated words (commands), which has very high rates of recognition, currently. However, we are still far from developing a system that could have a performance similar to the human being (automatic continuous speech recognition).

One of the great challenges of searches for continuous speech recognition is the large amount of pattern. The modern languages such as English, French, Spanish and Portuguese have approximately 500,000 words or patterns to be identified.

The purpose of this study is to use smaller units than the word such as phonemes, syllables and difones units as the basis for the speech recognition, aiming to recognize any words without necessarily using them. The main goal is to reduce the restriction imposed by the excessive amount of patterns. In order to validate this proposal, the system was tested in the isolated word recognition in dependent-case.

The phonemes characteristics of the Brazil's Portuguese language were used to developed the hierarchy decision system. These decisions are made through the use of neural networks SVM (Support Vector Machines).

The main speech features used were obtained from the Wavelet Packet Transform.The descriptors MFCC (Mel-Frequency Cepstral Coefficient) are also used in this work.

It was concluded that the method proposed in this work, showed good results in the steps of recognition of vowels, consonants (syllables) and words when compared with other existing methods in literature.

Keywords: *Speech Recognition, Wavelet Packet and Support Vector Machine.*

Capítulo I

1. Introdução: Processamento da Voz

Basicamente, os sistemas de processamento de voz estão divididos em três sub-campos: Codificação da fala (*Speech Coding*), Síntese da fala (*Speech Synthesis)* e Reconhecimento automático da fala (*Automatic Speech Recognition*).

Sistemas de codificação da fala englobam os processos nos quais a finalidade é obter uma representação compacta do sinal de voz. As técnicas de codificação do sinal de voz são usadas tanto para a transmissão quanto para o armazenamento compacto de sinais de voz. Uma das principais aplicações da codificação da fala é a transmissão do sinal de voz de forma eficiente.

A linha de pesquisa referente ao campo da síntese da fala se preocupa em gerar sons parecidos com a voz humana a partir do texto escrito, ou seja, conversão de texto em voz.

Os sistemas de reconhecimento automático da fala ou sistemas de reconhecimento de voz (SRV) têm seu enfoque voltado ao reconhecimento da voz do ser humano por máquinas inteligentes. Este trabalho tem seu foco principal voltado a esta área do processamento de voz.

Na Figura 1.1 é apresentado o esquema hierárquico dos sistemas de processamento de voz. Pode-se subdividir o campo do reconhecimento automático da fala em três áreas distintas: Reconhecimento do Locutor (pessoa que fala), Identificação da Linguagem (linguagem na qual se pronúncia) e Reconhecimento de Palavras (fala).

Como descrito anteriormente, o foco deste trabalho restringe-se ao reconhecimento de palavras. No entanto, alguns trabalhos referentes ao reconhecimento do locutor serão citados e servirão como fonte de dados para as análises desenvolvidas durante o trabalho.

O reconhecimento de palavras (fala) pode ser feito de dois modos diferentes: Modo dependente e independente do locutor. No modo dependente do locutor o sistema tem como objetivo reconhecer as palavras faladas somente por um locutor (pessoa), ou seja, o sistema é treinado e reconhece somente as palavras de uma pessoa específica. Já para o modo independente do locutor, o sistema é treinado com diversos locutores. A meta neste

caso é reconhecer as palavras pronunciadas por qualquer outra pessoa (locutor), diferente daquelas utilizadas no treinamento.

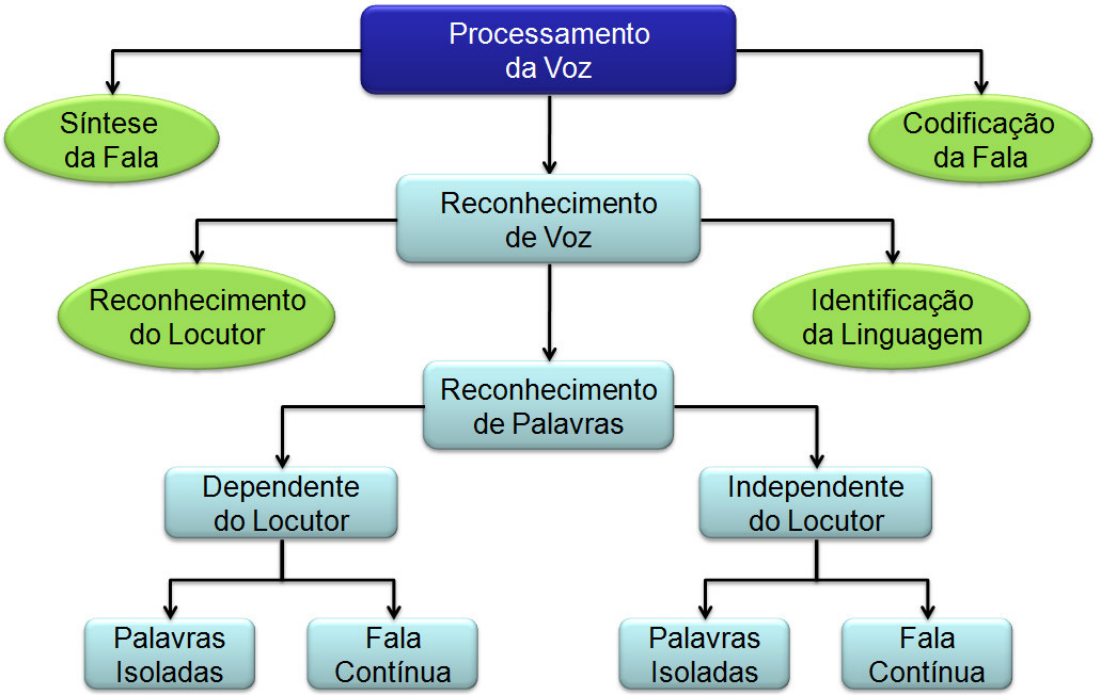

Figura 1.1: Esquema hierárquico dos sistemas de processamento de voz.

O objetivo final dos sistemas de reconhecimento de voz é a compreensão pela "máquina" do que foi pronunciado pelo locutor (pessoa que fala), ou seja, reconhecer o sinal de voz (pronúncia) transformando-o em texto ou em comandos para o acionamento de um ou diversos equipamentos eletro-mecânicos. Basicamente, os SRV (sistemas de reconhecimento de voz) visam o reconhecimento da palavra, seja ela isolada (comando) ou em um contexto de frase (fala contínua). No entanto, para reconhecer a palavra, muitos sistemas utilizam unidades menores, tais como: sílabas, fonemas, difones, trifones etc.

1.1 O Reconhecimento de Voz: Um breve histórico

O reconhecimento automático da voz por máquinas tem sido a meta de muitos pesquisadores por quase cinco décadas e tem inspirado maravilhas da ficção científica tais como: o computador HAL de Stanley Kubrick no famoso filme "*2001-Uma Odisséia no Espaço*" e o robô R2D2 de George Lucas no clássico filme "*Guerra nas Estrelas*". Entretanto, apesar do glamour das máquinas inteligentes que podem reconhecer as palavras faladas e

compreender seu significado, e apesar dos enormes esforços e gastos em pesquisas tentando criar tal máquina, ainda se está longe de alcançar a desejada meta de uma máquina poder entender um discurso falado de uma pessoa qualquer dentro de um universo de vários falantes, diferentes idiomas, diferente formas dos ambientes e diferentes níveis de ruídos [RABINER, 1993].

A realidade portanto é outra, ou seja, para o reconhecimento de uma simples palavra ou de uma frase falada continuamente (ritmo normal de conversação) é necessário um grande esforço computacional, onde diversas áreas do conhecimento são empregadas no treinamento e reconhecimento final. Até o presente momento, o uso da estrutura de sintaxe e a análise semântica no reconhecimento de voz são ainda questões em aberto [LEVISON, 2005]. Portanto, este trabalho tem o objetivo de ser uma contribuição ao universo de pesquisas desenvolvidas nos últimos 50 anos.

Basicamente, o desenvolvimento das pesquisas na área do reconhecimento de voz pode ser divido em décadas. Iniciando-se pelos anos 50 do século passado em que vários pesquisadores tentaram explorar as idéias fundamentais de acústica e fonética. Em 1952, nos laboratórios da Bell, Davis e Balashek [DAVIS, 1952] construíram um sistema para o reconhecimento de dígitos isolados (qualquer número de 0 a 9) utilizando apenas um único locutor. Em um esforço independente nos laboratórios da RCA, Olson e Belar [OLSON, 1956] tentaram reconhecer 10 sílabas diferentes faladas por uma pessoa. No final da década de 50, Fry e Denis [FRY, 1959] pesquisadores da *University College of England* construíram um reconhecedor de fonemas para reconhecer quatro vogais e nove consoantes. Eles usaram um analisador de espectro e uma combinação de padrões para fazer a decisão do reconhecimento. Outro esforço notável neste período foi o reconhecedor de vogais de Forgie & Forgie [FORGIE, 1959] construído no *MIT Lincoln Laboratories* em 1959, em que um analisador formado por um banco de filtros foi usado para prover a informação espectral.

Na década de 60 os laboratórios japoneses entraram na arena do reconhecimento de voz. O primeiro sistema japonês foi descrito por Suzuki e Nakata [SUZUKI, 1961] do Laboratório de pesquisas de Tóquio, o sistema era um reconhecedor de vogais. Outro esforço japonês na construção de um sistema de reconhecimento de voz foi o trabalho de Sakai e Doshita [SAKAI, 1962] da *Kyoto University*. Eles construíram um reconhecedor de fonemas que usava a análise de passagem por zero para fazer o reconhecimento de voz. O

terceiro esforço japonês foi o reconhecedor de dígitos de Nagata, Kato, Chiba e seus colegas de trabalho nos laboratórios da NEC (*Nippon Electric Company*) em 1963. Este trabalho foi talvez a mais notável tentativa no reconhecimento de voz da NEC, o qual iníciou um longo e altamente produtivo programa de pesquisas [NAGATA, 1963].

Ainda na década de 60, três projetos chave foram iniciados, os quais tiveram uma grande implicação nas pesquisas desenvolvidas no reconhecimento de voz nos 20 anos seguintes. O primeiro foi a pesquisa de Martin e seus colegas [MARTIN, 1964] dos laboratórios RCA (*Radio Corporation of America*) no final dos anos 60. Martin desenvolveu um conjunto de métodos elementares de normalização no tempo baseado na habilidade de detectar o início e o fim da fala. Quase ao mesmo tempo na União Soviética, Vintsyuk propôs o uso de métodos de programação dinâmica [VINTSYUK, 1968]. Outro êxito no final dos anos 60 foi a pesquisa pioneira de Reddy no campo do reconhecimento da fala contínua por meio de fonemas dinâmicos [REDDY, 1966]. As pesquisas de Reddy produziram o longo sucesso do programa de pesquisas do reconhecimento de voz da CMU (*Carnegie Mellon University*).

Na década de 70, as pesquisas de reconhecimento de voz tiveram um significativo avanço. Primeiro, a área de reconhecimento de palavras isoladas ou expressões discretas tornou-se viável com tecnologia baseada nos fundamentos estudados por Velicho e Zagoruyko na Rússia [VELICHKO, 1970], Sakoe e Chiba no Japão [SAKOE, 1978] e Itakura nos Estados Unidos [ITAKURA, 1975]. Os russos estudaram o uso de padrões de reconhecimento. Os japoneses pesquisaram como métodos de programação dinâmica podiam ser usados com sucesso. A pesquisa de Itakura apresentou idéias como o LPC (*linear predictive coding*). Finalmente, os laboratórios AT&T e Bell (EUA) iniciaram uma série de pesquisas visando fazer um sistema de reconhecimento de voz que pudesse entender uma pessoa falando.

Ao contrário dos anos 70, em que a meta era reconhecer uma palavra, nos anos 80 a meta era reconhecer a fala fluente de frases (fala contínua). Pode-se incluir nestas pesquisas uma realizada por Sakoe na NEC. Pesquisa esta iniciada ainda no final dos anos 70 [SAKOE, 1979]. As pesquisas nos anos 80 foram caracterizadas pelas aproximações estatísticas, especialmente o modelo de HMM (*Hidden Markov Model*). Desenvolvido nos anos 60, o HMM [BAUM, 1968] se tornou a ferramenta (de classificação) padrão nos sistemas de reconhecimento de voz nas décadas seguintes. Outra nova tecnologia introduzida nos anos 80 foi a aplicação de redes neurais em problemas de reconhecimento de voz. A era moderna

das redes neurais se iníciou em 1943 com o trabalho de McCulloch e Pitts [McCulloch, 1943]. Um grande avanço na área de reconhecimento de padrões foi alcançado em 1958 com o trabalho desenvolvido por Rosemblatt intitulado de *O perceptron* [Rosemblatt, 1958]. Nos anos 60 parecia que as redes neurais poderiam realizar qualquer coisa, no entanto o livro de Minsky e Papert [Minsky, 1969] demonstrou que existiam limites fundamentais para aquilo que os *perceptrons* de camada única podiam calcular. Somente em 1986, com o desenvolvimento do algoritmo da retro propagação (*back-propagation)* de Rumelhart, Hinton e Williams [Rumelhart, 1986] é que as redes neurais superaram a barreira imposta pelo trabalho de Minsky e Papert e se tornaram uma ferramenta utilizada nas mais variadas áreas de conhecimento, inclusive no reconhecimento de padrões de voz com muito sucesso.

Nos anos 90, as pesquisas continuaram com a busca de um sistema de reconhecimento de voz contínuo. Um exemplo destes sistemas é o DARPA (*Defense Advance Research Projetcs Agency*), o qual visava reconhecer continuadamente e sem erros palavras dentro de um arquivo de 1000 palavras. Também na década de 90, surgiu uma nova ferramenta desenvolvida por Vapnik e co-autores [Vapnik, 1992] chamada Máquinas de Vetor de Suporte (*Support Vector Machine* - SVM). O SVM é uma classe de redes de aprendizagem supervisionada poderosa do ponto de vista computacional, a qual é utilizada principalmente em reconhecimento de padrões e regressão. Uma característica das máquinas de vetor de suporte é a dimensão VC (*Vapnik-Chervonenkis*), a qual fornece uma medida da capacidade de uma rede neural de aprender a partir de um conjunto de exemplos [Vapnik, 1971].

Até o presente, a maioria dos sistemas de reconhecimento de voz tem como base os modelos de Markov - HMM em conjunto com os descritores MFCC (*Mel-Frequency Cepstral Coefficient*). Algumas versões são agora disponibilizadas comercialmente para o uso em computadores pessoais. Entretanto, seus desempenhos não são de confiança e deixam muito a desejar [Levison, 2005]. As pesquisas sobre reconhecimento de voz percorreram um grande e árduo caminho até os dias de hoje e, apesar de ainda ser uma questão em aberto no que diz respeito ao reconhecimento contínuo da fala, muitos avanços e bons resultados têm sido obtidos no campo do reconhecimento de voz restrito a palavras isoladas (comandos) e pequenas frases. Portanto, na atualidade, as pesquisas estão direcionadas ao reconhecimento de frases, ou seja, no reconhecimento da fala contínua.

No entanto, inúmeras pesquisas ainda estão focadas no reconhecimento de palavras isoladas (comandos), de fonemas e sílabas. Isto se dá por questões de mercado (principalmente telecomunicações e informática) e também devido ao fato de que o reconhecimento de unidades menores que palavra, pode ser o caminho para a resolução do problema do reconhecimento da fala contínua.

1.2 O Sistema de Reconhecimento de Voz - SRV

Um sistema de reconhecimento de voz é interdisciplinar, ou seja, várias áreas de conhecimento e de pesquisa são empregadas para formar o sistema por completo. Apesar de existirem dezenas de modelos diferentes, o formato básico é praticamente o mesmo em todos os sistemas, ou seja, um sistema de reconhecimento de voz é composto por 5 partes essenciais, as quais são:

- *a) Aquisição do sinal de voz.*
- *b) Pré-processamento do sinal.*
- *c) Extração de características do sinal de voz (descritores).*
- *d) Treinamento de um sistema classificador.*
- *e) Reconhecimento.*

As diferenças entre os sistemas de reconhecimento vão desde o objetivo a ser conquistado, ou seja, o que se pretende reconhecer (frase contínua ou comandos), passando por diferentes métodos de pré-processamento e extração de características até diferentes métodos de treinamento e classificação.

As cinco etapas descritas acima são apresentadas no diagrama da Figura 1.2. Basicamente, o processo total possui duas fases: Treinamento e Reconhecimento. Na fase do treinamento (Figura 1.2: a*, b, c, d*), o classificador é treinado com padrões específicos (descritores) para reconhecer um determinado padrão de voz. Na fase de reconhecimento (Figura 1.2: *a, b, c, e*), a máquina "inteligente" (classificador previamente treinado) é utilizada para fazer o reconhecimento do padrão de voz.

A etapa de aquisição do sinal (Figura 1.2-*a*) é responsável pela digitalização do sinal de voz, ou seja, a transformação do sinal acústico produzido pelo ser humano em um sinal digital que possa ser entendido pelo computador ou por um DSP (*Digital Signal Processor*).

A etapa de pré-processamento do sinal (Figura 1.2-*b*) é composta por vários processos, tais como: filtragem, pré-ênfase, normalização, janelamento etc. O objetivo desta etapa é eliminar ruídos, descontinuidades e quaisquer efeitos que possam prejudicar o desempenho do sistema.

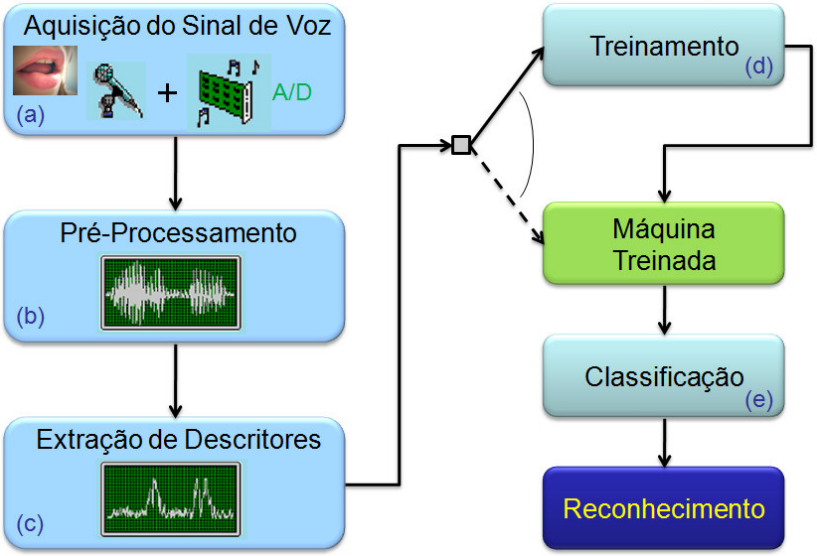

Figura 1.2: Diagrama dos sistemas de reconhecimento de voz.

A extração dos descritores do sinal de voz (Figura 1.2-*c*) é uma das partes mais importantes do sistema de reconhecimento, pois é neste estágio que se define quais as características que irão representar o sinal de voz tanto no treinamento quanto no reconhecimento.

O treinamento do classificador (Figura 1.2-*d*) é a etapa do sistema de reconhecimento de voz cujo objetivo é treinar uma "máquina inteligente" para que ela aprenda a reconhecer os descritores do sinal e conseqüentemente possa reconhecer o padrão de voz falado. Na etapa de classificação (Figura 1.2-*e*), utiliza-se a "máquina inteligente", previamente treinada, para fazer o reconhecimento através de uma decisão lógica dos resultados apresentados.

Para cada uma das etapas descritas acima, pode-se encontrar na literatura diferentes métodos que foram propostos e testados. No entanto, as etapas (*a*), (*b*) e (*c*) da Figura 1.2 são praticamente uniformizadas com pequenas diferenças entre algumas aplicações. As grandes diferenças encontram-se nas fases de extração dos descritores, treinamento e classificação.

1.3 O Estado da Arte dos SRV

Muitos trabalhos apresentam o estado da arte em um capítulo específico, descrevendo as citações, propostas e resultados de publicações com o objetivo de justificar a utilização de ferramentas e até a criação de novos métodos para resolução do problema. No entanto, devido à interdisciplinaridade natural do problema de reconhecimento de voz, neste trabalho optou-se pela apresentação do estado da arte no decorrer dos capítulos conforme o item ou ferramenta discutida. Devido a sua importância, alguns trabalhos serão citados mais de uma vez, pois apresentam contribuições em mais de uma área.

As duas ferramentas mais utilizadas nos sistemas de reconhecimento de voz atualmente são: MFCC (*Mel-Frequency Cepstral Coefficient*) como descritor do sinal de voz e HMM (*Hidden Markov Models*) como classificador. Um bom resumo sobre o método de cálculo dos descritores MFCC pode ser encontrado em [COMBRINCK, 1996]. Para um maior aprofundamento sobre os classificadores HMM, uma descrição completa pode ser encontrada em [RABINER, 1989]. No entanto, inúmeros trabalhos têm demonstrado que novas ferramentas para extração de descritores (em especial a *Wavelet* – Capítulo III) têm alcançado melhores resultados do que os coeficientes MFCC. Além disso, Russell e Bilmes declaram em seu editorial, que nos últimos anos reascendeu o interesse em classificadores que possam ir além da desempenho dos sistemas baseados nos modelos HMM [RUSSELL, 2003].

O trabalho de Richard Lippmann pode ser considerado como um levantamento do estado da arte dos sistemas de reconhecimento de voz até o ano de 1997 [LIPPMANN, 1997]. O objetivo do estudo foi determinar o quanto os avanços tecnológicos (até aquela data) tinham progredido na direção da meta principal dos sistemas de reconhecimento de voz (SRV), que é obter desempenho de reconhecimento próximo ou igual ao do ser humano. Em todas as análises apresentadas pelo autor, pode-se ver que naquele estágio (ano 1997) os sistemas de reconhecimento de voz ainda tinham um desempenho muito pobre em relação a capacidade humana de reconhecimento de voz. Lippmann considerou:

> *- Os resultados obtidos até o momento (1997) mostraram que os humanos têm um desempenho mais acurado que as máquinas quando usados modelos de baixo nível do tipo acústico-fonético. Estes resultados sugerem que as pesquisas devem estar voltadas*

para o desenvolvimento de máquinas que façam um melhor reconhecimento em baixo nível acústico-fonético, ou seja, fonemas, difones e trifones.

- Os humanos também possuem melhor desempenho de reconhecimento quanto o sinal contém ruídos. Assim, as pesquisas devem ser focadas para melhorar o desempenho do reconhecimento pelas máquinas na presença de ruído.

- A fala espontânea também derruba o desempenho dos sistemas de reconhecimento de voz. Isto ocorre devido ao fato que os sistemas são muito amarrados às restrições impostas pelo banco de dados treinado. Quando se muda o banco de dados, a máquina treinada com um banco de dados diferente tem desempenho muito ruim.

- A velocidade de adaptação a novos locutores (pessoas que falam) é muito rápida para os seres humanos. Em cerca de 2 a 4 segundos o ser humano consegue se adaptar a um novo locutor. Já as máquinas precisam treinar um amplo e demorado vocabulário para se adaptar a um novo locutor.

Por fim, Lippmann concluiu que ainda existe um problema que os sistemas de reconhecimento de voz devem resolver: os algoritmos não são capazes de reconhecer palavras novas sem ter que refazer o treinamento.

O resumo do estado da arte do reconhecimento de voz feito em 1997 por Lippmann, apesar de ter quase uma década, fez recomendações e observações que ainda valem para os dias de hoje. Pois, apesar dos muitos avanços e do surgimento de novas ferramentas, problemas como ruído e adaptação a novos locutores e novas palavras, ainda são problemas em aberto e, portanto, objeto de pesquisas nesta área.

1.4 Descrição do Trabalho e Contribuições

Em resumo, a proposta deste trabalho é utilizar unidades menores do que a palavra tais como: fonemas, difones, trifones e sílabas como unidades base para o reconhecimento da voz, utilizando a *Wavelet Packet* como descritor do sinal de voz e as Máquinas de Vetor de Suporte (SVM) como classificadores.

Limitou-se o foco deste trabalho ao reconhecimento de palavras isoladas (comandos) e ao modo dependente do locutor. O objetivo desta limitação é a validação desta nova proposta. Um estudo futuro poderá incluir a utilização desta nova proposta no reconhecimento contínuo da fala e no modo independente do locutor.

A principal contribuição deste trabalho está na elaboração de um novo Sistema de Decisão Hierárquica. Esta nova proposta tem como base as regras de classificação da fonética articulatória da língua portuguesa do Brasil. Basicamente, os fonemas são agrupados conforme suas classificações, ou seja, conforme a maneira de produção de cada fonema.

Nas fases de treinamento e classificação dos padrões, este trabalho apresenta outra contribuição através da utilização de fonemas, difones e sílabas como partes separadas (conjuntos), tendo como objetivo final o reconhecimento da palavra. Cada conjunto de reconhecimento pode ser considerado como uma máquina especialista, pois é treinado com propriedades exclusivas e com objetivo específico. Estas máquinas são constituídas basicamente por redes neurais do tipo Máquinas de Vetor de Suporte (SVM), as quais são agrupadas em uma estrutura similar a uma Máquina de Comitê [HAYKIN, 2001]. A utilização da transformada *Wavelet Packet* (WPT), em conjunto com os coeficientes MFCC, como descritores do sinal de voz também pode ser considerada como uma contribuição deste trabalho.

No Capítulo II são descritas as regras da fonética articulatória aplicadas a língua portuguesa. Além disso, neste capítulo são apresentadas as definições e a classificação dos fonemas, das sílabas e difones da língua Portuguesa do Brasil. Os modelos propostos para reconhecimento dos conjuntos de vogais e consoantes também são apresentados neste capítulo. Por fim, inúmeros trabalhos que justificam a utilização dos fonemas, difones e sílabas no reconhecimento de voz são apresentadas na seção sobre o estado da arte.

No Capítulo III é apresentada uma descrição detalhada da transformada *Wavelet* e *Wavelet Packet* (WPT), além dos descritores MFCC. Novamente, no final deste capítulo são apresentados inúmeros trabalhos que justificam a utilização da *Wavelet* como descritor do sinal voz, os quais serviram de base para o desenvolvimento deste trabalho.

No Capítulo IV é descrita a teoria necessária para o treinamento, classificação e decisão do sistema. Além das Máquinas de Vetor de Suporte (*Support Vector Machine*-SVM), neste capítulo é feita uma apresentação da teoria a respeito das Máquinas de Comitê e da teoria estatística da decisão.

No Capítulo V é apresentada em detalhes a proposta completa deste trabalho, o qual visa o reconhecimento de voz (palavra) através da utilização de unidades menores do que a

palavra. Neste capítulo são apresentadas ainda as ferramentas utilizadas na etapa de pré-processamento do sinal bem como o diagrama completo da nova Regra de Decisão Hierárquica.

No Capítulo VI são descritos as conclusões obtidas neste trabalho, bem como as contribuições e trabalhos futuros.

O Anexo I apresenta os experimentos com o reconhecimento de palavras isoladas, os quais foram importantes para a determinação das melhores *Wavelet* mãe para o reconhecimento de voz.

O Anexo II apresenta o diagrama de blocos completo do sistema desenvolvido. O Anexo III contém a lista das publicações em congressos e revistas, nacionais e internacionais, as quais foram obtidas durante a realização deste trabalho.

As referências citadas, tais como: publicações, livros e sites da internet, encontram-se no final deste trabalho e estão ordenadas por autor e em ordem alfabética. Este trabalho, segue as normas de apresentação de dissertações e teses do curso de pós-graduação em Engenharia Elétrica e Computação da UFRN (Universidade Federal do Rio Grande do Norte), bem como a norma NBR 6023.

Capítulo II

2. Fonética e Fonologia do Português Brasileiro

A fonética articulatória descreve a maneira natural com que os serem humanos articulam o trato vocal para a produção dos sons básicos (fonemas). Este arranjo articulatório natural e suas classificações foram utilizados como regra de construção e de decisão no sistema proposto por este trabalho.

Neste capítulo são descritos os conceitos da produção da fala na visão da Fonética e da Fonologia abrangendo os segmentos consonantais e vocálicos. Além disso, neste capítulo são descritos ainda as semivogais, encontros consonantais e a construção das sílabas e difones do Português Brasileiro. A seção final deste capítulo apresenta diversos trabalhos e publicações nos quais o foco foi o reconhecimento de unidades menores do que a palavra.

2.1 Fonética Articulatória

A Fonética é a ciência que apresenta os métodos para descrição, classificação e transcrição dos sons da fala. As principais áreas de estudo da fonética são:

- Fonética Articulatória: Compreende o estudo da produção da fala do ponto de vista fisiológico e articulatório;
- Fonética Auditiva: Compreende o estudo da percepção da fala;
- Fonética Acústica: Compreende o estudo das propriedades físicas dos sons da fala a partir da sua transmissão do falante ao ouvinte;
- Fonética Instrumental: Estuda as propriedades físicas da fala levando em consideração o apoio de instrumentos laboratoriais.

2.2 A Produção da Voz

O ser humano normal é capaz de emitir sons de qualquer língua. Para tanto ele faz uso de uma parte específica do corpo que é chamada de Aparelho Fonador. Existem três grupos

de órgãos do ser humano que são importantes na produção da fala, são eles: O Sistema Respiratório, o Sistema Fonatório e o Sistema Articulatório. A Figura 2.1 apresenta os três sistemas citados.

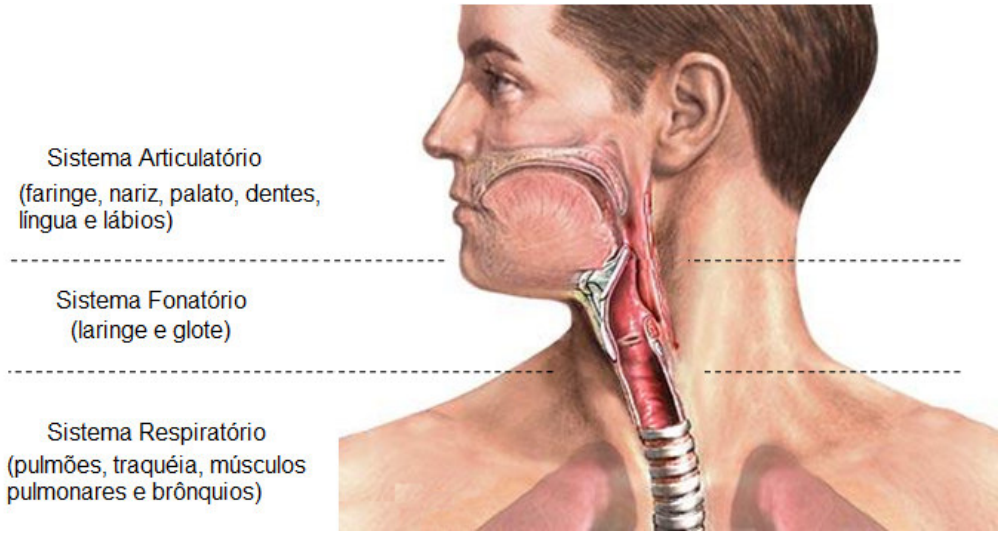

Figura 2.1: Os sistemas: Respiratório, Fonatório e Articulatório. Adaptação de [SILVA, 2003].

O sistema respiratório, composto pelos pulmões, músculos pulmonares, brônquios e traquéia, está localizado na parte inferior à glote. Sua função primária é a respiração.

O sistema fonatório é constituído pela laringe. Na laringe encontram-se as cordas vocais que são músculos estriados capazes de obstruir a passagem de ar. O espaço decorrente da não obstrução deste músculo é chamado de glote. A função primária da laringe é atuar como uma válvula que obstrui a entrada de comida nos pulmões. Isto é feito através do abaixamento da epiglote.

O sistema articulatório consiste da faringe, língua, nariz, dentes e lábios. As funções primárias deste sistema envolvem os atos de sugar, engolir, respirar, mastigar alimentos, além do paladar e olfato. A Figura 2.2 apresenta os pontos importantes do trato vocal, incluindo os articuladores passivos e ativos bem como as cavidades: oral, nasal, faringal e a glote ou cordas vocais.

Cada um dos articuladores (passivos e ativos) e as cavidades contribuem de uma forma ou de outra para a formação dos fonemas pelo ser humano.

Dependendo da língua nativa do locutor, alguns sons podem ou não ser produzidos. Por exemplo: um locutor nativo de língua inglesa aprende e, portanto, possui a capacidade de produzir cerca de doze vogais diferentes, enquanto que um nativo de língua espanhola

aprende somente a produzir cinco vogais. Isto explica a dificuldade de um nativo de língua espanhola ou latina de pronunciar muitas palavras da língua inglesa e vice-versa. No entanto, todo ser humano nasce com a capacidade de pronunciar quaisquer sons de qualquer língua.

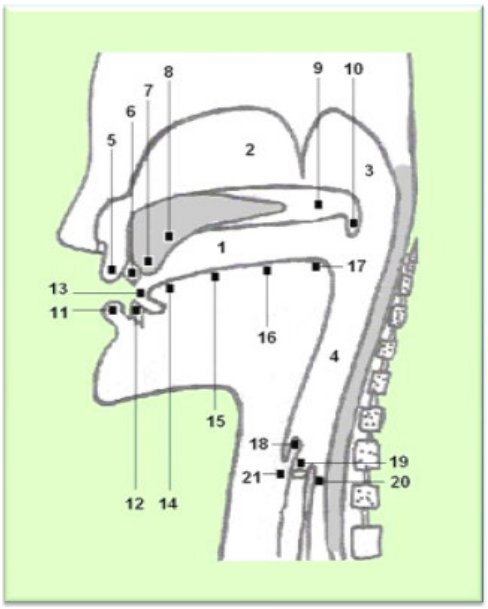

Figura 2.2: O trato vocal: cavidades e articuladores passivos e ativos. Adaptação de [Silva, 2003].

onde:

1- Cavidade Oral
2- Cavidade Nasal.
3- Cavidade Nasofaringal.
4- Cavidade Faringal.
5- Lábio Superior.
6- Dentes Superiores.
7- Alvéolos.
8- Palato duro.
9- Véu Palatino (palato mole).
10- Úvula.
11- Lábio Inferior
12- Dentes Inferiores.
13- Ápice da língua.
14- Lâmina da Língua.
15- Parte anterior da língua.
16- Parte média da língua.
17- Parte posterior da língua.
18- Epiglote.
19- Laringe.
20- Esôfago.
21- Glote.

Para dar ênfase a descrição fonética da língua portuguesa é preciso então estudar os símbolos fonéticos que a representam. Estes símbolos estão separados em duas categorias principais, chamadas de segmento consonantal (consoantes) e segmento vocálico (vogais).

2.3 Segmentos: Consonantal e Vocálico

O que difere os segmentos consonantais (consoantes) dos segmentos vocálicos (as vogais) é a maneira como o ar passa pelas cavidades supraglotais. Se o som é produzido com

algum tipo de obstrução da passagem de ar, podendo haver ou não fricção, então este som é consonantal. Por outro lado, se não houver obstrução a passagem de ar classifica-se este som como vocálico.

Alguns segmentos não são bem definidos, ou seja, não têm características bem definidas quanto à passagem do ar. Estes segmentos não são nem consoantes nem vogais, são semivogais ou *glide*. A Figura 2.3 apresenta um esquema da saída de ar na produção da voz pelo ser humano. Pode-se ver também a representação da vibração das cordas vocais (no caso de sons vocálicos ou sonoros).

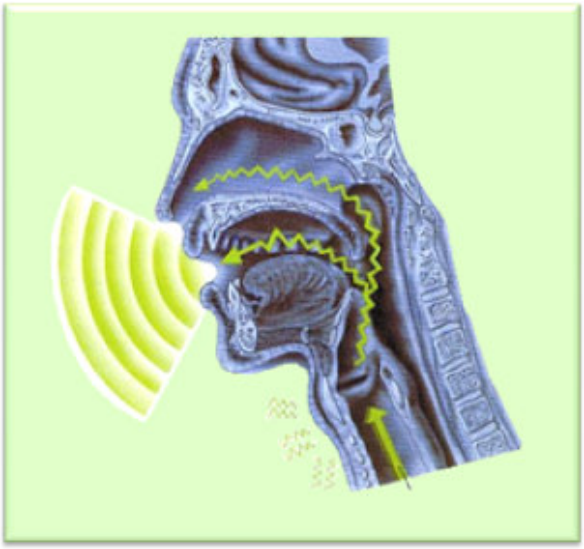

Figura 2.3: Saída de ar na produção da voz pelo ser humano. Adaptação de [SILVA, 2003].

2.3.1 Segmentos Consonantais

Para melhor entender a produção e conseqüente classificação dos segmentos consonantais é importante responder as seguintes questões:

- a) Qual o mecanismo da corrente de ar na produção do som?
- b) Há vibração das cordas vocais?
- c) O som é Oral ou Nasal?
- d) Quais os articuladores envolvidos na produção dos sons?
- e) Qual a maneira utilizada para obstruir a corrente de ar?

Questão (a): Qual o mecanismo da corrente de ar na produção do som?

R. Quando produzimos sons (fala) a corrente de ar é pulmonar e egressiva, ou seja, a corrente de ar sai dos pulmões.

Questão (b): Há vibração das cordas vocais?

R. Se houver vibração das cordas vocais (Figura 2.2 - 21) durante a produção do som, classifica-se o som como vozeado (Sonoro). Em contra partida, se não houver vibração das cordas vocais o som é classificado com não vozeado (Surdo).

No entanto, não existe um limite de separação entre um som sonoro e um surdo. Na verdade, as categorias vozeada e não vozeada podem ser interpretadas como limites de um contínuo que faz uma gradação de sons vozeados e sons não vozeados. Por exemplo: os sons dos fonemas consonantais /*b, d, g, v, ʒ e z* / no português são produzidos com a vibração das cordas vocais, já os sons dos fonemas /*p, t, k, f, ʃ e s* / são produzidos sem a vibração das cordas vocais. Para saber se um som é vozeado ou não, basta colocar a mão na garganta durante a pronúncia do som.

Questão (c): O som é oral ou nasal?

R. Conforme a posição do véu palatino pode-se classificar o som como nasal ou oral. No final do véu palatino está localizada a úvula (Figura 2.2 - 10). Se a úvula estiver para baixo o ar terá acesso à cavidade nasal, então se classifica o som como nasal. Ao contrário, se a úvula estiver levantada o som é classificado como oral, pois o ar passa somente pela cavidade da boca. No português brasileiro os sons consonantais [m, n, e *nh] são considerados nasais. As demais consoantes são orais.*

Antes de responder a questão (*d*), deve-se ter em mente que os articuladores podem ser passivos ou ativos. Os articuladores ativos são aqueles que possuem a propriedade de se mover, são eles: O lábio inferior, a língua, o véu palatino e as cordas vocais (Figura 2.2). A língua por sua vez é dividida em quatro partes: ápice (ponta), parte anterior, parte medial e parte posterior (Figura 2.4). Vale salientar, que para os segmentos consonantais da língua Portuguesa do Brasil não é relevante se o articulador ativo é o ápice ou lâmina da língua.

Já os articuladores passivos são aqueles que não podem efetuar movimento, são eles: O lábio superior, os dentes superiores e o céu da boca. O céu da boca também é dividido em quatro partes distintas: Alvéolos (perto dos dentes superiores), o palato duro, o véu palatino (ou palato mole) e a úvula. A Figura 2.4 mostra detalhes sobre partes da língua.

Questão (d): Quais os articuladores envolvidos na produção dos sons?

R. A relação do posicionamento entre os articuladores ativos e passivos possibilita a classificação dos fonemas conforme o Lugar de Articulação. O lugar de articulação é portanto uma maneira de classificar os segmentos consonantais. São oito categorias referentes ao lugar de articulação sendo: Bilabial, Labiodental, Dental, Alveolar, Alveopalatal, Palatal, Velar e Glotal.

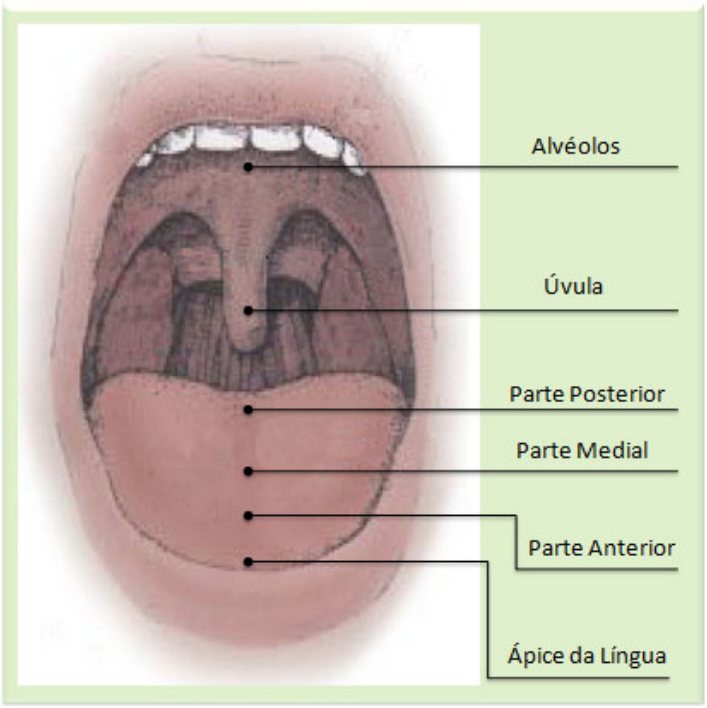

Figura 2.4: Esquema das partes da língua, alvéolos e úvula. Adaptada de (SILVA, 2003).

- Bilabial: O articulador ativo é o lábio inferior e o passivo o lábio superior. Exemplos: pá, má, bá.
- Labiodental: O articulador ativo é o lábio inferior e o passivo são os dentes incisivos superiores. Exemplos: fé, vá.
- Dental: O articulador ativo é a língua (ápice ou a lâmina) e o passivo são os dentes incisivos superiores. Exemplos: data, sapa.
- Alveolar: O articulador ativo é a língua (ápice ou a lâmina) e o passivo são os alvéolos. Exemplos: nada, lata.
- Alveopalatal (ou pós-velares): O articulador ativo é a parte anterior da língua e o passivo é a parte medial do palato duro. Exemplos: tia, dia, chá, já.
- Palatal: O articulador ativo é a parte média da língua e o passivo é a parte final palato duro. Exemplos: banha, palha.

➢ Velar: O articulador ativo é a parte posterior da língua e o passivo é a palato mole ou véu palatino. Exemplos: casa, gata, rata. A pronúncia do fonema /r/ em "rata" pode ser diferente conforme o dialeto do português.

➢ Glotal: Os músculos da glote comportam-se como articuladores. Exemplos: "rata", na pronúncia do dialeto de Belo Horizonte. Este som pode ser comparado ao ato de "escarrar" pronunciando o /r/ ao mesmo tempo.

Das categorias listadas acima, pode-se verificar que a Dental e a Alveolar são muito parecidas, e portanto, podem ser agrupadas em somente uma categoria chamada de Dental-Alveolar ou somente Alveolares.

Questão (e): Qual a maneira utilizada para obstruir a corrente de ar?

R. A maneira com que a corrente de ar é obstruída pelos articuladores proporciona a classificação chamada de Modo de Articulação. Este modo classifica os segmentos consonantais em oito categorias distintas, as quais são: Oclusivas, Nasal, Fricativa, Africada, Tepe, Vibrante, Retroflexa e Lateral.

➢ Oclusivas: Os articuladores produzem uma obstrução completa da passagem de ar. O véu palatino está levantado e o ar que vem dos pulmões encaminha-se totalmente para a cavidade oral. Portanto, as oclusivas são consoantes totalmente orais. Exemplos: pá, ta, cá, bar, dá, gol.

➢ Nasal: Os articuladores produzem uma obstrução completa da passagem de ar através da boca. Neste caso, o véu palatino está abaixado e o ar que vem dos pulmões segue para as cavidades oral e nasal. Exemplos: má, nada, banho.

➢ Fricativa: A aproximação dos articuladores produz uma fricção quando ocorre a passagem da corrente do ar. Neste caso, não há uma completa obstrução do ar, mas sim uma obstrução parcial que causa a fricção. Exemplos: fé, vá, sapa, zara, chá, já.

➢ Africada: É composta por dois movimentos: na fase inicial comporta-se como uma oclusiva impedindo a total passagem do ar. Em seguida, comporta-se como uma fricativa deixando a corrente de ar passar. Este caso é raro no português brasileiro e é encontrado somente nas sílabas [ti] de "tia" e [di] em "dia". As pronúncias são como "tchia" e "djia".

➢ Tepe: Também chamada de vibrante simples. Na Tepe, o articulador ativo toca rapidamente o articulador passivo ocorrendo uma rápida obstrução da passagem da

corrente de ar. O Tepe ocorre em português somente no uso do "r" em palavras como "cara", "brava" (pronunciando lentamente: "barava").

- Vibrante: Também chamada de vibrante múltipla. Neste caso, o articulador ativo toca algumas vezes o articulador passivo causando a vibração. Este tipo também é restrito ao som do "r" no português com presença somente na pronúncia de dois "erres" seguidos como em "marra".
- Retrofexa: O palato duro é o articulador passivo e a ponta da língua é o articulador ativo. Esta classe também é restrita ao som do "r" no português e aparece somente em palavras como "carta". O som aqui é aproximado a um "caipira" falando a palavra "mar".
- Laterais: O articulador ativo (língua medial) toca o passivo (palato duro) e a corrente de ar é obstruída na linha central do trato vocal. O ar será expelido por ambos os lados desta obstrução, tendo portanto uma saída lateral. As laterais ocorrem no português em exemplos tais como: "lá", "palha".

Pode-se ver claramente que nesta classificação existem quatro classes que são de grande importância: Oclusivas, Nasais, Fricativas e Laterais. Já as classificações Tepe, Vibrante e Retroflexa são importantes para diferenciar as diferentes pronúncias da letra "r".

Especificamente, neste trabalho, considerar-se-á somente dois tipos de "erres", o simples "r" e o duplo "rr", pois suas características são bem distintas e capazes de englobar as outras classificações.

Quanto a classe Africada, sua importância é pequena visto que a mesma aparece somente em sílabas que iniciam por "ti" ou "di". A classe Africada pode ser considerada como uma representação de um dialeto do Português, pois em determinadas regiões do País como no Norte e Nordeste ela é pouco utilizada. No nordeste e no norte do País predomina a pronúncia do /di/ e /ti/ na forma original.

Por fim, após a resposta das cinco questões acima, conclui-se que a classificação dos segmentos consonantais pode ser feita da seguinte maneira:

Oral/Nasal + Modo de Articulação + Lugar de Articulação + Grau de Vozeamento.

Exemplos: /*p*/ : oral, oclusiva, bilabial, não vozeada.
/v/ : oral, fricativa, labiodental, vozeada.
/m/: nasal, bilabial, vozeada.

Pode-se notar que muitas configurações de articuladores entre ativos e passivos não existem. Por exemplo: não existe um fonema que seja ao mesmo tempo oclusivo e palatal e da mesma forma não existe um fonema que seja bilabial e fricativo.

A Tabela 2.1 apresenta a classificação dos fonemas para o Português do Brasil. Esta tabela é parte da tabela internacional do alfabeto de fonética revisado em 1996 [SILVA, 2003]. A tabela completa do alfabeto fonético internacional pode ser encontrada na internet em [IPA, 2008]. A Tabela 2.1 apresenta somente os símbolos fonéticos consonantais relevantes à língua portuguesa.

Tabela 2.1: Símbolos fonéticos consonantais do Português do Brasil. Adaptada de [IPA, 2008].

Lugar de Articulação → / ↓ Maneira de Articulação		Bilabial	Labiodental	Dental ou Alveolar	Alveopalatal	Palatal	Velar	Glotal
Oclusiva	Não vozeada	p		t			k	
	Vozeada	b		d			g	
Africada	Não vozeada				tʃ			
	Vozeada				dʒ			
Fricativa	Não vozeada		f	s	ʃ		x	h
	Vozeada		v	z	ʒ		ɣ	ɦ
Nasal	Voz.	m		n		ɲ		
Tepe	Voz.			ɾ				
Vibrante	Voz.			ř				
Retroflexa	Voz.			ɹ				
Lateral	Voz.			l		ʎ		

Muitos símbolos fonéticos (som da letra) são idênticos aos símbolos ortográficos (letras), como por exemplo a letra “p” que tem o símbolo fonético /p/. Deste modo, quando se quer apresentar o símbolo ortográfico deve-se colocar o símbolo entre aspas (“”).

Já para representar o símbolo fonético, coloca-se o mesmo entre barras (//) ou entre colchetes como, por exemplo, em /’pa/ ou [‘pa]. O apóstrofo dentro dos colchetes indica a sílaba tônica. Com objetivo de padronizar a apresentação, as letras serão sempre apresentadas entre aspas e os fonemas serão apresentados entre barras.

Para melhor exemplificar a utilização dos símbolos fonéticos, a Tabela 2.2 apresenta a classificação com exemplos ortográficos e fonéticos para cada símbolo consonantal. Na Tabela 2.2, a classificação do segmento consonantal obedece a seqüência: Oral ou Nasal, Modo de Articulação + Lugar de Articulação + Grau de Vozeamento. Os símbolos da Tabela 2.2 estão dispostos na mesma seqüência da Tabela 2.1.

Tabela 2.2: Classificação dos símbolos fonéticos consonantais do Português, exemplos.

Símbolo	Classificação	Ex. Ortográfico	Fonemas
/ p /	Oral, Oclusiva, Bilabial, Não vozeada	"pata"	/'pata/
/ b /	Oral, Oclusiva, Bilabial, Vozeada	"bala"	/'bala/
/ t /	Oral, Alveolar, Bilabial, Não vozeada	"tapa"	/'tapa/
/ d /	Oral, Alveolar, Bilabial, Vozeada	"data"	/'data/
/ k /	Oral, Oclusiva, Velar, Não vozeada	"capa"	/'capa/
/ g /	Oral, Oclusiva, Velar, Vozeada	"gata"	/'gata/
/ tʃ /	Oral, Africada, Alveopatal, Não vozeada	"tia"	/'tʃia/
/ dʒ /	Oral, Africada, Alveopatal, Vozeada	"dia"	/'dʒia/
/ f /	Oral, Fricativa, Labiodental, Desvoz.	"faca"	/'faka/
/ v /	Oral, Fricativa, Labiodental, Vozeada	"vaca"	/'vaka/
/ s /	Oral, Fricativa, Alveolar, Não vozeada	"sala", "caça" , "paz"	/'sala/, /'kasa/, /'pas/
/ z /	Oral, Fricativa, Alveolar, Vozeada	"Zapata", "casa"	/za'pata/, /'kaza/
/ ʃ /	Oral, Fricativa, Alveopalatal, Não vozeada	"chá"	/'ʃa/
/ ʒ /	Oral, Fricativa, Alveopalatal, Vozeada	"já"	/'ʒa/
/X/, /ɣ/ *[1]	Oral, Fricativa, Velar, Desvoz./ Vozeada	"rata", "marra", "mar"	/'Xata/, /'maXa/, /'maX/
/ h /, / ɦ/ *[2]	Oral, Fricativa, Glotal, Desvoz. / Vozeada	"rata", "marra", "mar"	/'hata/, /'maha/, /'mah/
/ m /	Nasal, Bilabial, Vozeada	"mala"	/'mala/
/ n /	Nasal, Alveolar, Vozeada	"nada"	/'nada/
/ ɲ /	Nasal, Palatal, Vozeada	"banha"	/'bãɲa/
/ ɾ / *[3]	Oral, Tepe, Alveolar, Vozeada	"cara", "arara"	/'kaɾa/, /aɾaɾa/
/ ř / *[4]	Oral, Vibrante, Alveolar, Vozeada	"rata", "marra"	/řata/, /'mař a/
/ ɹ / *[5]	Oral, Retroflexa, Alveolar, Vozeada	"mar"	/'maɹ/
/ l /	Oral, Lateral, Alveolar, Vozeada	"lata" , "plana"	/'lata/, /plana/
/ ʎ /	Oral, Lateral, Palatal, Vozeada	"malha"	/'maʎa/

*(*1): O fonema / X / é a pronúncia de "rr" como se estivesse normal, ou seja, sem usar a garganta quanto se está pronunciando o som. Na Tabela 2.2 que o fonema / X / representa o "r" não vozeado e o /ɣ/ representa o mesmo fonema só que vozeado.*

*(*2): O fonema / h / é a pronúncia de "rr" como se estivesse "escarrando" ao mesmo tempo em que se pronúncia o som. Note na tabela que o fonema / h / representa o "rr" não vozeado e o / ɦ/ representa o mesmo fonema só que vozeado.*

*(*3): O fonema /ɾ/ é a pronúncia do "r" com o uso de um toque (fricção rápida) entre língua e os alvéolos (ou os dentes superiores). Exemplo: pronúncia do "r" em "arara".*

*(*4): O fonema /ř/ é a pronúncia do "rr" com o uso de vários toques (fricção) entre a língua e os alvéolos (ou os dentes superiores).*

*(*5): O fonema /ɹ / é a pronúncia do "r" com como se estivesse sendo falado por um "caipira". A ponta (ápice) da língua toca o palato duro (céu da boca). Este fonema aparece praticamente no final das sílabas.*

Como pode ser visto na primeira coluna da Tabela 2.2, existem sete fonemas que servem para identificar as diferentes pronúncias da letra "r" no português brasileiro. As observações (*) ajudam a diferenciar a pronúncia de cada um dos fonemas que representam os sons da letra "r".

Se todos os fonemas referentes às letras "r" e "rr" fossem agrupados, existiriam somente 20 fonemas consonantais para a língua portuguesa do Brasil. Entretanto, seria um erro juntar todos estes fonemas em somente uma classe, pois os sons destes fonemas são produzidos por diferentes articuladores e com diferentes tipos de obstrução da corrente de ar, deste modo, possuem características sonoras diferentes. No entanto, neste trabalho restringiu-se ao uso de apenas dois tipos de "r_s". Um alveolar /ɾ/ que representa a letra "r" , e outro glotal /ɦ/ que representa o som das letras "rr", ambos vozeados.

O fato de se juntar os fonemas /X/, /ɣ/, / h / e / ɦ/ em apenas um (/ ɦ/) esta na questão de que estes fonemas não formam classes diferentes, mas sim maneiras diferentes (muito parecidas) de se fazer a mesma pronúncia (sotaque). Já os fonemas / ɾ̆ / e / ɹ / não serão utilizados devido a sua rara utilização. Assim, para efeitos de classificação utilizou-se neste trabalho 19 classes consonantais diferentes.

2.3.2 Sistema Hierárquico Consonantal

Com base nas classificações das consoantes apresentadas na seção anterior, buscou-se encontrar o melhor modelo para a construção de um sistema hierárquico capaz de reconhecer com eficiência os 19 segmentos ou fonemas consonantais.

Durante a realização deste trabalho, diversas estruturas foram testadas. Inicialmente, procurou-se separar as consoantes em grandes grupos tais como: Vozeadas e Não vozeadas, Nasais e Orais etc. No entanto, os resultados não foram promissores. Após a análise das formas de onda e dos descritores extraídos (ver Capítulo V), decidiu-se pela criação de um sistema hierárquico composto, ou seja, primeiramente o sistema divide as consoantes pela Maneira de Articulação, em seguida faz-se a classificação pelo Lugar de Articulação.

Deste modo, as consoantes foram separadas em sete grupos principais conforme o Modo ou Maneira de Articulação sendo: Nasais, Laterais, Erres (“r”, “rr”), Oclusivas Vozeadas, Oclusivas Não vozeadas, Fricativas Vozeadas e Fricativas Não vozeadas. Nesta etapa, cada grupo é treinado na estratégia de “um contra todos”. Assim, somente um grupo será vencedor nesta etapa. Cada grupo ou classe possui dois ou três fonemas e a decisão final é feita através da estratégia de “um contra um”, ou pela combinação dos resultados para o caso de haver mais de dois fonemas. Por exemplo, se o grupo das Laterais for o vencedor na primeira etapa, então somente o item 11 será utilizado para decisão final entre os fonemas /l/ e /ʎ/ (“lh”). Se o grupo Nasal for o vencedor, então somente os itens 08, 09 e 10 serão testados para decidir entre os três possíveis fonemas nasais /m/, /n/ e /ɲ/ (“nh”). Os resultados obtidos no reconhecimento dos segmentos consonantais (duas etapas) podem ser vistos no Capítulo V. A Figura 2.5 apresenta a configuração da estrutura hierárquica de decisão para as consoantes, a qual será melhor detalhada no Capítulo V.

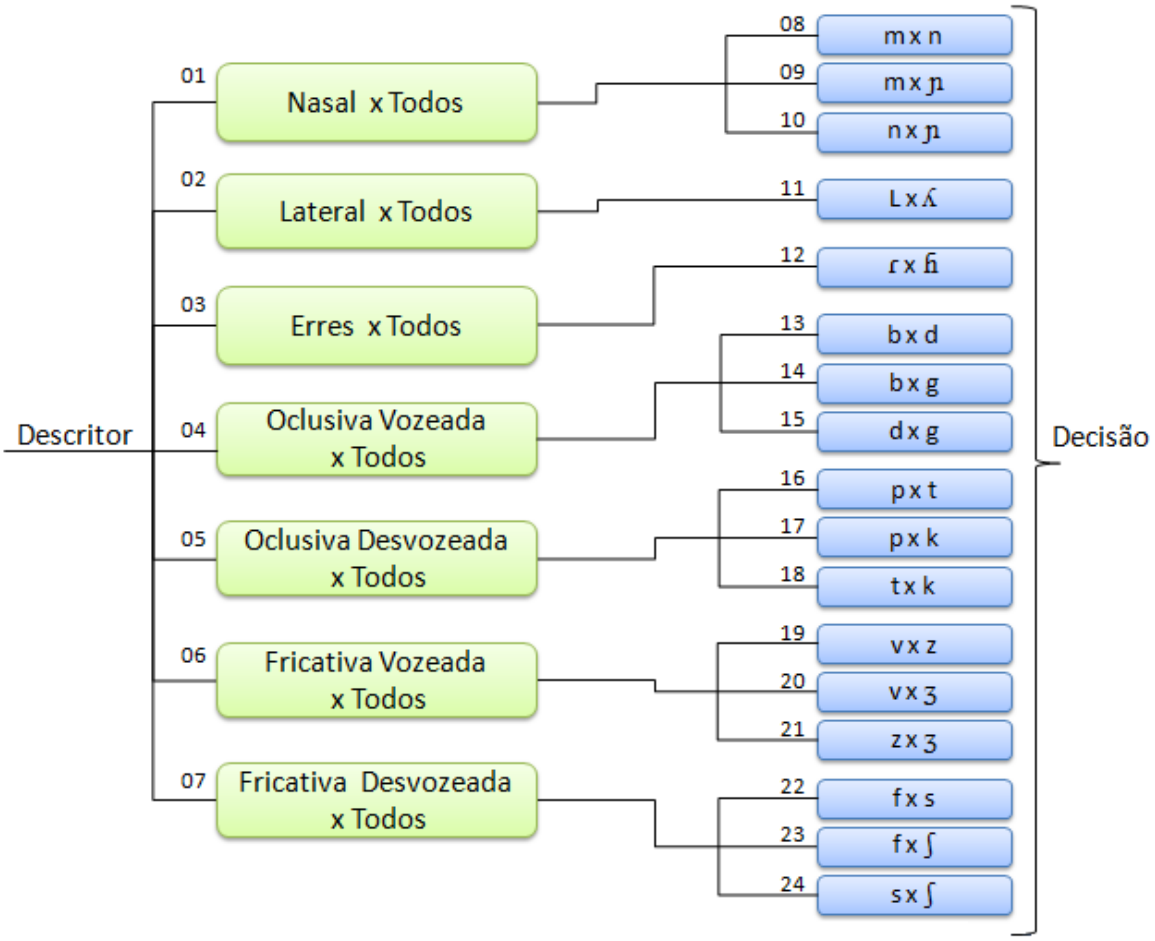

Figura 2.5: Estrutura hierárquica de decisão para as consoantes com duas etapas distintas.

2.3.3 Segmentos Vocálicos

Como citado anteriormente, os segmentos vocálicos são caracterizados por não obstruir a passagem de ar quando da pronúncia do som. Os segmentos vocálicos (vogais) são classificados levando-se em consideração os seguintes aspectos:

- Posição da língua em termos da altura (Vertical);
- Posição da língua em termos anterior, central ou posterior (Horizontal);
- Arredondamento ou não dos lábios.

A altura da língua representa a dimensão vertical ocupada pela língua dentro da cavidade bucal. A classificação é feita em graus de altura sendo: Alta, Média Alta, Média Baixa e Baixa.

A anterioridade ou posterioridade da língua refere-se a posição da mesma em relação a dimensão horizontal. Divide-se a cavidade bucal em três partes simétricas, sendo: uma parte localizada a frente da cavidade bucal (anterior), uma parte central e uma parte localizada no final da cavidade bucal (posterior).

A Figura 2.6 apresenta o triângulo formado pelo posicionamento da língua quando da produção das sete vogais orais na língua portuguesa.

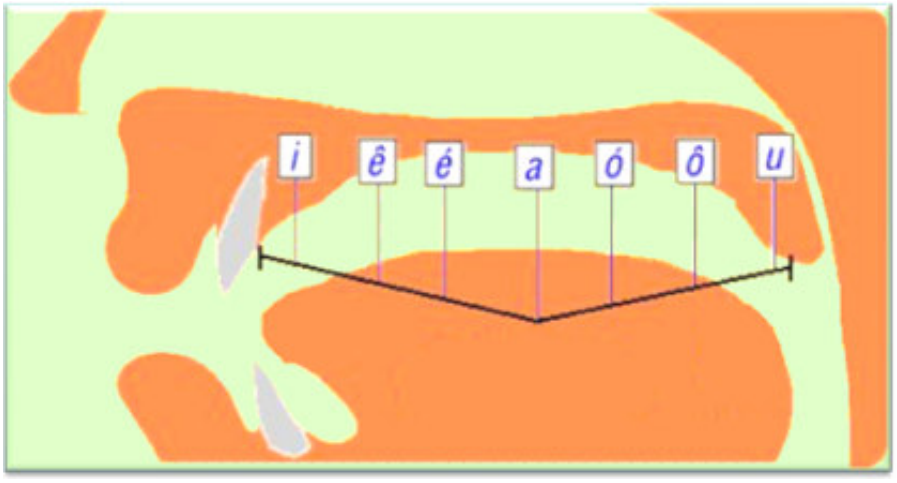

Figura 2.6: Posicionamento da língua na produção das vogais orais da língua portuguesa.

Quanto ao arredondamento dos lábios tem-se duas classificações possíveis: Lábios estendidos e Lábios arredondados.

Além das sete vogais orais, o português possui ainda mais cinco tipos de vogais chamadas de nasais. Para representar estas vogais nasalizadas acrescenta-se o símbolo "~"

sobre as cinco vogais / ã, ẽ, ĩ , õ , ũ /. Este fato ocorre principalmente quando as vogais são seguidas pelas letras “m”, “n” ou “nh”. Em alguns casos, o som nasalizado ocorre sem a presença das letras “m”, “n” ou “nh” como por exemplo na sílaba /fã/. A Tabela 2.3 apresenta os segmentos vocálicos da língua portuguesa do Brasil.

Tabela 2.3: Exemplos dos símbolos fonéticos vocálicos do Português do Brasil.

Fonema	Exemplo Ortográfico	Fonética
/ a /	“pá”	/‘pa/
/ ɛ /	“pé”	/‘pɛ/
/ i /	“vi”	/‘vi/
/ ɔ /	“avó”	/a‘vɔ/
/ u /	“tu”	/tu/
/ e /	“ipê”	/i‘pe/
/ o /	“avô”	/a‘vo/
/ ã /	“lã”	/lã/
/ ẽ /	"entre"	/ẽtɾe/
/ĩ /	“impar”	/ĩpaɾ/
/ õ /	“onze”	/õze/
/ ũ /	"um”	/ũ/

A Tabela 2.4 apresenta as vogais orais e nasais com a respectiva classificação quanto à posição da língua vertical e horizontal, além do arredondamento ou não dos lábios. É fácil notar que as vogais orais e nasais são produzidas com o mesmo posicionamento da língua, sendo diferenciadas apenas pela saída de ar (pelo nariz ou pela boca).

Tabela 2.4: Classificação dos símbolos fonéticos vocálicos do Português do Brasil.

Horizontal →	Anterior		Central		Posterior	
↓ Vertical	Arred.	Não-Arred.	Arred.	Não-Arred.	Arred.	Não-Arred.
Alta		/i/, /ĩ /			/u/, / ũ /	
Média-Alta		/e/, / ẽ /			/o/, / õ /	
Média-Baixa		/ ɛ /			/ ɔ /	
Baixa				/a/, / ã /		

A posição horizontal da língua é descrita como “anterior” quando se pronuncia os fonemas /i/, /ĩ/, /e/, /ẽ/ e /ɛ/ (lábios não arredondados). A posição da língua será posterior se for pronunciado um dos fonemas /u/, /ũ/, /o/, /õ/ e /ɔ/ (lábios arredondados). Já quando a posição for central, têm-se os fonemas /a/ e /ã/. A posição vertical da língua é descrita

como alta para os fonemas /i/, /ĩ/, /u/ e /ũ/, média-alta para os fonemas /e/, /ẽ/, /o/ e /õ/, média-baixa para os fonemas /ɛ/ e /ɔ/ e baixa para os fonemas /a/ e /ã/.

Como exemplo, pode-se citar os fonemas /a/ e /ã/ que possuem a seguinte classificação: Baixa, Central e de lábios estendidos (Não Arredondados). A única diferença entre os dois está no fato do primeiro ser oral e do segundo ser nasal.

2.3.4 Sistema Hierárquico para as Vogais

Com base nas classificações das vogais apresentadas na seção anterior, a exemplo das consoantes, buscou-se encontrar o melhor modelo para a construção de um sistema hierárquico capaz de reconhecer com eficiência os 12 segmentos ou fonemas vocálicos.

Novamente, diversas estruturas foram testadas. A melhor estrutura, ou seja, aquela que apresentou os melhores resultados nos experimentos (Capítulo V) é apresentada na Figura 2.7.

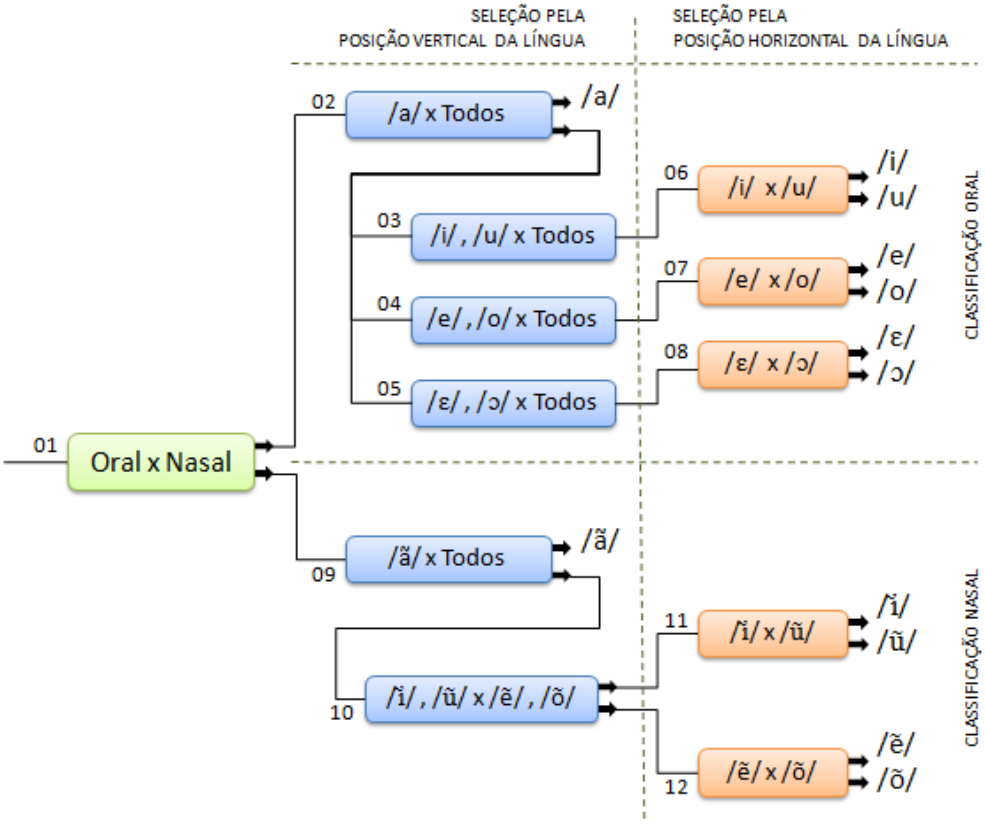

Figura 2.7: Sistema hierárquico para o reconhecimento das vogais.

O sistema apresentado na Figura 2.7 faz o reconhecimento *frame* a *frame* ou janela a janela do sinal. Na primeira decisão, o item especialista 01 irá separar os fonemas orais dos

nasais. Se o *frame* for identificado como oral, então o processo segue com as decisões 02 a 08. Se o *frame* for identificado como nasal então o processo de classificação é direcionado as decisões 09 a 12.

Os fonemas vocálicos orais são classificados inicialmente a partir da posição vertical da língua. Neste caso, utiliza-se os itens especialistas 02, 03, 04 e 05. Em seguida, as seleções 06, 07 e 08 fazem a classificação final pela posição horizontal da língua.

A seleção nasal utiliza os mesmos critérios da oral, ou seja, primeiramente faz classificação pela posição vertical em seguida pela posição horizontal da língua. Mais detalhes sobre o funcionamento desta estrutura hierárquica e os resultados obtidos nos experimentos serão apresentados no Capítulo V.

Devido ao fato de que o reconhecimento é feito *frame* a *frame, e*sta estrutura hierárquica pode ser utilizada também para identificação de encontros vocálicos chamados de *glide* ou semivogais.

2.3.5 Semivogais ou *Glides*

Do ponto de vista fonético, o que caracteriza um segmento como vocálico (vogal) ou consonantal (consoante) é o fato de haver ou não obstrução da passagem da corrente de ar pelo trato vocal. As semivogais podem apresentar características fonéticas de ambos segmentos vocálicos e consonantais, daí a necessidade de sua classificação ser diferenciada. Somente duas vogais são utilizadas como semivogais, são elas: o "i" e o "u". Com objetivo de diferenciar as semivogais das vogais, utiliza-se os símbolos fonéticos /y/ e /w/ para representar as semivogais das letras "i" e "u", respectivamente. No entanto, as letras "e, i, o, u, m e n", podem representar o som das semivogais /y/ e /w/, conforme a Tabela 2.5.

Tabela 2.5: Formação das semivogais do Português do Brasil.

Letras	Fonemas	Exemplo Ortográfico	Representação Fonética
"i" "e"	/y/	Boi Pães	/boy/ /pãys/
"o" "u"	/w/	Cão Touro	/kãw/ /towɾo/
m*	/y/ /w/	Sentem Cantam	/sẽtẽy/ /kãtãw/
n*	/y/	Hífen	/ifẽy/
* São semivogais somente nos encontros: *"am", " em"* e *"en"* (no fim da palavra)			

As semivogais ou *glides* aparecem sempre quando existe um encontro de duas vogais (neste caso letras), as quais não podem ser separadas dentro da sílaba. Este encontro é chamado de Ditongo. Por sua vez, os ditongos podem ser classificados de crescente ou decrescente, dependo da posição da vogal na sílaba. Se a vogal vier antes da semivogal o ditongo é decrescente. Já se a vogal vier depois da semivogal, o ditongo é crescente. Os ditongos podem ser ainda orais ou nasais.

Quando o encontro das vogais pode ser separado em duas sílabas, deve-se classificar o encontro como hiato. Existe o caso em que ocorre o encontro de duas semivogais com uma vogal. Este encontro vocálico é chamado de tritongo. Alguns exemplos de ditongos crescentes e decrescentes (orais e nasais) e tritongos são apresentados a seguir.

Exemplos de Ditongos Crescentes Orais:

/ya/ história, pátria, área, nívea, ígnia;
/ye/ Iemanjá, cárie;
/yɛ/ quieto, dieta;
/yo/ iodo;
/yu/ médio, áureo, níveo;
/wa/ égua, régua, magoa, nódoa;
/we/ lingüeta, coelho;
/wɛ/ goela, equestre;
/wi/ linguiça;
/wo/ aquoso;
/wɔ/ quota;
/wu/ vácuo, oblíquo.

Exemplos de Ditongos Crescentes Nasais:

/yã/ criança, ianque;
/yẽ/ lêmem (Arábia);
/yõ/ ionte;
/wã/ quando, quantidade;
/wẽ/ freqüente, qüinqüênio, cinqüenta;
/wĩ/ arquindo.

Exemplos de Ditongos Decrescentes Orais:

/ay/ caixa, pai, mais;
/aw/ pau, pausa, náufrago;
/ey/ lei, peito, jeito, feito, deixar;
/ɛy/ anéis, papéis, fiéis, coronéis;
/ew/ leu, breu, seu, teu;
/ɛw/ réu, véu, chapéu;
/iw/ viu, partiu, sentiu;

/ɔy/ heróis, rói, anzóis, faróis;
/oy/ foice, coisa, boi, coitado;
/uy/ azuis, fui, intuito, influi.

Exemplos de Ditongos Decrescentes Nasais:

/ãy/ cãibra, alemães, mãe;
/eĩ/ tem, cem, também;
/õy/ balões, sermões;
/uy/ muito;
/ãw/ cantam, falam, cão, pão.

Exemplos de Tritongos Orais:

/way/ Uruguai, quais, iguais, Paraguai;
/wey/ agüei, enxagüei, averigueis;
/wiw/ delinquiu;
/wow/ apaziguou, averiguou.

Exemplos de Tritongos Nasais:

/wey/ delínquem, enxágüem;
/wãw/ quão, mínguam, saguão;
/wõy/ saguões.

2.4 As Sílabas

A junção silábica (ou sílaba) ocorre quando se faz a junção de fonemas, ou seja, concatenação de um ou mais fonemas. Esta junção forma as sílabas as quais podem ser formadas de muitas maneiras.

O mais importante é que toda sílaba em português tem que ter uma vogal como parte central, ou seja, não existe nenhuma sílaba na língua portuguesa na qual não se tenha pelo menos uma vogal. Além disso, a vogal é geralmente o centro da sílaba e a parte com maior duração e energia [MAIA, 1985].

Estes itens, citados acima, são de suma importância no desenvolvimento prático deste trabalho. As formas mais comuns de formação das sílabas são apresentadas na Tabela 2.6.

Além das combinações apresentadas na Tabela 2.6, outras combinações mais raras são possíveis em nosso idioma. A palavra "*script*", por exemplo, apresenta sílaba com a combinação CCVCC.

Sílabas formadas somente por consoantes são bastante raras e praticamente inexistem em nossa língua. Geralmente, estas palavras são derivadas de outras línguas, por exemplo:

Stress e *Script.* A grande maioria das sílabas que aparecem na língua portuguesa são do tipo: CV, VC CCV e CVC [MAIA, 1985].

Tabela 2.6: Tipos de sílabas do Português Brasileiro.

Sílaba	Formação	Exemplo
V	Apenas uma vogal	"a"
CV	Consoante + vogal	"pá"
VC	Vogal + consoante	"ar"
VS	Vogal + semivogal	"ou"
CCV	Consoante + consoante + vogal	"Praga"
CSV	Consoante + semivogal + vogal	"Cópia"
CVC	Consoante + vogal + consoante	"Foz"
CVS	Consoante + vogal + semivogal	"Vai"
SVS	Semivogal + vogal + semivogal	"Uai"
VCC	Vogal + consoante + consoante	"Abstrato"
VSC	Vogal + semivogal + consoante	"Eis"
CCVC	Consoante + consoante + vogal + consoante	"Triste"
CCVS	Consoante + consoante + vogal + semivogal	"Plausível"
CSVS	Consoante + semivogal + vogal + semivogal	"Paraguai"
CVSC	Consoante + vogal + semivogal + consoante	"Meus"
CCVSC	2 Consoantes + vogal + semivogal + consoante	"Comprais"
CSVSC	Consoante + semiv. + vogal + semiv. + consoante	"Iguais"

A partir da análise das combinações dos fonemas nas sílabas, pode-se chegar a algumas conclusões:

- A vogal é a base de qualquer sílaba.
- Em torno da vogal gravitam semivogais e consoantes.
- As semivogais se ligam diretamente a uma vogal, antes ou depois desta.
- A consoante pode ocorrer adjacente a outra consoante, mas não há sílabas com três consoantes seguidas.
- A consoante pode se ligar com outra consoante, com uma semivogal ou com uma vogal.

A partir das regras de formação de sílabas, pode-se formar um modelo geral para a composição da sílaba conforme a Tabela 2.7, onde “C”, “S” e “V” representam as consoantes, semi-vogais e vogais, respectivamente.

Tabela 2.7: Modelo geral para formação de sílabas na língua portuguesa do Brasil.

Palavra	Tipo de Sílaba	Início	C	C	S	V	S	C	C	Fim
Abdicar	VC					a		b		dicar
Bueiro	VS	Bu				e	i			ro
Prato	CV	Pra		t		o				
Praga	CCV		p	r		a				ga
Cortês	CVC			c		o		r		tês
Abstrato	VCC					a		b	s	trato
Triste	CCVC		t	r		i		s		te
Plausível	CCVS		p	l		a	u			sível
Enxaguei	CSVS	Enxa		g	u	e	i			
Normais	CVSC	Nor		m		a	i	s		
Comprais	CCVSC	Com	p	r		a	i	s		
Script	CCVCC	S	c	R		i		p	t	

Alguns foneticistas afirmam que a sílaba é a unidade de emissão do aparelho fonador [MAIA, 1985]. Ao que tudo indica, existe uma correspondência entre os movimentos musculares do aparelho fonador e a emissão de sílabas. Nesse sentido, a sílaba pode ser considerada com a unidade fisiológica de pronúncia.

O que caracteriza a sílaba na abordagem fisiológica é a emissão de um conjunto de fonemas em um único movimento expiratório do aparelho fonador. Portanto, a sílaba pode ser interpretada como um movimento de força muscular que intensifica-se atingindo um limite máximo, após o qual ocorrerá a redução progressiva desta força [CAGLIARI, 1981].

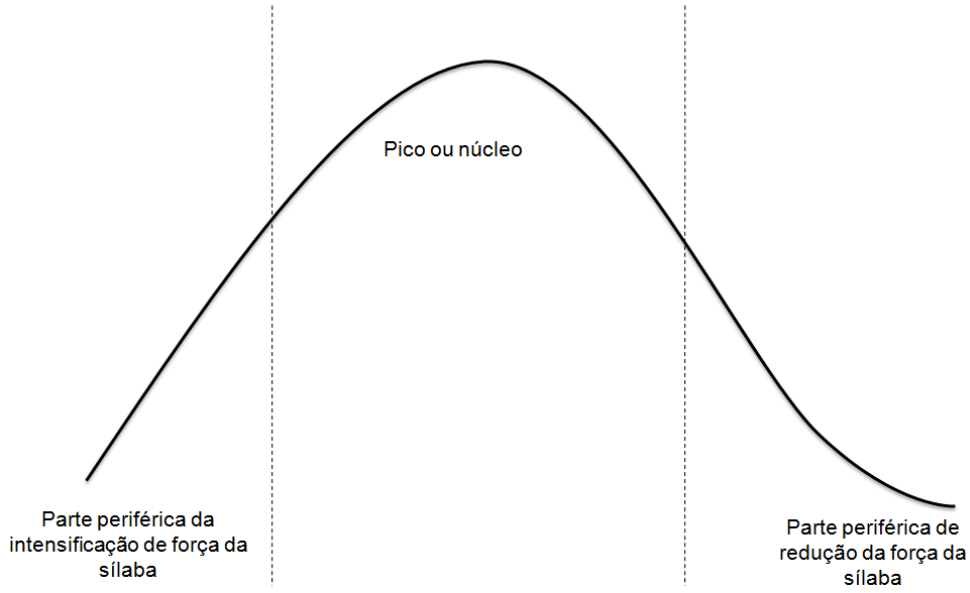

Figura 2.8: Esquema da força muscular empregada na produção de uma sílaba [MAIA, 1985].

Na Figura 2.8, pode-se ver que o pico ou núcleo da força muscular empregada na produção de uma sílaba ocorre praticamente no meio do gráfico. Este ponto central coincide com o ponto central do sinal da vogal que existe na sílaba. Diante disto, teoricamente pode-se usar a vogal (com auxílio de outras ferramentas) na separação das sílabas.

2.5 Difones

Para a fonética, um difone é um par de fonemas adjacentes. O termo difone também é geralmente usado para se referir a uma transição entre dois fonemas. Basicamente, um difone é um encontro de dois fonemas na mesma sílaba ou em sílabas adjacentes. Exemplo: "casa": /ka-sa/, difones: [ka] – [as] – [sa].

Do exemplo acima, pode-se ver que a palavra "casa" possui três difones, ou seja, dois difones referentes aos encontros fonéticos das sílabas inicial e final [ka] e [sa], e o encontro entre os fonemas /a/ e /s/ das diferentes sílabas adjacentes [as].

Neste trabalho especificamente, os difones são utilizados no reconhecimento dos fonemas que antecedem (pré-vocálicos) ou que aparecem após a vogal (pós-vocálicos), ou seja, as consoantes que aparecem antes e depois das vogais. A utilização dos difones para o reconhecimento destas consoantes se dá pelo fato de ser praticamente impossível separar alguns segmentos consonantais devido ao pouco tempo (muitas vezes cerca de apenas 20ms a 30ms) de duração dos mesmos [MAIA, 1985]. Um encontro de três fonemas é chamado de trifone. Os trifones são de muita importância no reconhecimento de palavras na língua inglesa [YOUNG, 1996].

2.6 Estado da Arte: Unidades menores do que a palavra

Para formulação de um sistema de reconhecimento de voz, deve-se primeiro decidir qual a unidade a ser reconhecida, ou seja, reconhecer diretamente a palavra ou utilizar as unidades fonéticas menores, tais como: sílabas, trifones, difones ou fonemas. Uma língua natural, tal como o português, possui cerca de 500.000 palavras.

Assim, sistemas de reconhecimento restritos as palavras exigem grandes quantidades de processamento e de armazenamento o que torna difícil ou até inviabiliza o reconhecimento contínuo em tempo real.

Nos últimos anos, os esforços têm se voltado às unidades fonéticas menores do que a palavra. Um exemplo é o trabalho de Alcaim *et* al [ALCAIM, 2001] onde as sílabas foram utilizadas como unidades de reconhecimento no modo dependente do locutor. No entanto, os autores alertam que as sílabas são atraentes somente quando o número de padrões a serem treinados é pequeno. As sílabas podem ter mais de 2000 padrões e não são muito úteis em línguas como o inglês que não possui uma divisão silábica trivial. Para sistemas de língua inglesa, os trifones são as unidades mais utilizadas, mas são de difícil treinamento [YOUNG, 1996].

A seguir são apresentados alguns trabalhos, das três últimas décadas, cujo foco das pesquisas é o reconhecimento de fonemas ou unidades menores do que a palavra utilizando as mais diversas técnicas.

Osamu Fugimura [FUGIMURA, 1975] propôs que a unidade básica para o reconhecimento de voz fosse a sílaba ao invés da palavra ou do fonema. Fugimura explica que o reconhecimento de palavras isoladas já era possível com a tecnologia da época (em 1975) e que o grande empecilho para o uso da palavra como unidade padrão era a grande quantidade de padrões. O autor também considerou que, na época, não havia possibilidade de usar os fonemas como unidade padrão no reconhecimento de voz devido a pouca duração (tempo) do fonema e de sua variação conforme a posição dentro da palavra.

Davis e Mermelstein [DAVIS, 1980] fizeram um importante trabalho de reconhecimento de sílabas simples formadas por CVC (consoante-vogal-consoante). Os autores compararam o desempenho de vários descritores.

Fujisaki e Tominaga [FUJISAKI, 1982] trabalharam com o reconhecimento de fonemas sonoros */p/*, */b/* e */g/*. Os autores concluíram que a primeira formante F_1 contribui muito pouco para separação das consoantes citadas e que as formantes F_2, F_3 e F_4 são suficientes para separar estas consoantes.

Rosenberg *et al* [ROSENBERG, 1983] propuseram o uso de "*demisyllable*" (metade da sílaba) como unidade fonética para o reconhecimento de voz. Apesar do esforço em tentar provar a viabilidade da utilização da meia sílaba, os resultados mostraram que a taxa de erro de reconhecimento variou de 18 a 33%, enquanto que a taxa de erro quando utilizadas as palavras variou de 9 a 15%. Uma possível explicação para esta alta taxa de erro talvez esteja na dificuldade de separação das sílabas na língua inglesa, explicou o autor.

Raman e Yegnanarayana [RAMAN, 1984] fizeram um interessante trabalho sobre o reconhecimento de palavras isoladas da língua indiana (Hindi) com alto grau de similaridade. Os autores concluíram que a melhor forma de distinguir palavras com similaridade fonética é através da utilização da parte consonantal da sílaba. O problema é que a parte consonantal de uma sílaba tem pouca duração, cerca de 20ms a 30ms. O mesmo acontece na língua portuguesa [MAIA, 1985].

Kepuska e Gowdy [KEPUSKA, 1989] propuseram a utilização da rede neural SOM (*Self-Organizing Map*), a qual é uma rede neural de treinamento não supervisionado, no reconhecimento de fonemas. O objetivo do trabalho foi investigar a utilidade do SOM na representação de fonemas que variam devido a articulação. Segundo os autores, o problema do reconhecimento de fonemas está no efeito chamado *"overlapping*" (superposição), isto é, em uma dada janela do sinal de voz podem estar presentes características de dois fonemas devido aos efeitos de articulação do trato vocal.

Meng et al [MENG, 1991] fizeram um importante trabalho de comparação do desempenho das características espectrais *versus* as características fonéticas. Como resultado, os autores concluíram que o desempenho de reconhecimento das características espectrais é praticamente igual ao das características fonéticas.

Davenport e Garudadri [DAVENPORT, 1991] utilizaram a *Wavelet* para extrair características acústico-fonéticas das palavras. Foram realizados dois testes: um com fonemas sonoros e não sonoros (surdos) e outro com fonemas fricativos e não fricativos. O destaque foi a taxa de acerto no reconhecimento das vogais com 95,7% .

Malbos *et al* [MALBOS, 1994] realizaram o reconhecimento das consoantes oclusivas (/p/, /*k*/, /*t*/, /*b*/, /*g*/ e /*d*/) da língua francesa utilizando a transformada *Wavelet*. Foram obtidos bons resultados na classificação das consoantes oclusivas não vozeadas (/*p*/, /*k*/ e /*t*/), alcançando a taxa de 95,5% de acerto no reconhecimento.

Rangoussi e Delopoulos [RANGOUSSI, 1995] analisaram o reconhecimento de fonemas não vozeados (surdos) e oclusivos /*k*/, /*p*/ e /*t*/. Os autores afirmaram que os fonemas vozeados são fáceis de classificar, pois possuem uma natureza quase periódica, enquanto que os não-vozeados são mais difíceis de classificar devido a sua natureza irregular. Como resultado, os autores encontraram uma média de classificação com 94% de acerto para estes fonemas específicos.

Tan *et al* [TAN, 1996] testaram o desempenho da *Wavelet* comparada com a transformada de Fourier (FFT) usada com escala Mel no reconhecimento de fonemas. A *Wavelet* apresentou melhor desempenho.

Marchesi *et al* [MARCHESI, 1996] fizeram um estudo de reconhecimento de vogais orais do Português Brasileiro, utilizando as freqüências fundamentais como descritores. Como resultado, obtiveram uma taxa de erro no reconhecimento de 2,5% para o fonema /*a*/, de 0,2% para o fonema /o/ e 0% para os outros fonemas vocálicos.

Long e Datta [LONG, 1996] procuram encontrar qual a *Wavelet* que melhor descreve as características fonéticas de um dado fonema. Os autores determinaram que a *Wavelet* do tipo *Daubechies* representa melhor os fonemas fricativos, devido a sua estrutura ter a forma parecida com a de um sinal de ruído. Já a *Wavelet* do tipo *Morlet* é mais apropriada para representar os fonemas vozeados (que usam as cordas vocais), pois estes fonemas têm comportamento quase-periódico, tal como esta *Wavelet.*

Abdelatty *et al* [ABDELATTY, 1998] investigaram o uso das características acústico-fonéticas para o reconhecimento das consoantes fricativas para língua inglesa. Os resultados sugerem a possibilidade da classificação dos fonemas fricativos por meio das características acústico-fonéticas relacionadas ao local de articulação. Em um trabalho seguinte, Abdelatty *et al* [ABDELATTY, 1999] separaram os fonemas em quatro classes: Oclusivas, Fricativos, Nasais e Vogais. Os algoritmos de classificação foram baseados nas características acústico-fonéticas e como resultado a classificação apresentou taxas de acerto de 92% em média.

Outro trabalho que utilizou as consoantes oclusivas (/*p*/, /*k*/ e /*t*/) foi realizado por Lukasik [LUKASIK, 2000]. Em todos os testes, o autor usou a WPT (*Wavelet Packet Transform*), a qual obteve melhores resultados do que os coeficientes *cepstrais* de Fourier (MFCC - *Mel-Frequency Cepstral Coefficient)*.

Deshmukh *et al* [DESHMUKH, 2002] utilizaram os parâmetros acústico-fonéticos no reconhecimento de voz com bons resultados. Estes parâmetros descrevem a posição dos articuladores na produção do sinal de voz.

A publicação de Rodrigues e Yehia [RODRIGUES, 2002] tinha como objetivo identificar qual descritor apresentava melhor robustez no caso de reconhecimento de voz com diferentes locutores (caso independente do locutor). As vogais orais do português foram utilizadas como unidades de reconhecimento.

O trabalho de Farooq e Datta [FAROOQ, 2003] trata do reconhecimento de fonemas utilizando a *Wavelet Packet*. Como resultado, foi obtido um aumento na taxa de reconhecimento de aproximadamente 10% em comparação com resultados anteriores os quais foram citados no referido trabalho.

Os trabalhos apresentados acima são apenas uma pequena amostra cronológica dos inúmeros esforços feitos nas últimas três décadas no sentido de encontrar uma melhor forma de enfrentar o problema do reconhecimento de voz. Estes trabalhos enfatizam que a utilização de unidades menores do que a palavra e das características acústico fonéticas são viáveis do ponto de vista das taxas de reconhecimento obtidas.

Capítulo III

3. Descritores do Sinal de Voz: *Wavelet Packet*

Como descrito no Capítulo II, a voz humana é produzida pela vibração do ar que é expulso dos pulmões, passa pelas cordas vocais e é modificado pela boca, lábios e a língua. Todo este sistema de produção da voz é chamado de trato vocal. Cada ser humano possui um sistema de produção da voz (trato vocal) diferente de qualquer outro. Simplificando, as cordas vocais de cada ser humano são diferentes, tal como as digitais. Este fato não restringe-se somente às cordas vocais, pois todas as cavidades e articuladores passivos e ativos, que colaboram para formação do sinal de voz, são diferentes em cada ser humano.

Esta característica é muito importante quando o objetivo é identificar uma determinada pessoa pela análise do sinal de voz. Esta linha de pesquisa é denominada de Reconhecimento do Locutor (pessoas), e apesar de ter uma definição simples, tem um caráter interdisciplinar e alto nível de complexidade. Um importante trabalho sobre esta linha de pesquisa pode ser encontrado em [FECHINE, 2000].

Da mesma forma que a diferença entre os tratos vocais dos seres humanos ajuda na identificação individual do locutor, dificulta quando pretende-se reconhecer a palavra que um determinado individuo está falando. Para aumentar a complexidade, uma única pessoa pode pronunciar uma mesma palavra de várias maneiras diferentes dependendo dos fatores emocionais, físicos, locais, de saúde etc., presentes no momento da pronúncia.

Assim, dois sinais de voz (no domínio do tempo) de uma mesma palavra são diferentes. Isto independe se a palavra foi pronunciada pela mesma pessoa ou não. Portanto, para fazer o reconhecimento de uma palavra é preciso encontrar determinadas características dentro do sinal de voz que se repitam quando esta mesma palavra for pronunciada. Estas características do sinal voz permitem classificar um determinado sinal de voz, ou seja, descrevem um determinado padrão de voz. Por este motivo, estas características são chamadas de descritores do sinal de voz.

Diversas são as ferramentas matemáticas, transformadas e métodos computacionais utilizados para extração de descritores do sinal de voz. Rabiner e Juang em seu livro

Fundamentals of Speech Recognition [RABINER, 1993] fizeram um minucioso estudo sobre diversas técnicas de extração de descritores, tais como: Análise Espectral, Transformada de Fourier, Banco de Filtros, LPC (*Linear Predictive Coding*), Autocorrelação, Características Acústicas, etc., além da ferramenta mais utilizada pelos pesquisadores no reconhecimento de voz, a MFCC (*Mel-Frequency Cepstral Coefficient*).

Não é o foco deste trabalho encontrar a melhor ferramenta para extração dos descritores do sinal de voz. Deste modo, duas ferramentas importantes são apresentadas neste capítulo: MFCC e Transformada *Wavelet*. A primeira (MFCC) por ser a ferramenta padrão utilizada no reconhecimento de voz e, portanto, é uma ferramenta base para comparação de resultados. Já a Transformada *Wavelet*, além de ser um dos focos deste trabalho, tem sido nos últimos anos objeto de estudo de inúmeras pesquisas.

Conforme pode ser visto na última seção deste capítulo, que descreve o estado da arte dos descritores do sinal de voz, a Transformada *Wavelet* apresenta, em muitos casos, melhores resultados do que os descritores MFCC no reconhecimento de fonemas, sílabas e palavras isoladas.

3.1 Transformada de Fourier

Coube a Jean-Baptiste Joseph Fourier, um matemático e físico francês, o mérito pela investigação sobre a decomposição de funções periódicas em séries trigonométricas convergentes, chamadas séries de Fourier, e a sua aplicação aos problemas da condução do calor. A Transformada de Fourier foi designada em sua homenagem.

A análise de Fourier é uma das técnicas matemáticas com maior número de aplicações práticas. Além de ser utilizada extensivamente no cálculo numérico e nas mais diversas áreas das ciências aplicadas e engenharias, a análise de Fourier constitui ainda a base do processamento de sinais. A denominação "Transformada de Fourier" refere-se à Transformada de Fourier para funções contínuas e representa qualquer função integrável $x(t)$ como a soma de exponenciais complexas com freqüência angular (ω) e amplitude complexa $F(\omega)$ conforme,

$$F(\omega) = \int_{-\infty}^{+\infty} x(t) \, . \, e^{-i\omega t} dt \qquad (3.1)$$

onde $x(t)$ é um sinal periódico, ω é a freqüência e $i = \sqrt{-1}$ e $e^{-i\omega t} = cos(\omega) - isen(\omega)$.

Assim a Equação (3.1) pode ser interpretada como o sinal $x(t)$ sendo decomposto em senos e cossenos de diferentes freqüências.

Pode-se obter o sinal original $x(t)$ através da transformada inversa de Fourier dada pela Equação (3.2).

$$f(t) = \mathcal{F}^{-1}\big(F(\omega)\big) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(\omega)\,.\,e^{i\omega t}dt \qquad (3.2)$$

A Transformada de Fourier permite a análise de características não percebidas diretamente no domínio original do sinal (domínio do tempo). No entanto, a decomposição espectral através da Transformada de Fourier não possibilita a determinação completa da relação espaço freqüência de $x(t)$, produzindo uma descrição espectral global do sinal, ou seja, a transformada de Fourier diz quais as freqüências que estão presentes no sinal, mas não onde elas se encontram.

Em um esforço para corrigir esta deficiência, Dennis Gabor adaptou a transformada de Fourier para analisar somente uma seção pequena do sinal em um determinando momento e chamou esta técnica de "janelamento" do sinal. A adaptação de Gabor, chamada de *Short Time Fourier Transform* (STFT), representa o sinal em uma função bidimensional do tempo e da freqüência [RABINER, 1993].

A STFT representa uma combinação do compromisso entre as bases do tempo e freqüência de um sinal, ou seja, fornece alguma informação sobre ambos, mostrando quando e em que freqüências um evento do sinal ocorre. Entretanto, pode-se somente obter esta informação com precisão limitada, e essa precisão é determinada pelo tamanho da janela. O inconveniente da STFT é que uma vez que se escolhe um tamanho particular para a janela do tempo, esta janela será a mesma para todas as freqüências. A Figura 3.1 apresenta a forma gráfica da análise feita pela STFT e a correspondente representação na escala tempo/freqüência.

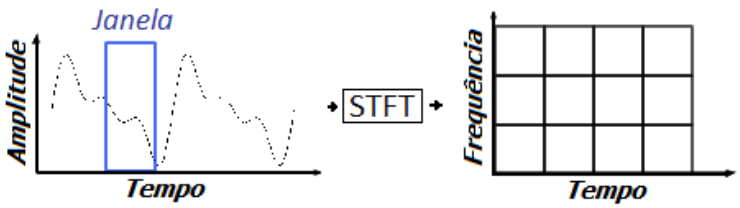

Figura 3.1: Diagrama da análise feita pela STFT.

No entanto, sabe-se que muitos sinais exigem uma aproximação mais flexível onde seja possível variar o tamanho da janela para determinar mais precisamente as informações no tempo ou na freqüência. Esta flexibilidade é uma das características apresentadas pela *Wavelet.* Para melhor ilustrar esta flexibilidade, a Figura 3.2 apresenta a análise espectral através da Transformada de Fourier e da Transformada *Wavelet* de dois sinais.

O primeiro sinal, Figura 3.2 (a), consiste da superposição de duas freqüências (sen 10t e sen 20t). Já o segundo sinal (b), consiste das mesmas freqüências aplicadas separadamente a cada uma das metades da duração do sinal.

As Figuras 3.2 (c) e (d) mostram os espectros dos dois sinais obtidos através da Transformada de Fourier, ou seja, $|f(\omega)|^2 \times \omega$, de (a) e (b) respectivamente.

Finalmente, as Figuras 3.2 (d) e (e) mostram a magnitude da Transformada *Wavelet* dos mesmos sinais (usando para isso a *Wavelet* mãe do tipo Morlet).

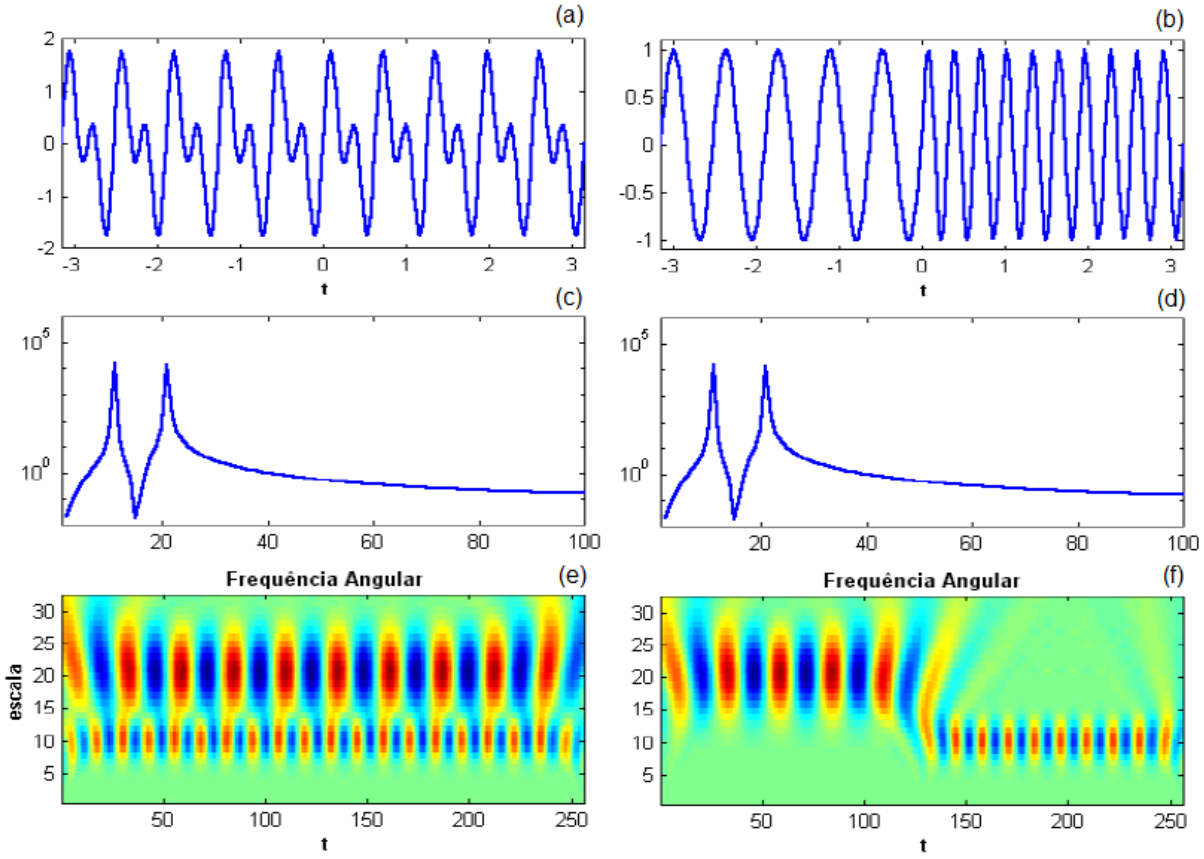

Figura 3.2: Análise de dois sinais usando as transformadas de Fourier e *Wavelet* [PROTAZIO, 2002].

Pode-se observar, que o espectro de frequência obtido através da transformada de Fourier é praticamente o mesmo para os dois sinais conforme as Figuras 3.2 (c) e (d). Já o

espectro obtido pela transformada *Wavelet* consegue diferenciar os dois sinais inclusive determinando a região de mudança de um sinal para o outro, conforme pode ser visto na Figura 3.2 (e).

3.1.1 DFT: Transformada Discreta de Fourier

Quanto deseja-se utilizar a transformada de Fourier em computadores é preciso discretizar o sinal $x(t)$ em um sinal x_k, sendo que o sinal existe somente para os valores discretos de $k = 0, 1, 2, \dots, N$. Onde N é o tamanho da amostra do sinal ou tamanho da janela (quantidade de pontos do sinal amostrado).

Para trabalhar com um sinal discreto, utiliza-se a Transformada Discreta de Fourier (DFT – *Discret Fourier Transform*) dada pela Equação (3.3).

$$f_j = \sum_{k=0}^{n-1} x_k e^{\frac{2\pi i}{n} jk} \qquad j = 0, 1, 2, \dots, n-1 \tag{3.3}$$

Da mesma forma, o sinal original x_k pode ser reconstruído com a transformada discreta inversa de Fourier, conforme Equação (3.4).

$$x_k = \frac{1}{n} \sum_{j=0}^{n-1} f_j \, e^{\frac{-2\pi i}{n} jk} \qquad k = 0, 1, 2, \dots, n-1 \tag{3.4}$$

No entanto, a transformada inversa discreta não pode reproduzir o domínio do tempo inteiro, a menos que a entrada seja periódica. Portanto, diz-se freqüentemente que o DFT é uma transformada para a análise de Fourier de funções discretas no tempo e de domínio finito. Um problema da transformada de Fourier é a sua complexidade computacional. Para a computação da transformada de Fourier são necessários $O(n^2)$ operações. Para resolver este problema foi desenvolvido o algoritmo FFT (*Fast Fourier Transform*). Este algoritmo reduz a complexidade para $O(n \log n)$. O algoritmo da FFT foi desenvolvido por Cooley-Tukey [COOLEY, 1965], e é usado como base para o cálculo dos coeficientes MFCC.

3.2 Os Descritores MFCC

Os coeficientes Mel-Cepestrais surgiram devido aos estudos na área de psicoacústica (ciência que estuda a percepção auditiva humana), os quais mostraram que a percepção

humana das freqüências de tons puros não segue uma escala linear. Isto estimulou a idéia de serem definidas freqüências subjetivas de tons puros, da seguinte forma: para cada tom com freqüência *f*, medida em Hz, define-se um tom subjetivo medido em uma escala que se chama "escala Mel".

Para definir os coeficientes MFCCs, é necessário descrever sobre cinco itens importantes: A Transformada Rápida de Fourier (FFT), os coeficientes Cepstrais, a escala Mel, o banco de filtros triangular espaçados pela escala Mel e a Transformada Discreta do Cosseno (DCT).

Para se obter os coeficientes no domínio da freqüência, utiliza-se a Transformada Rápida de Fourier do sinal (FFT). Como citado anteriormente, a FFT é um algoritmo que reduz a complexidade computacional na obtenção dos coeficientes da Transformada de Fourier. Uma descrição sobre o algoritmo FFT incluindo a teoria, diferentes propostas de implementação e aplicações, pode ser encontrada no livro de Brigham [BRIGHAM, 1998].

Devido a sua grande popularidade, diversos *softwares* de simulação e programação possuem *Toolboxes* que permitem o cálculo da FFT, o Matlab®[1] é um destes programas.

A representação cepstral ou "C*epstrum"* de um sinal pode ser definida como a Transformada de Fourier do logaritmo da Transformada de Fourier do Sinal de voz. A sequência de cálculo do *cepstrum* (Cps) pode dada por,

$$Sinal \rightarrow TF \rightarrow abs[] \rightarrow log \rightarrow (fase) \rightarrow TF \Rightarrow Cepstrum$$

ou pela Equação (3.5) conforme,

$$Cps = TF(log(|TF(x_k)|) + j2\pi m) \tag{3.5}$$

onde, TF é a transformada de Fourier, x_k é o sinal de voz discreto e m é o inteiro exigido para encontrar corretamente o ângulo ou a parte imaginária do log da função complexa.

O Mel é uma unidade de medida da freqüência percebida de um tom. Como referência, definiu-se a freqüência de 1 kHz, com potência 40 dB acima do limiar mínimo de

[1] Criado pela *MathWorks® Inc.*, o MATLAB é um software que permite: a manipulação de matrizes, a criação de gráficos de funções e de dados, a criação e execução de algoritmos, além de possuir uma vasta gama de funções pré-definidas chamadas de *Toolboxs* de varias áreas da engenharia, estatística, física etc.

audição do ouvido humano como 1000 mels. Os outros valores subjetivos foram obtidos através de experimentos. Estes experimentos permitiram verificar que, o mapeamento entre a escala de freqüência real em Hz e a escala de freqüências percebida em Mel, é aproximadamente linear abaixo e logarítmica acima dos 1000 Hz.

Portanto, a escala Mel faz com que as faixas de freqüência sejam posicionadas em uma escala logarítmica, a qual se aproxima da resposta do sistema auditivo humano. A escala Mel foi originalmente introduzida por Stevens e Volkman [STEVENS, 1937]. A Equação (3.6) faz a conversão de Hz para Mel e a Equação (3.7) de Mel para Hz, respectivamente.

$$m = 1127{,}01048\, log_e\left(1 + \frac{f}{700}\right) \tag{3.6}$$

$$f = 700\left(e^{\frac{m}{1127{,}01048}} - 1\right) \tag{3.7}$$

A Figura 3.3 apresenta a escala Mel. Pode-se ver que abaixo da freqüência de 1000 Hz o espaçamento dos tons é praticamente linear e acima logarítmico.

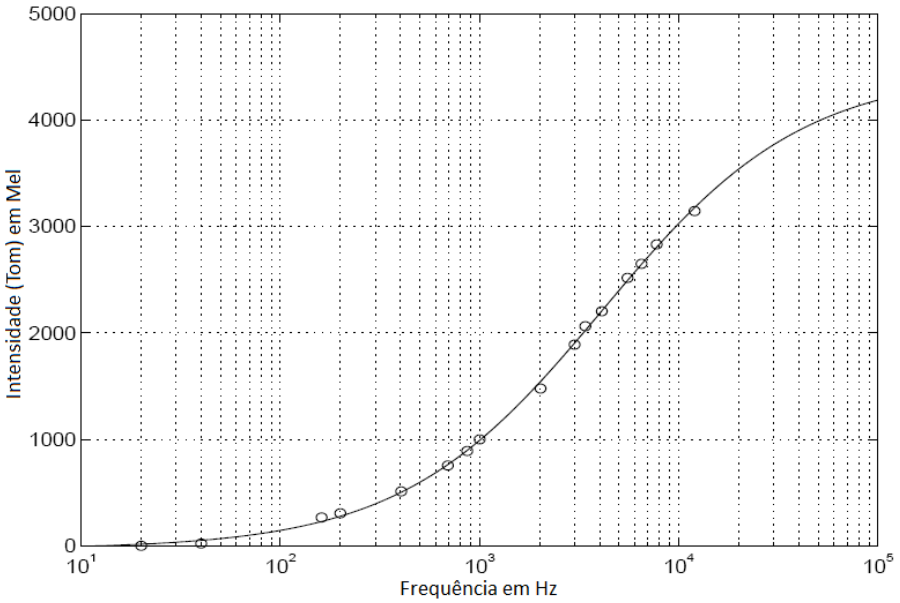

Figura 3.3: Escala Mel [COMBRINCK, 1996].

A maior utilidade da escala Mel está na criação do banco de filtros Mel. O banco de filtros Mel consiste na sobreposição de filtros triangulares. Estes filtros têm freqüências centrais espaçadas linearmente e a largura de banda é espaçada conforme a escala mel. Para a faixa de freqüências de interesse da voz humana, geralmente utiliza-se de 12 a 30 filtros.

A Figura 3.4 apresenta o banco de filtros mel com 18 bandas. Até 1000 Hz os filtros possuem espaçamento linear, após este valor o espaçamento é logaritmo.

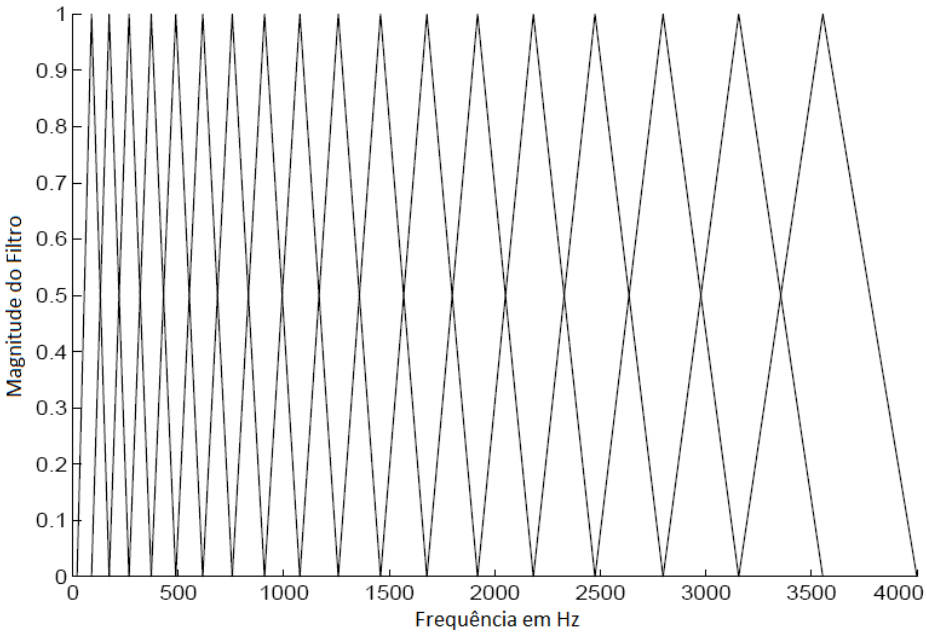

Figura 3.4: Banco de filtros triangular em escala Mel [COMBRINCK, 1996].

A Transformada Discreta do Cosseno (DCT) é muito utilizada na compressão de dados. Um típico exemplo é formato JPEG (*Joint Photographics Experts Group*) de compressão de imagens que utiliza a DCT com base para compressão [BLINN, 1993]. Pode-se considerar a DCT como similar à Transformada Discreta de Fourier (DFT), mas neste caso utilizando somente números reais.

Existem oito variações padrões para a DCT. A forma definida pela Equação (3.8) é a mais utilizada para os sinais de voz. Sendo X_k os coeficientes resultantes da transformada discreta do cosseno de um sinal x_n , temos:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{n}\left(n+\frac{1}{2}\right)k\right] \qquad k = 0, 1, \ldots, N-1 \qquad (3.8)$$

Basicamente, o algoritmo para obtenção dos coeficientes MFCC pode ser resumido através dos passos a seguir:

- Calcular a Transformada Rápida de Fourier (FFT) do sinal (espectro do sinal).
- Obter as amplitudes do logaritmo do espectro do sinal, usando o banco de filtros triangular em escala mel.

- Aplicar a Transformada Discreta do Cosseno (DCT) às amplitudes do logaritmo (como se fosse um sinal).
- Os Coeficientes MFCC são as amplitudes obtidas no passo anterior.

3.3 A Transformada *Wavelet*

A primeira menção às *Wavelets* aconteceu em 1909 por A. Haar [HAAR, 1910]. No entanto, as *Wavelets* de Haar ficaram no anonimato por muitos anos e por um período muito longo elas continuaram a ser a única base ortonormal de *Wavelet* conhecida.

A fase moderna das *Wavelets* iniciou-se nos anos 80 com vários trabalhos, mas foi a publicação de Ingrid Daubechies [DAUBECHIES, 1988] que ascendeu o interesse pela *Wavelet* em aplicações nas áreas da engenharia, processamento de sinais, estatística e análise numérica.

Inicialmente, os pesquisadores usaram a termo em francês *"ondelette"* que significa onda pequena. Mais tarde, a palavra foi traduzida (ou neste caso transferida) para o inglês por uma substituição do termo francês *"onde"* pelo inglês "*wave*", formando assim a palavra *"Wavelet"*. Em português denomina-se a *Wavelet* pelo termo "ondaleta". Neste trabalho, será utilizado o termo em inglês por ser o mais utilizado na literatura.

O objetivo inicial das pesquisas sobre as *Wavelets* (na era moderna) era criar um conjunto de funções base e transformadas, as quais dão uma descrição (informação) eficiente e prática sobre a função ou sinal analisado. Se o sinal é representado como uma função do tempo (ex. sinal de voz), as *Wavelets* proporcionam uma localização eficiente em ambas as escalas de tempo e freqüência [GOWDY, 2000].

Outra idéia central é o conceito da análise de múltipla resolução (multiresolução) onde a decomposição do sinal é feita em termos de dois filtros, Aproximação e Detalhe. A decomposição em multiresolução permite separar as componentes do sinal de um modo superior a qualquer outra ferramenta ou método de análise, processamento ou compressão do sinal [BURRUS, 1998]. Devido a esta habilidade da *Wavelet* em decompor um determinado sinal em independentes escalas e aproximações, Burke chamou as *Wavelets* de "Microscópio Matemático" [BURKE, 1994]. Portanto, pode-se definir a *Wavelet* (ou ondaleta) como uma função capaz de decompor outras funções no domínio da freqüência, de forma a permitir a

análise destas funções em diferentes escalas de freqüência e de tempo. Esta decomposição de um sinal pela *Wavelet* é conhecida como Transformada *Wavelet* e tem suas variantes Contínuas e Discretas.

3.3.1 Escala e Translação

Na análise de Fourier, um sinal a ser analisado é decomposto por uma combinação linear de ondas senoidais de diferentes freqüências. O benefício desta técnica está no fato de que todas as freqüências contidas no sinal podem ser facilmente obtidas. Como a forma de onda senoidal tem um caráter global, a informação sobre as freqüências obtidas serão de caráter global também. Deste modo, se o sinal a ser analisado é estacionário ou periódico, a transformada de Fourier é totalmente suficiente para definir o espectro do sinal, pois as freqüências contidas em um sinal estacionário não mudam com o passar do tempo.

Entretanto, muitas vezes é preciso determinar o local onde as freqüências aparecem, neste caso será difícil extrair esta informação através da transformada de Fourier. Este é o caso dos sinais não estacionários (ex. sinal de voz). Aqui, principalmente, é que as *Wavelets* apresentam todo seu poder de processamento e análise [BURRUS, 1998].

A análise feita pela Transformada *Wavelet* (WT-*Wavelet Transform*) envolve a decomposição de um sinal em um conjunto de Aproximações e Detalhes que são, simplesmente, nada mais do que padrões em diferentes escalas e posições (translações) de uma determinada função "mãe" ou *Wavelet* mãe $\psi(x)$. Estes padrões são construídos a partir de uma função mãe $\psi(x)$ através da obtenção de diversas escalas $\psi(2^{-j}x)$ e de translações $\psi(2^{-j}x - k)$. O conceito de escala pode ser melhor visualizado com o exemplo a seguir. A Figura 3.5 apresenta duas ondas senoidais $f(x) = \sin x$ e $g(x) = sin2^3x$.

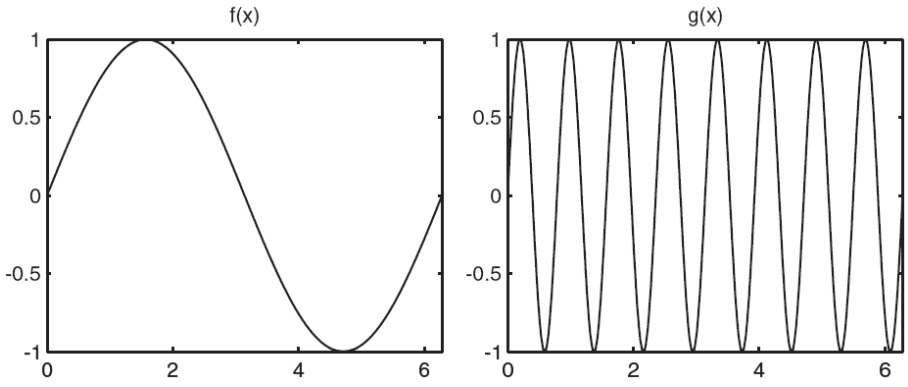

Figura 3.5: Duas ondas senoidais com diferentes freqüências (escalas).

Pode-se ver que $g(x)$ tem uma freqüência maior do que $f(x)$. Este conceito de freqüência, na teoria da *Wavelet,* pode ser repassado à idéia de escala. A organização do parâmetro escala está relacionada ao nível j, indicado por 2^{-j}. Assim, considerando que $\psi(x)$ é o padrão na escala "0" (escala inicial), então $\psi(2^{-j}x)$ é o padrão na $j_{ésima}$ escala. Pode-se também chamar $\psi(2^{-j}x)$ de resolução dada por 2^{j}. Deste modo, a resolução aumenta conforme a escala j diminui. Quanto maior a resolução, menores e mais finos são os detalhes que podem ser acessados.

Uma das características das *Wavelet* mãe $\psi(x)$ é o suporte compacto. Matematicamente, significa que uma determinada função $\psi(x)$ tem suporte compacto se $\psi(x)$ for igual a "0" fora do intervalo [a, b]. Em outras palavras, a *Wavelet* tem a habilidade de truncar o sinal em torno de uma localização especial. Um exemplo deste tipo de forma de onda (genérico) é apresentado na Figura 3.6.

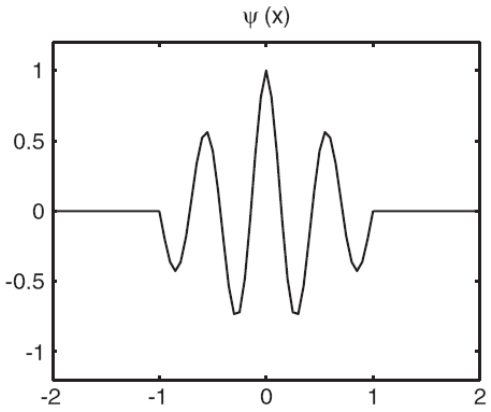

Figura 3.6: Modelo da forma de onda genérica de uma *Wavelet.*

Uma vez que a escala tem significado similar à freqüência, pode-se entender que para função original de uma *Wavelet* $\psi(x)$, a escala é responsável por medir ou analisar a informação da freqüência do sinal perto do tempo "0". Se o objetivo for encontrar ou analisar a informação da freqüência perto de um tempo k, basta apenas transladar a função em $\psi(2^{-j}x - k)$.

Geralmente, todo o sinal ou fenômeno físico com energia finita pode ser representado por uma função integrável quadrada, isto é, diz-se que um sinal $f(x)$ é de energia finita se a função do sinal satisfaz o quadrado integrável conforme a Equação (3.9).

$$\int_{-\infty}^{+\infty} |f(x)|^2 dx < \infty \qquad (3.9)$$

Matematicamente, as funções de quadrado integrável são denominadas por $L^2(\mathbb{R})$, ou seja, $f(x) \in L^2(\mathbb{R})$. O espaço $L^2(\mathbb{R})$ é particularmente importante ao processamento de sinais. Este é o espaço de todas as funções $f(x)$ que possuem uma integral bem definida do quadrado do módulo da função. O "L" significa a integral *Labesgue*[2], o "2" define a integral do quadrado do módulo da função e $\mathbb{R}$ expressa que a variável independente x pertence ao conjunto dos números reais.

3.3.2 A *Wavelet* Haar

Devido à simplicidade da sua forma de onda, utiliza-se nesta seção a *Wavelet* Haar para introduzir os conceitos básicos das *Wavelets*. Sendo $\mathbb{Z}$ o conjunto de números inteiros (positivos e negativos), isto é, $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$, Haar propôs em 1910 a função $\psi(x)$ descrita pela Equação (3.10).

$$\psi(x) = \begin{cases} -1 & se\ 0 \leq x \leq \frac{1}{2} \\ 1 & se\ \frac{1}{2} \leq x \leq 1 \\ 0 & outros\ valores \end{cases} \tag{3.10}$$

A Figura 3.7 apresenta função $\psi(x)$ da *Wavelet* de Haar.

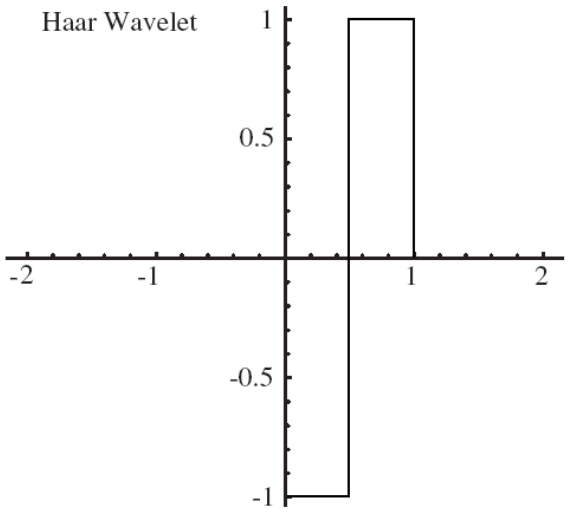

Figura 3.7: Função *Wavelet* de Haar.

[2] Na matemática, a integral de uma função não negativa pode ser considerada como a área entre o gráfico dessa função e o eixo x (abscissa). A integração de Lebesgue é uma construção matemática que estende a integral a uma classe maior de funções e igualmente estende os domínios em que estas funções podem ser definidas.

Para cada par de inteiros $j, k \in \mathbb{Z}$, Haar construiu um modelo padrão com escalas 2^j e translação k, dada pela Equação (3.11).

$$\psi_{j,k}(x) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{x - 2^j k}{2^j}\right) \tag{3.11}$$

Em seguida Haar formou um conjunto $\mathcal{H}$ destes padrões conforme a Equação (3.12) a seguir.

$$\mathcal{H} = \{\psi_{j,k}(x) \mid j, k = \cdots, -2, -1, 0, 1, 2, \cdots\} \tag{3.12}$$

Assim, genericamente, para qualquer valor inteiro (positivo ou negativo) de j e k, o padrao $\psi_{j,k}(x)$ é uma função com os valores e intervalos específicos, dados pela Equação (3.13) a seguir,

$$\psi_{j,k}(x) \begin{cases} -\frac{1}{\sqrt{2^j}} & para\ o\ intervalo\ \left[2^j k\ , 2^j k + 2^{j-1}\right] \\ \frac{1}{\sqrt{2^j}} & para\ o\ intervalo\ \left[2^j k + 2^{j-1}, 2^j (k+1)\right] \\ 0 & fora\ do\ intervalo\ \left[2^j k, 2^j (k+1)\right] \end{cases} \tag{3.13}$$

onde, j é chamado de nível de escala ou simplesmente nível e k é chamado de localização.

A Figura 3.8 apresenta quatro exemplos de diferentes escalas e localizações da *Wavelet* Haar, as quais foram obtidas a partir da Equação 3.13, sendo os valores:

$$\psi_{0,-2}(x) = \begin{cases} -1 & -2 \le x < -\frac{3}{2} \\ 1 & -\frac{3}{2} \le x < -1 \\ 0 & outros\ valores \end{cases} \qquad \psi_{1,1}(x) = \begin{cases} -\frac{1}{\sqrt{2}} & 2 \le x < 3 \\ \frac{1}{\sqrt{2}} & 3 \le x < 4 \\ 0 & outros\ valores \end{cases}$$

$$\psi_{-2,1}(x) = \begin{cases} -2 & \frac{1}{4} \le x < \frac{3}{8} \\ 2 & -\frac{3}{2} \le x < \frac{1}{2} \\ 0 & outros\ valores \end{cases} \qquad \psi_{3,-1}(x) = \begin{cases} -\frac{1}{2\sqrt{2}} & -8 \le x < -4 \\ \frac{1}{2\sqrt{2}} & -4 \le x < 0 \\ 0 & outros\ valores \end{cases}$$

Os exemplos da Figura 3.8 mostram que quanto maior for o nível de j, mais larga ou maior é o espaçamento da onda $\psi_{j,k}(x)$, e quanto menor o nível de j mais estreita é a onda $\psi_{j,k}(x)$. Portanto, um valor pequeno do nível j corresponde a um fina resolução em $\psi_{j,k}(x)$ e, vice-versa, uma resolução grosseira de $\psi_{j,k}(x)$ está associada com um valor alto do nível

j. A Figura 3.8 demonstra também que os valores de k são responsáveis pelo deslocamento ou translação de $\psi_{j,k}(x)$.

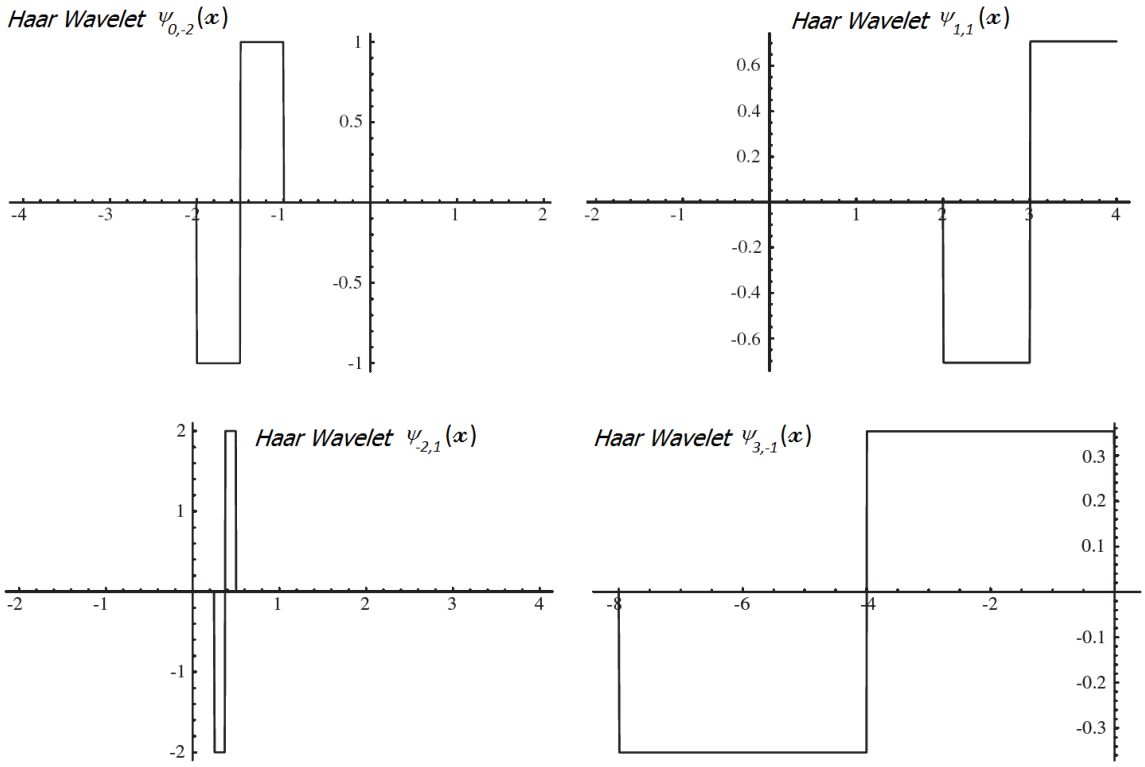

Figura 3.8: Padrões da *Wavelet* Haar para diferentes escalas e localizações.

Outra importante característica das *Wavelets* é a ortogonalidade. Considerando inicialmente um conjunto de funções com energia finita $L^2(\mathbb{R})$, pode-se definir o produto interno de duas funções com energia finita, $f(x)$ $g(x)$ d$L^2(\mathbb{R})$ como sendo:

$$\langle f(x), g(x) \rangle = \int_{-\infty}^{+\infty} f(x)g(x)dx \tag{3.14}$$

O produto interno $\langle f(x), g(x) \rangle$ pode ser usado para medir a similaridade de dois sinais $f(x)$ $g(x)$, ou seja, quanto maior a magnitude do produto interno maior é a similaridade entre eles.

Se o produto interno de duas funções é igual a 0, então estas função são ditas ortogonais ou não similares. O produto interno de uma função $f(x)$ com ela mesma é chamada de energia da função. Deste modo, com a computação da integral da Equação 3.14, pode-se provar que cada padrão Haar $\psi_{j,k}(x)$, definido pela Equação 3.11, tem energia igual a 1 e o produto interno de quaisquer dois diferentes padrões é igual a 0.

Com estas duas propriedades, pode-se dizer que o conjunto $\mathcal{H}$ de funções Haar consiste em uma base ortonormal de $L^2(\mathbb{R})$ chamada de *Wavelet* base de Haar. Assim, a *Wavelet* base de Haar favorece a prova de que qualquer função $f(x)$ de energia finita pode ser decomposta em uma combinação linear dada pela Equação (3.15).

$$f(x) = \sum_{j=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \langle f(x), \psi_{j,k}(x) \rangle \, \psi_{j,k}(x) \tag{3.15}$$

A Equação (3.15) permite dizer que a família Haar é completa e suficiente para representar qualquer sinal de energia finita. Deste modo, pode-se dizer que a expansão dos coeficientes $\langle f(x), \psi_{j,k}(x) \rangle$ de $\psi_{j,k}(x)$, na Equação (3.15), fornece a informação de quanto o sinal $f(x)$ é similar ao padrão $\psi_{j,k}(x)$, ou em outras palavras, pode-se determinar através de $\langle f(x), \psi_{j,k}(x) \rangle$ o quanto do padrão $\psi_{j,k}(x)$, o sinal $f(x)$ possui.

Outra propriedade importante da função *Wavelet* é o momento *Vanish*, ou momento nulo, dada pela Equação (3.16). Uma função *Wavelet* $\psi(x)$ possui "p" momentos nulos se,

$$\int_{-\infty}^{+\infty} x^k \psi(x) dx = 0 \qquad para \; 0 \leq k < p \tag{3.16}$$

Quanto mais momentos nulos ("p" grande) uma *Wavelet* possuir, menores serão os coeficientes da *Wavelet* correspondentes às partes de $f(x)$ que são suaves, ou seja, os coeficientes de *Wavelet* serão apreciáveis onde $f(x)$ não for suave, o que permite usar *Wavelets* para detectar singularidades de $f(x)$.

3.3.3 Transformada Contínua da *Wavelet*

Sendo uma função $f(t)$ de uma variável contínua no domínio do tempo, a Transformada Contínua da *Wavelet* é definida pela Equação (3.17):

$$\Upsilon(a,b) = \int_{-\infty}^{+\infty} f(t) \, \psi_{a,b}(t) dt \tag{3.17}$$

onde $\psi_{a,b}$ é a função base *Wavelet* ou *Wavelet* mãe *e* os parâmetros a e b são os fatores de escala e translação respectivamente. A Transformada Inversa Contínua é dada por:

$$f(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Upsilon(a,b) \, \psi_{a,b}(t) da db \tag{3.18}$$

A função $\psi_{a,b}$, (*Wavelet* mãe) é definida através das propriedades de escala e translação conforme a Equação (3.18).

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{3.19}$$

onde *a, b ∈ R, a ≠0.* O fator de multiplicação $\frac{1}{\sqrt{a}}$ é utilizado para a normalização da energia através das diferentes escalas [BURRUS, 1998]. Se o valor de a for maior do que zero então a operação corresponde à dilatação. Por outro lado, se o valor de a for menor que zero então a operação corresponde à contração de $f(t)$.

Pode-se ver claramente a similaridade entre as Equações 3.19 e 3.11 que definem a *Wavelet* mãe genérica e a Haar, respectivamente. Vale salientar aqui, que as variáveis de escala e translação no modo contínuo são representadas por $[a\ b]$, enquanto que no modo discreto são representadas por $[j\ k]$.

3.3.4 Transformada Discreta da *Wavelet*

A meta agora é gerar um conjunto de funções de expansão tal que, qualquer sinal em $L^2(\mathbb{R})$ possa ser representado pelas séries da Equação (3.19) dada por:

$$f(t) = \sum_{j=0}^{N}\sum_{k=0}^{N} a_{j,k}2^{\frac{j}{2}}\psi\left(2^j t - k\right) \tag{3.20}$$

ou usando a Equação (3.20),

$$f(t) = \sum_{j=0}^{N}\sum_{k=0}^{N} a_{j,k}\psi(t) \tag{3.21}$$

onde o conjunto de coeficientes (bidimensional) $a_{j,k}$ são chamados de Transformada Discreta da *Wavelet* (DWT – *Discrete Wavelet Transform*) de $f(t)$.

Na prática, a DWT de um determinado sinal (x) é calculada passando este sinal através de uma série de filtros.

Inicialmente, as amostras são passadas através de um filtro passa-baixa com resposta ao impulso (g) resultando em uma convolução dos dois, dada pela Equação (3.21):

$$y[n] = (x * g)[n] = \sum_{k=-\infty}^{+\infty} x[k]g[n-k] \tag{3.22}$$

O sinal é decomposto também, simultaneamente, usando um filtro passa-alta (h). As saídas dos filtros, passa-baixa e alta, fornecem os coeficientes chamados de Aproximação e Detalhe, respectivamente. É importante que os dois filtros estejam relacionados entre si, e quando este fato ocorre, o filtro é chamado de filtro em espelho de quadratura.

Dado que a metade das freqüências do sinal foi removida, a metade das amostras pode ser rejeitada de acordo com a regra ou Teorema de Nyquist[3]. As saídas dos filtros são então amostradas em uma forma de divisão por 2. As Equações (3.22) e (3.23) apresentam a forma dos dois filtros passa-baixa e passa-alta:

$$y_{baixa}[n] = \sum_{k=-\infty}^{+\infty} x[k]g[2n-k] \tag{3.23}$$

$$y_{alta}[n] = \sum_{k=-\infty}^{+\infty} x[k]h[2n-k] \tag{3.24}$$

Esta decomposição divide pela metade a resolução no tempo devido ao fato de que somente a metade da saída de cada filtro caracteriza o sinal.

Entretanto, cada saída tem a metade da banda de freqüência da entrada, assim a resolução da freqüência foi dobrada. A Figura 3.9 apresenta o digrama de blocos dos filtros descritos acima. O operador $(\downarrow 2)$ é o operador de sub-amostragem (do inglês *downsampling*). Este operador aplicado a uma função discreta (uma seqüência) reduz o seu número de elementos pela metade, recuperando apenas os elementos em posições pares.

Um aumento na resolução da freqüência pode ser obtido através repetição da decomposição da Figura 3.8 para os próximos níveis.

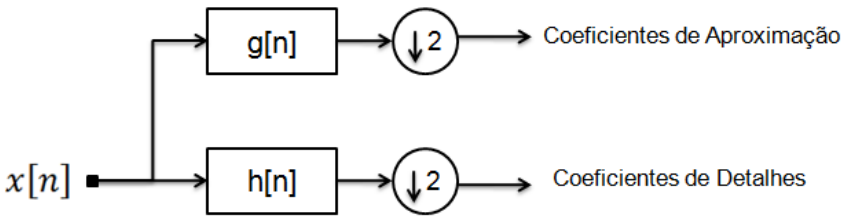

Figura 3.9: Diagrama de blocos da decomposição DWT.

[3] O teorema da amostragem de Nyquist-Shannon é fundamental no campo da teoria de informação. Essencialmente, significa que um sinal analógico que foi digitalizado pode ser perfeitamente reconstruído se a taxa de amostragem for no minimo o dobro da freqüência mais elevada do sinal original.

A Figura 3.10 apresenta esta árvore binária com três níveis de decomposição a qual é conhecida como banco de filtros.

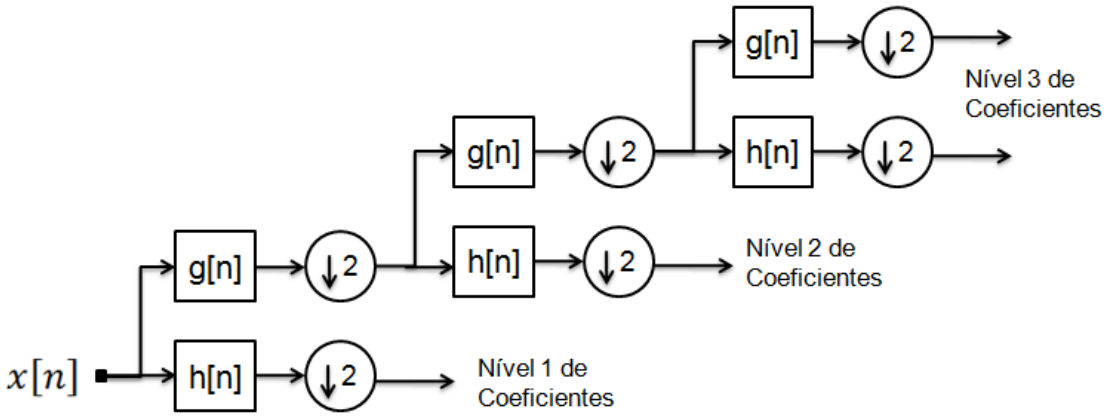

Figura 3.10: Um banco de filtro com 3 níveis.

Deste modo, os coeficientes da aproximação são novamente decompostos por filtros passa-baixa e alta. Isto é representado através de uma árvore binária, onde os ramos representam um sub-espaço com uma diferente localização de tempo-freqüência.

A decomposição (Figura 3.10) feita pela DWT é também chamada de análise de multiresolução, ou análise de múltipla resolução, a qual é a base para a construção da *Wavelet Packet*.

3.4 Transformada *Wavelet Packet*

A Transformada *Wavelet Packet* (WPT- *Wavelet Packet Transform*) é a principal ferramenta utilizada para obtenção dos descritores do sinal de voz neste trabalho. Diversos trabalhos focados no reconhecimento de voz utilizam a *Wavelet Packet* como descritor. Alguns destes trabalhos são citados e comentados na seção final deste capítulo.

A WPT foi proposta por Ronald Coifman [COIFMAN, 1992] para permitir uma resolução de freqüência fina e ajustável para as altas freqüências.

A *Wavelet Packet* também gera uma rica estrutura que permite a adaptação a sinais particulares ou classes de sinais como a voz [JIANG, 2003].

O custo desta rica estrutura é a complexidade computacional de O(N log(N)), similar à FFT, em contraste à clássica transformada *Wavelet* na qual é O(N). A *Wavelet Packet* é uma ampliação do conceito da Transformada *Wavelet* Discreta, na qual a resolução

tempo/freqüência pode ser escolhida de acordo com o sinal. Isto é realizado dentro dos limites do princípio de incerteza de Heisenberg[4] [BURRUS, 1998].

Na análise *Wavelet* Discreta, o sinal é dividido em coeficientes de aproximação e detalhes (Figura 3.10) e somente os coeficientes de aproximação são divididos novamente. Por outro lado, na análise da *Wavelet Packet,* tanto os coeficientes de aproximação e detalhes podem ser decompostos em qualquer nível. Deste modo, cada nível terá 2^n conjuntos de coeficientes ou bandas. Desta forma, tem-se o sinal representado em bandas de freqüências com diferentes resoluções. O resultado produz o que é chamado de árvore de decomposição *Wavelet Packet* [COIFMAN, 1992].

A Figura 3.11 apresenta a árvore de decomposição de um sinal obtida a partir da WPT com três níveis de decomposição, onde $g[n]$ e h[n] são os filtros passa-baixa e alta respectivamente e $x(n)$ é o sinal a ser decomposto.

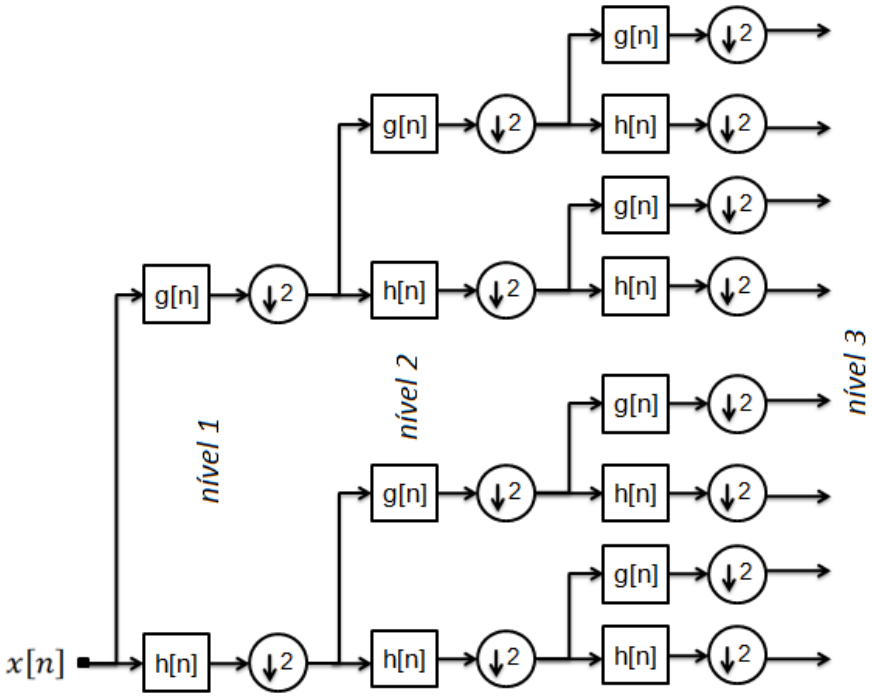

Figura 3.11: Decomposição *Wavelet Packet* com 3 níveis.

Comparando com a Figura 3.10, que representa a decomposição da DWT, pode-se ver que neste caso ambos os filtros passa-baixa e alta são decompostos a cada novo nível.

[4] O princípio da incerteza de Werner Heisenberg, estabelece que é impossível conhecer simultaneamente a posição e a energia de uma partícula tal como o elétron. O princípio da incerteza pode ser assim interpretado: quanto mais de perto tentamos olhar uma partícula diminuta, tanto mais difusa se torna a visão da mesma.

Do ponto de vista do tamanho do descritor obtido, onde se deseja que a dimensão do descritor seja a menor possível, a WPT não produz o melhor resultado, pois a quantidade de bases ou bandas da *Wavelet Packet* aumenta em uma potência de 2 a cada novo nível. Por exemplo: Se um sinal for decomposto em 7 níveis, somente o último nível terá $2^7 = 128$ bandas e ao total serão 254 bandas se todos os níveis forem contabilizados. Este problema é contornado neste trabalho através da utilização de uma escala aproximada, a escala mel. A utilização desta escala reduz sensivelmente a quantidade de bandas utilizadas e conseqüentemente a dimensão do descritor.

A Figura 3.12 apresenta a mesma decomposição da Figura 3.11, mas agora na posição vertical. Esta representação dos níveis e bandas da *Wavelet Packet* permite uma melhor visualização quando existem quatro níveis ou mais (caso deste trabalho).

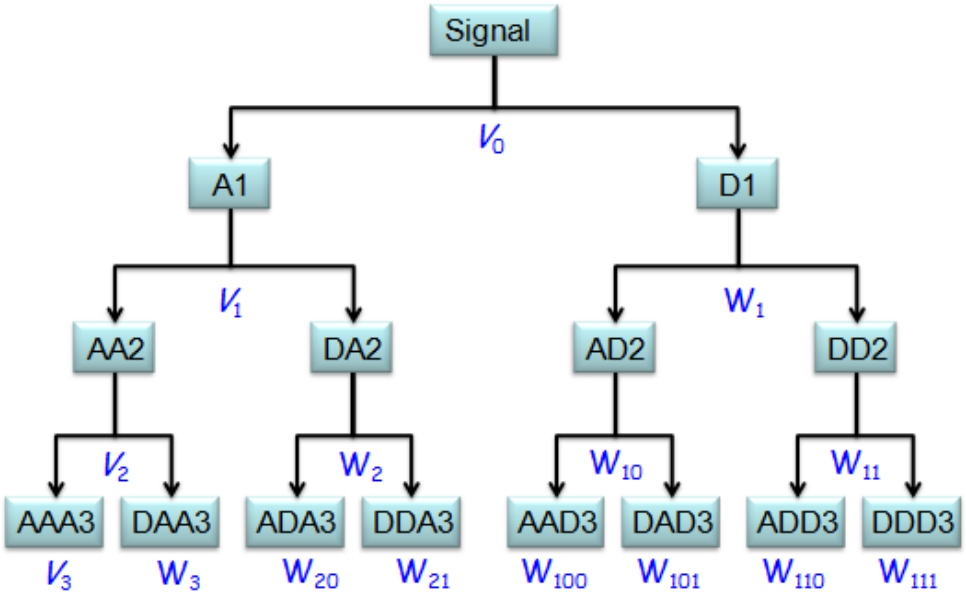

Figura 3.12: Árvore binária da transformada *Wavelet Packet* com 3 níveis.

As bandas de Aproximação são rotuladas pela letra "A", enquanto que as bandas de Detalhe pela letra "D". No nível 1, por exemplo, tem-se A1 e D1 como aproximação e detalhe, respectivamente. O nível 2 possui quatro bandas, sendo duas resultantes da decomposição de A1 (AA2 e DA2) e duas resultantes de D1 (AD2 e DD2). Este processo pode ser repetido até o nível de resolução desejado, conforme o sinal a ser analisado. Os espaços de cada nível são denotados por *V* e as subdivisões são denotadas por *W,* assim,

$$V_3 \subset V_2 \subset V_1 \subset V_0$$

A Figura 3.13 apresenta o espaço de decomposição referente aos três níveis da *Wavelet Packet*.

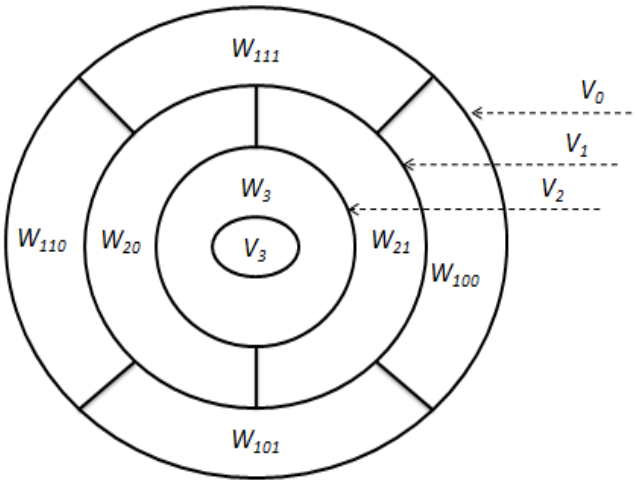

Figura 3.13: Sub-espaços de decomposição da *Wavelet Packet* com três níveis.

Já a Figura 3.14 apresenta a resposta em freqüência das bandas referentes aos espaços e subespaços definidos por esta *Wavelet Packet* com três níveis.

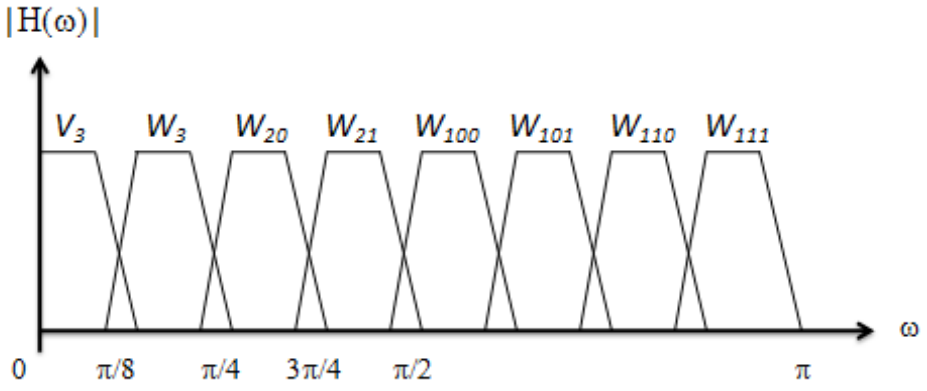

Figura 3.14: Resposta em freqüência do banco de filtros da *Wavelet Packet* com 3 níveis.

A Figura 3.15 apresenta uma comparação gráfica entre as decomposições da Transformada de Fourier, da STFT (*Short Time Fourier Transform*) e a Transformada *Wavelet Packet*. Pode-se ver claramente que a STFT faz a análise do sinal no tempo e na freqüência mas, devido ao tamanho da janela ser sempre fixa, haverá sempre o seguinte problema: Para se ter uma boa resolução em baixas freqüências deve-se utilizar um tamanho de janela pequeno. No entanto, isto acarretará uma perda na análise em altas freqüências, o mesmo ocorrerá no caso inverso. A análise da WPT permite o uso de intervalos de tempo longos,

onde se deseja uma informação de baixa freqüência mais precisa, e regiões mais curtas no tempo onde se quer a informação de alta freqüência.

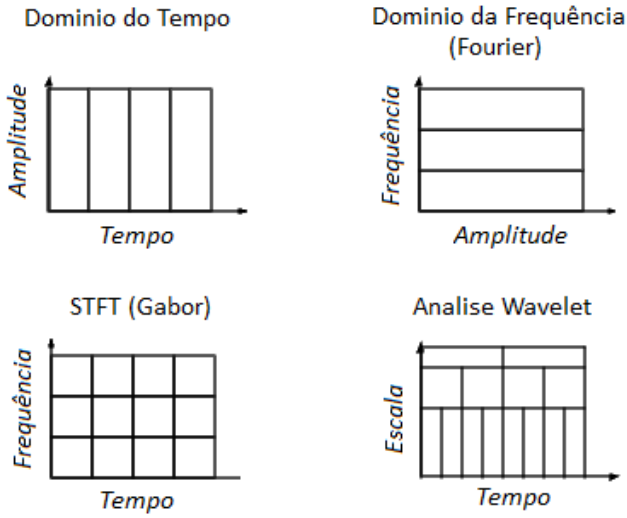

Figura 3.15: Comparação gráfica entre as análises: transformada de Fourier, STFT e WPT.

3.4.1 Descritores da Voz Utilizando a *Wavelet Packet* e a Escala Mel

Conforme visto na Figura 3.12, o sinal é decomposto tanto em alta como em baixa freqüência pela *Wavelet Packet*. Deste modo, a cada novo nível (n) de decomposição, 2^n bandas de decomposição são acrescentadas. O nível 1 possui 2 bandas, o nível 2 possui 4 bandas, o nível 3 possui 8 bandas e assim por diante. Portanto, o descritor obtido pela *Wavelet Packet* possui a característica de alta dimensionalidade. Com objetivo de reduzir a dimensionalidade Gowdy e Tufekci [GOWDY, 2000], Farooq e Datta [FAROOQ, 2003] e Chan *et al* [CHAN, 2001] utilizaram bandas da *Wavelet Packet* escolhidas com auxílio da escala Mel.

Devido ao fato de que a *Wavelet Packet* divide o sinal em 2^n bandas (decomposição binária), suas bandas não são espaçadas exatamente igual à escala Mel (linear e exponencial). Neste trabalho, 26 bandas da *WP* foram escolhidas com o intuito de representar aproximadamente a escala Mel (faixa da voz, Figura 3.3).

A escala Mel possui as 10 bandas iniciais com largura de 100 Hz espaçadas linearmente até a freqüência de 1.000 Hz. Como o sinal de voz, neste trabalho, é amostrado com uma freqüência de 22.050 Hz, então, no oitavo nível da *Wavelet Packet,* cada banda deverá ter uma frequência de 86 Hz. Deste modo, as 12 bandas iniciais do oitavo nível totalizam 1032 Hz. As bandas subseqüentes foram escolhidas de forma a possuírem uma largura de banda aproximadamente igual a escala Mel, conforme à Tabela 3.1.

Tabela 3.1: Bandas da WP escolhidas para representar o sinal, próximas a escala Mel.

Bandas	*Wavelet Packet* (Hz)		Bandas	Escala Mel (Hz)	
	Faixa Freq.	Largura Banda		Faixa Freq.	Largura Banda
1	86	86	1	100	100
2	172	86	2	200	100
3	258	86	3	300	100
4	344	86	4	400	100
5	430	86	5	500	100
6	516	86	6	600	100
7	602	86	7	700	100
8	688	86	8	800	100
9	774	86	9	900	100
10	860	86	10	1000	100
11	946	86			
12	1032	86			
13	1206	172	11	1149	160
14	1378	172	12	1320	184
15	1550	172	13	1516	211
16	1722	172	14	1741	242
17	2067	345	1'5	2000	278
18	2412	345	16	2297	320
19	2757	345	17	2639	367
20	3102	345	18	3031	422
21	3447	345	19	3482	484
22	4134	689	20	4000	556
23	4823	689	21	4595	639
24	5512	689	22	5278	734
25	6201	689	23	6063	843
26	6890	689	24	6964	969
27	8269	1378	25	8000	1113

Deste modo, as bandas escolhidas foram:

- Nível 8: 12 bandas de 86Hz (8-0 a 8-11); Faixa de 0 - 1032Hz.
- Nível 7: 04 bandas de 172Hz (7-6 a 7-11); Faixa de 1032 - 1720Hz.
- Nível 6: 05 bandas de 344Hz (6-6 a 6-11); Faixa de 1720 - 3440Hz.
- Nível 5: 05 bandas de 689Hz (5-6 a 5-9); Faixa de 3440 - 6890Hz.
- Nível 4: 01 bandas de 1378Hz (4-5); Faixa de 6890 - 8268Hz.

Pode-se ver, que ao total foram escolhidas 27 bandas da Wavelet Packet. A banda 1 foi eliminada pois pode conter as freqüências do sinal elétrico da rede de 60Hz. Deste modo, restaram 26 bandas conforme a Figura 3.16.

Para cada banda foi calculada a energia [SARIKAYA, 1998], [ROCHA, 2005]. Desde modo, cada banda tornou-se um coeficiente do descritor. Assim a decomposição feita pela *Wavelet Packet,* em conjunto com a escala Mel, forma um descritor com 26 componentes ou coeficientes que representam a energia de cada banda escolhida.

No desenvolvimento deste trabalho, inúmeras configurações com diferentes níveis da *Wavelet Packet* foram testadas. A Figura 3.16 mostra a configuração que apresentou os melhores resultados, e portanto, foi utilizada nos testes práticos deste trabalho. Oito níveis de decomposição são necessários para se obter as bandas desejadas. Esta estrutura com 8 níveis possui 560 bandas. Este é um dos fatores que ainda limita a utilização da *Wavelet Packet* em sistemas de tempo real, pois o tempo de processamento destes oito níveis (560 bandas) torna-se demasiadamente grande. Para tentar reduzir o tamanho do descritor e diminuir o tempo de processamento, mas sem afetar o desempenho, é que seu utilizou as 26 bandas escolhidas através da escala Mel aproximada (Tabela 3.1).

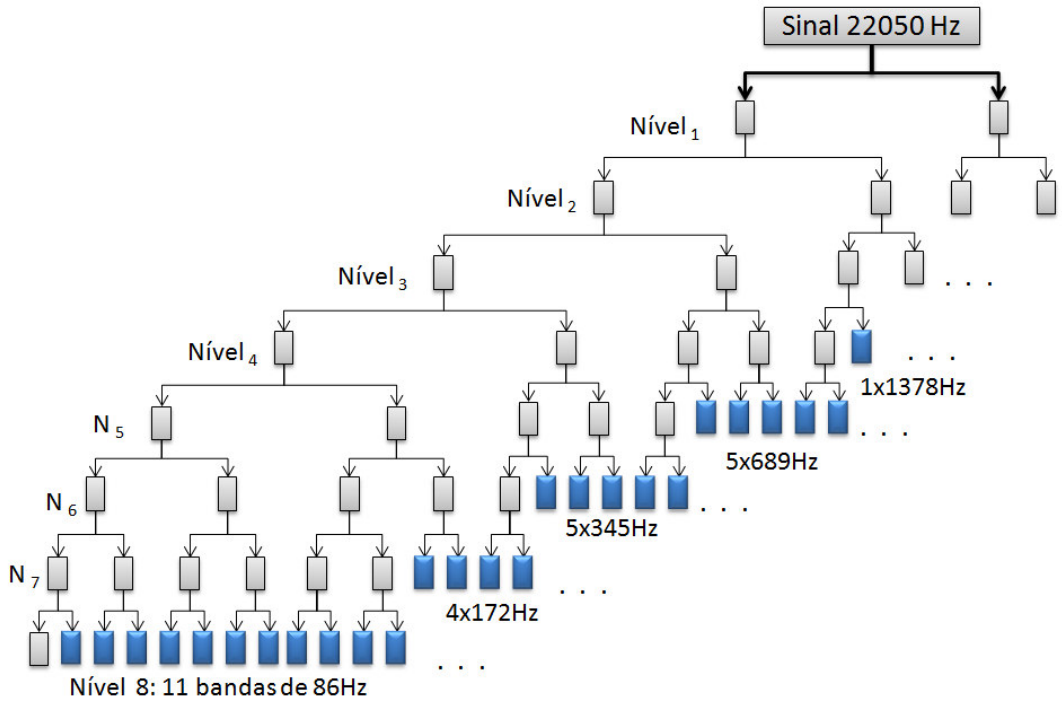

Figura 3.16: Estrutura para obtenção dos descritores através da *Wavelet Packet.*

3.5 Tipos de *Wavelet* Mãe

Dependendo das características do sinal a ser analisado, uma diferente *Wavelet* mãe pode ser utilizada. As *Wavelets* apresentadas na Figura 3.17 (a) são devidas a Ingrid

Daubechies. Estas *Wavelets* mãe são ortogonais, de suporte compacto, mas assimétricas. Os nomes das *Wavelets* da família *Daubechies* são descritos por "db_N", onde N é a ordem da *Wavelet.* A *Wavelet* db1 não é apresentada pois é da mesma forma que a *Wavelet* de Haar.

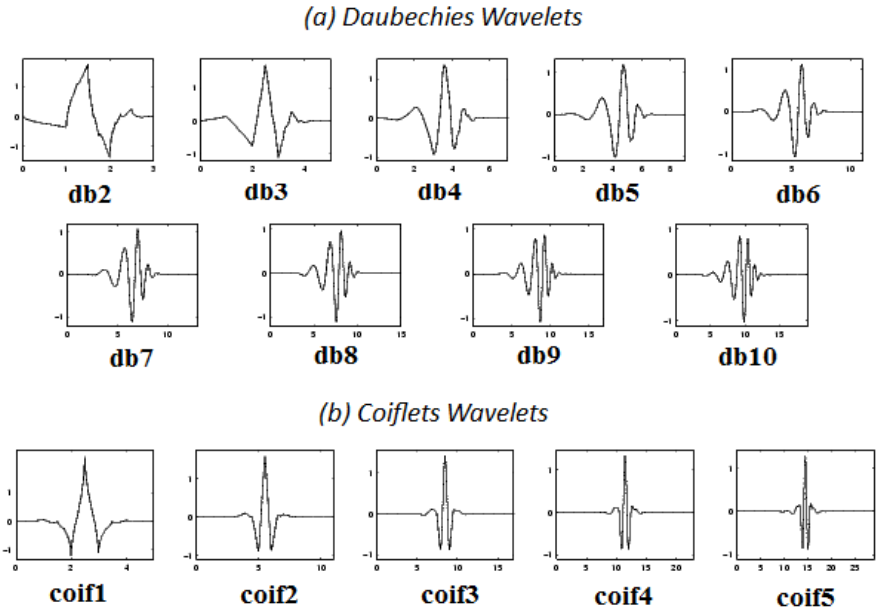

Figura 3.17: Exemplos de *Wavelet* mãe: Daubechies e Coiflets.

As funções *Coiflet* $\psi(x)$ apresentadas na Figura 3.17 (b), também foram construídas por Ingrid *Daubechies*, mas agora a pedido de R. Coifman [COIFMAN, 1992]. As *Coiflets* são mais simétricas do que as *Daubechies* e são denominadas pelos índices "$coif_N$".

Já as *Symlets* são *Wavelets* quase simétricas que também foram propostas *Daubechies* como modificações à família do "db". As propriedades das famílias "db" e "sym" são similares. A Figura 3.18 apresenta as *Symlets* denominadas pelos índices "sym_N".

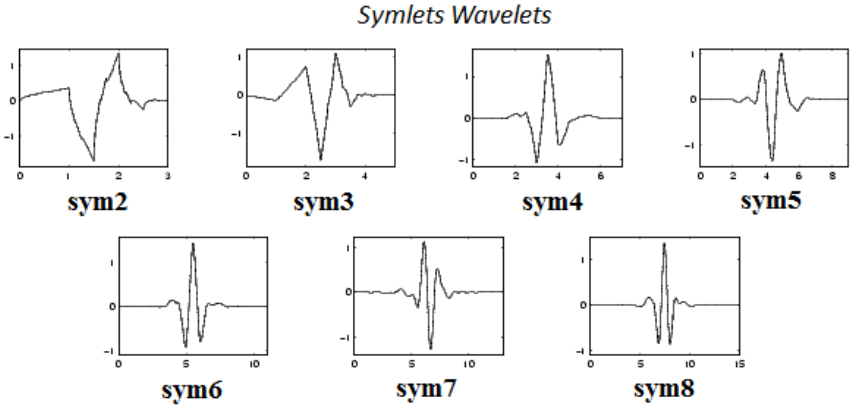

Figura 3.18: Exemplos de *Wavelet* mãe: Symlets *Wavelets.*

A Figura 3.19 (a, b e c) apresenta as *Wavelets Morlet, Mexican Hat* (MexH) e Meyer, respectivamente. Apesar das três possuírem características semelhantes como regularidade e simetria, somente a Meyer possui transformada discreta.

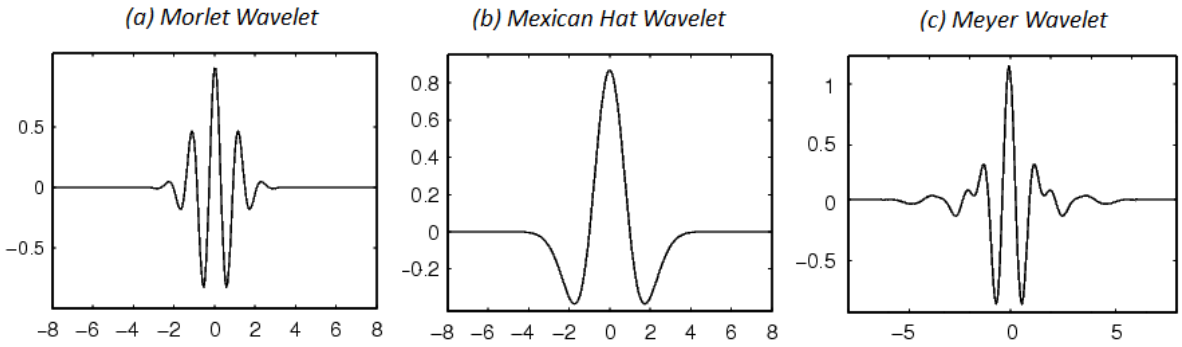

Figura 3.19: Exemplos de *Wavelet* mãe: *Morlet*, *Mexican Hat* e *Meyer.*

A Tabela 3.2 apresenta um resumo das famílias de *Wavelets* com as respectivas características de cada uma delas. Existem ainda outros tipos de *Wavelet* mãe tais como: Biortonormais e as Complexas.

Tabela 3.2: Famílias de *Wavelets* com suas respectivas propriedades.

Propriedades	Morlet	MexH	Meyer	Haar	db_N	Sym_N	$coif_N$	Biort
Regularidade	•	•	•					
Regularidade Arbitrária					•	•	•	•
Suporte compacto ortogonal				•	•	•	•	
Suporte compacto biortogonal								•
Simétrica	•	•	•	•				•
Assimétrica					•			
Quase Simétrica						•	•	
Momentos nulos arbitrários					•	•	•	•
Momentos nulos (função escala)							•	
Existência da função escala			•	•	•	•	•	•
Análise Ortogonal			•	•	•	•	•	
Análise Biortogonal								•
Reconstrução exata		•	•	•	•	•	•	•
Filtros FIR				•	•	•	•	•
Transformada Contínua	•	•	•	•	•	•	•	•
Transformada Discreta			•	•	•	•	•	•
Algoritmo para cálculo rápido				•	•	•	•	•
Expressão explícita	•	•		•				

Além disso, toda a análise em uma dimensão (1D) pode ser expandida para duas dimensões (2D), como no caso da utilização das *Wavelets* no processamento de imagens.

Para o caso do reconhecimento de voz, como pode ser visto na seção a seguir, as *Wavelets* Daubechies 5, Meyer e Coiflet 5 são as mais utilizadas.

3.6 Estado da Arte: Descritores do Sinal de Voz

Muitas pesquisas foram realizadas na busca do melhor descritor para o sinal de voz. Pode-se dizer que, apesar da grande utilização dos descritores MFCC na literatura, a resposta para esta questão ainda está em aberto, pois nos últimos anos muitos trabalhos demonstraram que as *Wavelets* podem ir além. Uma boa perspectiva histórica sobre as *Wavelets* pode ser encontrada no livro de Meyer [MEYER, 1993]. Algumas conclusões podem ser tiradas das pesquisas realizadas nas últimas três décadas e, deste modo, justificar a utilização das ferramentas descritas neste capítulo no sistema de reconhecimento de voz desenvolvido neste trabalho.

O livro de Pedro Morettin [MORETTIN, 1999] faz uma comparação da análise de sinais usando a transformada de Fourier e as *Wavelets*, bem como apresenta uma descrição matemática completa destas transformadas.

Os trabalhos de Tan *et al* [TAN, 1996], Long e Datta [LONG, 1996], Sarikaya e Hansen [SARIKAYA, 2000] e Kim *et al* [KIM, 2000] demonstraram que a *Wavelet* é apropriada para análise da voz. Além disso, os resultados obtidos nestes trabalhos mostraram que a transformada *Wavelet* apresenta um desempenho no reconhecimento melhor quando comparada com o descritor tradicional MFCC.

Wendt e Petropulu [WENDT, 1996], Hossain *et al* [HOSSAIN, 1999] e Yang *et al* [YANG, 2001] concluíram em seus trabalhos que as *Wavelets* tem comportamento parecido com o ouvido humano e por este motivo podem descrever mais precisamente os sinais de voz. Os trabalhos citados a seguir estão em ordem cronológica.

Davis e Mermelstein [DAVIS, 1980] fizeram um importante trabalho de comparação do desempenho de descritores paramétricos no reconhecimento de voz contínua. O trabalho tinha como base o reconhecimento de monossílabas com dependência do locutor. Os autores compararam o desempenho de 5 (cinco) descritores: MFCC, LFCC (*Linear Frequency Cepstrum Coefficients*), LPC, RC (*Reflection Coefficients*) e LPCC (*Linear Predictive Cepstrum Coefficients*). O objetivo era reconhecer sílabas simples formadas por CVC (consoante-vogal-

consoante). Em todos os experimentos os coeficientes MFCC apresentaram melhores resultados.

Günther Ruske apresentou um estudo do reconhecimento de sílabas da língua alemã [RUSKE, 1982]. O autor utilizou 14 coeficientes extraídos do LPC (*Linear Prective Coding*). Como resultado, obteve-se em média 59,4% de acerto no reconhecimento, sendo um bom indicativo de que os coeficientes LPC não representam bem as palavras, mas sim o locutor.

Davenport e Garudadri [DAVENPORT, 1991] utilizaram a *Wavelet* para extrair características acústico-fonéticas das palavras. Foram utilizadas 2 sub-bandas da *Wavelet* tipo Daubechies, totalizando 4 bandas de freqüências: 0-500Hz, 500-1kHz, 1kHz-2kHz e 2kHz-4kHz. Em cada banda foi extraída a média da energia e a taxa de mudança da energia, totalizando 8 valores para cada janela de 5ms do sinal. Foram obtidos em média 84% de acerto no reconhecimento para todos os fonemas. O destaque foi o reconhecimento das vogais que obteve 95,7% de acerto.

Timothy Anderson [ANDERSON, 1991] fez a comparação entre dois modelos de descritores: o primeiro foi baseado no modelo auditivo (biológico) ou acústico desenvolvido por Karen Payton [PAYTON, 1985] e o segundo tem como base a análise espectral DTF (*Discrete Fourier Transform*). O autor também comparou o desempenho de dois classificadores o SOM e *Kmeans.* Os resultados para este dois modelos de classificação foram semelhantes. Já para os descritores, o autor concluiu que os descritores baseados no modelo acústico do ouvido humano têm melhor desempenho do que os descritores espectrais, mas são muito difíceis de ser obter.

A publicação de Priebe e Baugh [PRIEBE, 1994] é um estudo muito importante sobre o uso da *Wavelet* como descritor de sinais. Além de fazer uma descrição matemática sucinta da *Wavelet,* os autores comparam seu desempenho teórico com a STFT (*Short-Time Fourier Transform*). A *Wavelet* apresentou melhor desempenho do que a transformada de Fourier quando aplicada em sinais.

Malbos *et al* [MALBOS, 1994] realizaram o reconhecimento das consoantes oclusivas [p, k, t, b, g, d] da língua francesa utilizando a transformada *Wavelet*. Os autores propuseram um método baseado na correlação da *Wavelet* com o próprio sinal. Foram obtidos bons resultados na classificação das consoantes oclusivas não vozeadas [p, k e t], alcançando 95,5% de reconhecimento.

Long e Datta [LONG, 1996] realizaram um estudo que procurava encontrar qual a melhor *Wavelet* que descreve as características fonéticas de um dado fonema. Os autores determinaram que a *Wavelet* do tipo Daubechies representa melhor os fonemas fricativos, devido a sua estrutura ter a forma parecida com a de um sinal de ruído. Já a *Wavelet* do tipo Morlet é mais apropriada para representar os fonemas vozeados (que usam as cordas vocais), pois estes fonemas têm comportamento quase-periódico, tal como esta *Wavelet.*

Le-Tien Thuong [THUONG, 1997] fez um interessante estudo de comparação do desempenho de diferentes *Wavelet* mãe sendo: Morlet, Spline, Mexican-Hat e Morlet. O autor concluiu que a *Wavelet* que melhor representa o sinal de voz é a Morlet.

O trabalho de Sarikaya e Gowdy [SARIKAYA, 1998] testou quatro novos descritores baseados nas sub-bandas da *Wavelet*. Estes novos descritores foram comparados com o tradicional MFCC na classificação de voz sob condições de "stress", ou seja, sinais de voz gravados em condições de "raiva", "rapidez", "lentidão", "questionamento" etc. Os autores utilizaram a *Wavelet* mãe do tipo Daubechies através da *Wavelet Packet Transform*. Os resultados revelaram que o melhor descritor foi o baseado na energia de cada sub-banda. O desempenho deste descritor superou todos os outros inclusive o tradicional MFCC.

Lukasik [LUKASIK, 2000] utilizou três métodos para extração de características do sinal de voz utilizando a WPT (*Wavelet Packet Transform*), com o objetivo de efetuar a classificação das consoantes oclusivas /p/, /t/ e /k/. O autor utilizou como classificador a rede neural MLP e testou a desempenho de várias *Wavelets* mãe para cada método. O autor concluiu que em todos os testes a WPT obteve melhores resultados do que os coeficientes cepstrais de Fourier.

O trabalho de Sarikaya e Hansen [SARIKAYA, 2000] realizou uma comparação dos descritores tradicionais MFCC com os obtidos através da transformada *Wavelet* e a *Wavelet Packet.* A *Wavelet* mãe utilizada foi a *Daubechies.* Dos resultados apresentados, vale salientar que em todas as condições testadas, os descritores baseados na *Wavelet,* apresentaram melhor desempenho do que os MFCC.

Kim et al [KIM, 2000] utilizaram o classificador HMM para o reconhecimento de palavras de zero a nove (dígitos) na língua Coreana. Os testes foram feitos para vários tipos de *Wavelets* mãe, ortogonais e biortogonais, e comparados com os descritores LPC e MFCC. Praticamente, todas as configurações de *Wavelet* tiveram desempenho igual, ou melhor,

que os descritores LPC e MFCC, sendo que as *Wavelets* ímpares tiveram melhor desempenho, fato que foi atribuído pelos autores ao melhor desempenho das *Wavelet* ímpares em alta freqüência.

Farooq e Datta [FAROOQ, 2003] apresentaram uma nova configuração para a *Wavelet Packet* (WP). Esta configuração aproxima a escala de filtros Mel. Foram obtidas 24 bandas de freqüências sendo: doze de 125 Hz, na faixa de 0 a 1375 Hz. Seis bandas de 250 Hz, situada na faixa entre 1375 e 3000 Hz. Duas bandas de 500 Hz, para a faixa de 3 kHz a 4 kHz. Por fim, quatro bandas de largura de 1 kHz dentro da faixa de 4 a 8kHz. Os resultados foram comparados com os descritores MFCC e em quase todos os casos houve um empate nas taxas de reconhecimento.

Chan *et al* [CHAN, 2001] também usaram a *Wavelet Packet* distribuída em uma escala parecida com a escala Mel no reconhecimento de voz de fonemas da língua inglesa. Os autores utilizaram o filtro de Winner para fazer a redução de ruído do sinal em quatro bandas da *Wavelet Packet*. Em todos os tipos de fonemas analisados, a *Wavelet Packet* teve melhor rendimento do que a MFCC, ocorrendo somente um empate nos fonemas oclusivos.

Jiang *et al* [JIANG, 2003] utilizaram a *Wavelet Packet* como descritor para o reconhecimento de voz de palavras isoladas (dígitos de 0 a 9) na língua inglesa. Os autores também procuraram estabelecer a árvore de decomposição da *Wavelet Packet* de forma que a mesma ficasse parecida com a escala Mel. Para determinar os descritores, os autores utilizaram a energia e o logaritmo de cada sub-banda. Também foram utilizados os coeficientes MFCC, para efeito de comparação. Em ambos os casos (com ruído e sem ruído) o desempenho da *Wavelet Packet* foi melhor do que os coeficientes MFCC.

Rocha *et al* [ROCHA, 2005] usaram a *Wavelet* com 6 sub-bandas no reconhecimento de comandos (palavras isoladas). Como descritores, foram utilizadas as energias de cada sub-bandas da *Wavelet* totalizando 6 coeficientes para cada *frame* (janela). Como as palavras possuem tamanhos diferentes os autores fizeram um sistema de janelamento dinâmico. O banco de vozes era composto de 2100 padrões de um mesmo locutor e 2331 locuções de locutores diferentes. Foram feitos 2 testes, sendo um independente do locutor e outro dependente do locutor, tendo como resultado um acerto de 93,22% e 99,37%, respectivamente.

Capítulo IV

4. Treinamento, Classificação e Decisão

Este capítulo tem como objetivo principal descrever as ferramentas utilizadas tanto no treinamento quanto na classificação dos padrões de voz. Basicamente, três principais teorias são descritas neste capítulo: o SVM (*Support Vector Machine*), as Máquinas de Comitê e a teoria sobre a Análise Estatística da Decisão.

Como descrito no Capítulo I, os modelos escondidos de Markov - HMM (*Hidden Markov Models*) dominam a grande maioria das pesquisas sobre reconhecimento e voz. Uma explanação completa sobre este método, desenvolvimento e aplicações podem ser encontradas em Rabiner [RABINER, 1993]. Os trabalhos de Waibel *et al* [WAIBEL, 1989], Tan *et al* [TAN, 1996], Sarikaya e Hansen [SARIKAYA, 2000], Zhu e Alwan [ZHU, 2000], Kim *et al* [KIM, 2000], Alcaim e Santos [ALCAIM, 2001], Chan *et al* [CHAN, 2001], Hosom [HOSOM, 2002] e Jiang *et al* [JIANG, 2003] são alguns exemplos de publicações que utilizam os modelos HMM no reconhecimento de voz.

No entanto, em seu editorial Russell e Bilmes [RUSSEL, 2003] declaram que nos últimos anos reascendeu o interesse em classificadores que possam ir além do desempenho dos sistemas baseados nos modelos de HMM. Aliado a isto, diversos trabalhos publicados recentemente utilizaram as Máquinas de Vetor de Suporte no reconhecimento de voz com sucesso, tais como: [CLARKSON, 1999], [ABE, 2002], [DESHMUKH, 2002], [PICONE, 2002], [JUNEJA, 2003], [ABDULLA, 2003], [MARINHO,2004] e [MPORAS, 2006].

4.1 Máquinas de Vetor de Suporte - SVM

A teoria sobre o SVM (*Support Vector Machine*) foi introduzida pela primeira vez por Vapnik [VAPNIK, 1992]. As Máquinas de Vetor de Suporte representam uma nova abordagem para a classificação de padrões e, recentemente, tem atraído um grande interesse junto à comunidade científica, especialmente nas áreas de classificação, regressão e aprendizagem de máquinas [BURGES, 1998]. Esta abordagem possui uma forte ligação com a teoria da aprendizagem estatística e a teoria da minimização estrutural do risco [CRISTIANINI, 2003].

Basicamente, o SVM faz o mapeamento do espaço de entrada para um espaço de alta dimensionalidade e a partir do cálculo de um hiperplano de separação ótimo, neste novo espaço, o SVM aprende a fronteira entre as regiões pertencentes a duas classes. Este hiperplano de separação é escolhido de forma que ele maximiza a distância de separação entre as amostras de treinamento mais próximas [HAYKIN, 2001].

4.1.1 Introdução Geométrica ao SVM

Os modelos de SVM foram inicialmente definidos para a classificação de classes linearmente separáveis. Um exemplo de duas classes linearmente separáveis é apresentado na Figura 4.1. Pode-se ver nesta figura duas classes de objetos bidimensionais nominados por [+1] e [-1].

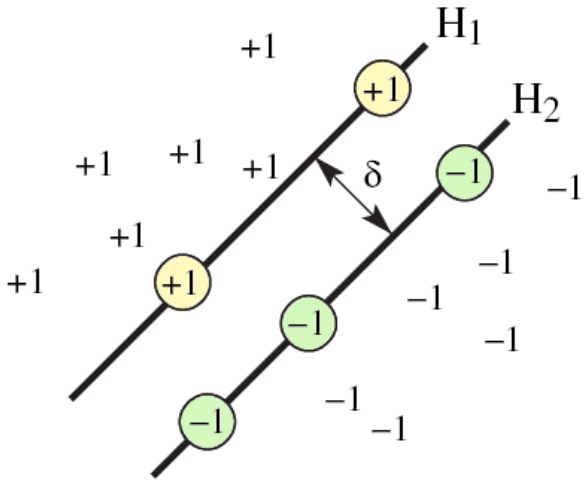

Figura 4.1: Formação do hiperplano de separação através dos vetores de suporte.

É relativamente fácil encontrar neste exemplo uma linha de separação para estas duas classes. Para este caso especial de duas classes linearmente separáveis, o SVM encontra um único hiperplano com máxima margem de separação denotada por δ. Este hiperplano estará localizado entre as linhas H_1 e H_2, e será ótimo se a distância para as duas linhas for máxima. A linha H_1 define a borda ou fronteira com os objetos da classe [+1]. Já a linha H_2 define a borda ou fronteira com os objetos da classe [-1]. Pode-se ver que dois objetos da classe [+1] definem a linha de fronteira H_1. Para a fronteira H_2 existem três objetos da classe [-1] que definem esta linha. Estes objetos marcados com um círculo na Figura 4.1 são chamados de vetores de suporte (*Support Vectors)*.

Esta é uma característica especial do SVM, ou seja, a solução para o problema de classificação é representada pelos vetores de suporte, os quais são fundamentais na determinação do hiperplano de separação com margem máxima.

O SVM pode ser igualmente usado para separar classes que não podem ser separadas com um classificador linear (Figura 4.2-esquerda), ou seja, padrões não-linearmente separáveis. Nesses casos, as coordenadas dos objetos são mapeadas do espaço de entrada para um espaço de característica usando funções não-lineares chamadas de funções de característica Φ. O espaço de característica é um espaço de alta dimensionalidade em que as duas classes podem ser separadas por um classificador linear (Figura 4.2-direita).

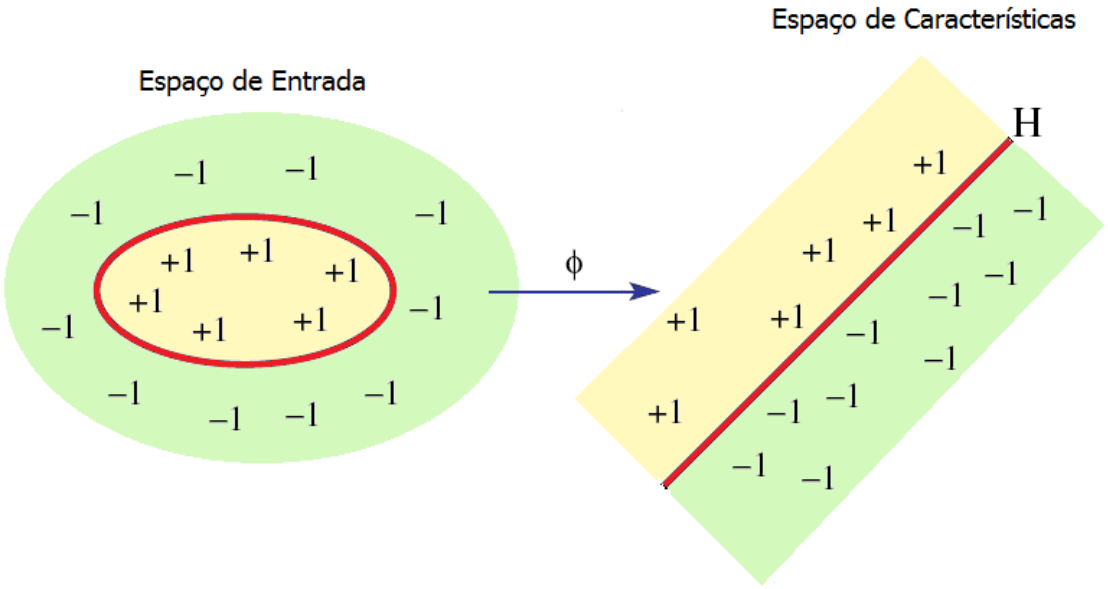

Figura 4.2: Padrões não-linearmente separáveis: separação linear no espaço de características.

Devido à alta dimensionalidade do espaço de características, não é prático usar diretamente a função de característica Φ para encontrar o hiperplano de separação. Ao invés disso, o mapeamento não-linear induzido pela função de característica é computado com o auxílio de funções não-lineares especiais chamadas de *Kernel*. O *Kernel* tem a vantagem de operar no espaço de entrada, onde a solução do problema de classificação é feito por meio da soma ponderada da função *Kernel* avaliada pelos vetores de suporte. Simplificando, o *Kernel* possibilita a construção de um hiperplano de separação ótimo no espaço de características sem ter que considerar o próprio espaço de características de forma explicita [HAYKIN, 2001].

A idéia principal do SVM é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima. Para isso a máquina faz uso de duas teorias: Dimensão VC e a Minimização Estrutural do Risco.

A dimensão VC (Vapnik e Chervonenkis) é o número máximo de exemplos de treinamento que podem ser aprendidos pela máquina sem erro.

Em um problema de aprendizagem supervisionada deseja-se obter o melhor desempenho de generalização adequando-se a capacidade da máquina com a quantidade disponível de dados de treinamento para o problema em questão. O método de minimização estrutural do risco fornece um procedimento indutivo, no qual utilizando a dimensão VC como uma variável de controle, pode-se obter este melhor desempenho. Resumindo, as SVMs possuem características importantes que justificam a sua utilização:

- Boa capacidade de generalização;
- Robustez diante de objetos de dimensões elevadas;
- Convexidade da função objetivo, ou seja, possui apenas um mínimo global;
- Capacidade de lidar com dados ruidosos e;
- Uma base teórica bem estabelecida na Matemática e Estatística.

4.1.2 A Dimensão VC

A dimensão VC denominada em homenagem aos seus criadores Vapnik e Chervonenkis, pode ser definida como a medida da capacidade de aprendizado de uma classe de funções que classifica corretamente o maior número de amostras de treinamento [CRISTIANINI, 2003].

A Figura 4.3 exemplifica esse conceito, a dimensão VC é igual a 3 pois este é o maior número de classes que podem ser separadas por uma reta para qualquer padrão de classificação binária que as amostras podem admitir. Por indução, a dimensão VC para funções lineares no $\Re^n$, com $n \geq 2$, é $n + 1$.

Existem classes de funções com valores de dimensão VC (capacidades) diferentes, tais como: lineares, exponenciais, polinomiais. Uma capacidade maior indica que é possível construir máquinas bem mais complexas, mas com tendência ao sobre-ajuste, causando a perda da generalização, e quanto menor for essa capacidade maior será a restrição no que a máquina pode fazer [HAYKIN, 2001]. Para resolver este problema utiliza-se a minimização do

risco estrutural (SRM – *Structural Risk Minimization)* a fim de encontrar um valor ótimo para esta capacidade.

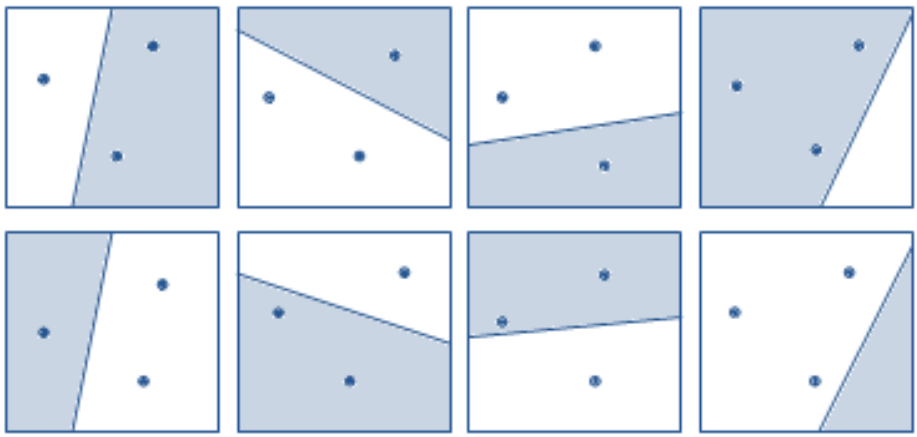

Figura 4.3: Dimensão VC: possíveis separações de três pontos por uma reta.

4.1.3 Minimização do Risco Estrutural

A minimização do risco estrutural, desenvolvida por Vapnik [VAPNIK, 1982], é uma forma de tratar o problema da escolha de uma dimensão apropriada. Dada uma estrutura na qual o conjunto com possíveis classes (hipóteses) foi divido em subconjuntos dado por:

$$F_1 \subset F_2 \subset \cdots \subset F_k \subset \cdots$$

onde para cada subconjunto F_k tem-se uma dimensão VC h_k. Cada subconjunto possui então a propriedade $h_k \leq h_{k+1}$.

A técnica de minimização do risco estrutural consiste em encontrar o subconjunto de funções que minimiza o "limite superior de risco" (erro de generalização). A solução pode ser encontrada a partir do treinamento de uma série de máquinas, uma para cada subconjunto, com o objetivo de minimizar o risco empírico [Haykin, 2001]. A máquina a ser escolhida será aquela cuja soma do risco empírico e da razão $\frac{h}{n}$ for a menor. O termo $\frac{h}{n}$ indica que a "capacidade da máquina" é diretamente proporcional a dimensão VC representada por h, e inversamente proporcional ao número de exemplos de treinamento n. Para um subconjunto particular F_k, sendo $\hat{f}_k$ o classificador com o menor risco empírico, à medida que k cresce o risco empírico de $\hat{f}_k$ diminui, uma vez que a complexidade do conjunto de classificadores é maior. Contudo, o termo de capacidade aumenta com k, resultando um valor ótimo $\bar{k}$ em

que se obtém uma soma mínima do risco empírico e do termo de capacidade, minimizando assim o limite sobre o risco esperado. Os conceitos acima mencionados podem ser visualizados na Figura 4.4.

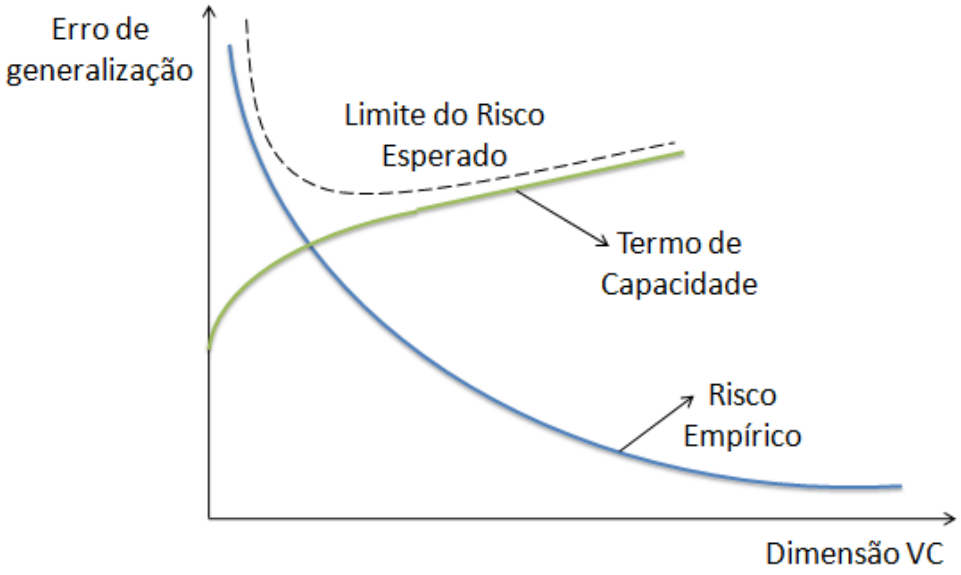

Figura 4.4: Princípio de minimização do risco estrutural.

4.1.4 Hiperplano Ótimo para Padrões Linearmente Separáveis

Nesta seção é feita uma pressuposição de um problema com duas classes linearmente separáveis, o objetivo é demonstrar a idéia básica ou os conceitos fundamentais das Máquinas de Vetor de Suporte em um cenário simples, mas completo do ponto de vista matemático. Na seção seguinte, estes conceitos serão expandidos para o caso de padrões não-linearmente separáveis.

Considerando uma amostra de treinamento $\{(x_i, d_i)\}_{i=1}^{N}$, onde x_i é o padrão de entrada para o i-ésimo exemplo e d_i é a resposta desejada correspondente. Inicialmente, assume-se que estes padrões representam duas classes distintas “linearmente separáveis”. A equação de uma superfície de decisão na forma de um hiperplano que realiza esta separação é dada pela Equação (4.1), sendo:

$$w^T x + b = 0 \tag{4.1}$$

onde x é um vetor de entrada, w é um vetor peso ajustável e b é um bias ou tendência. Pode-se assim escrever:

$$\begin{aligned} w^T x_i + b \geq 0 \qquad & para\ d_i = +1 \\ w^T x_i + b < 0 \qquad & para\ d_i = -1 \end{aligned} \tag{4.2}$$

A margem de separação (representada por ρ) é a distância entre o hiperplano definido na Equação (4.1) e o ponto de dado mais próximo, isto para um vetor peso w e bias b específicos. O objetivo de uma máquina de vetor de suporte é encontrar o hiperplano particular para o qual a margem de separação ρ é máxima. Sob esta condição, a superfície de decisão é referida como um Hiperplano ótimo.

A Figura 4.5 ilustra a construção geométrica de um hiperplano ótimo para um espaço de entrada bidimensional.

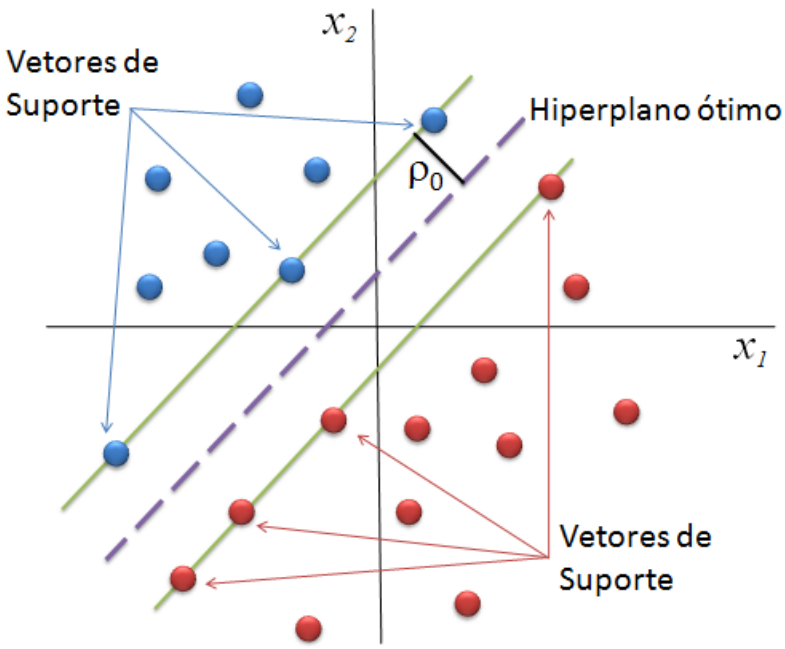

Figura 4.5: Hiperplano ótimo para padrões linearmente separáveis. Adaptação de [HAYKIN, 2001].

Considerando ainda que $\boldsymbol{w}_0$ e b_0 representam os valores ótimos do *vetor peso* e do bias, respectivamente. O hiperplano ótimo representa uma superfície de decisão linear multidimensional no espaço de entrada e é definido reescrevendo a Equação (4.1):

$$w_0^T \boldsymbol{x} + b_0 = 0 \tag{4.3}$$

Assim, a função discriminante que fornece uma medida algébrica da *distância* de $\boldsymbol{x}$ até o hiperplano [DUDA, 1973] é dada pela Equação (4.4).

$$g(x) = w_0^T x + b_0 \tag{4.4}$$

Um melhor modo de expressar $\boldsymbol{x}$ é feito pela Equação (4.5).

$$x = x_p + r\frac{w_o}{\|w_o\|} \tag{4.5}$$

onde $\boldsymbol{x}_p$ é a projeção normal de $\boldsymbol{x}$ sobre o hiperplano ótimo, e r é a distância algébrica desejada; r é positivo se $\boldsymbol{x}$ estiver no lado positivo do hiperplano ótimo e negativo se $\boldsymbol{x}$ estiver no lado negativo. Uma vez que por definição $g(\boldsymbol{x}_p) = 0$, então resulta que

$$g(x) = w_0^T x + b_0 = r\|w_0\| \tag{4.6}$$

ou

$$r = \frac{g(x)}{\|w_o\|} \tag{4.7}$$

Através da Figura 4.6, pode-se mostrar que a distância da origem (i.e., $\boldsymbol{x}$ = 0) até o hiperplano ótimo é dada por $b_0/\|w_0\|$. Se $b_0 > 0$, a origem está no lado positivo do hiperplano ótimo; se $b_0 < 0$ ela está no lado negativo. Se $b_0 = 0$, o hiperplano ótimo passa origem.

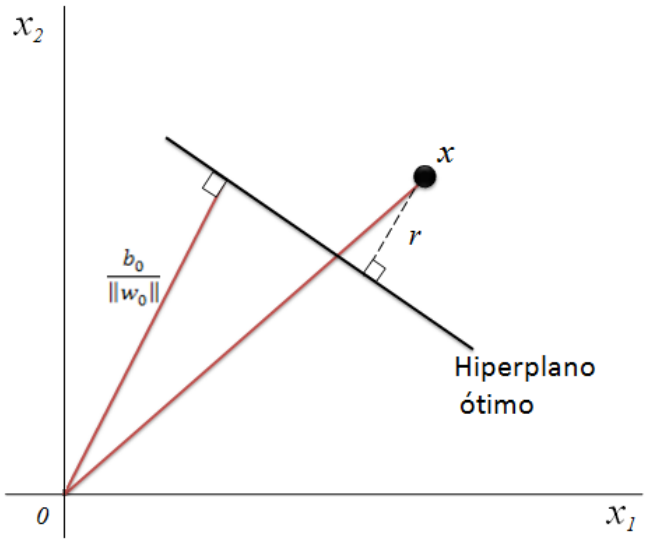

Figura 4.6: Distâncias algébricas de um ponto até o hiperplano ótimo para um caso bidimensional.

Assim, dado um conjunto de treinamento $\Im = \{(x_i, d_i)\}$, a questão agora é encontrar os parâmetros $\boldsymbol{w}_0$ e b_0 para o hiperplano ótimo. Com base nos dados da Figura 4.6, pode-se ver que o par ($\boldsymbol{w}_0$, b_0) deve satisfazer a restrição imposta pela Equação (4.8) dada por:

$$\begin{aligned} w_0^T x + b_0 &\geq +1 \qquad para\ d_i = +1 \\ w_0^T x + b_0 &\leq -1 \qquad para\ d_i = -1 \end{aligned} \tag{4.8}$$

Os pontos de dados particulares ($\boldsymbol{x}_i,\ d_i$) para a Equação (4.8) são chamados de *vetores de suporte*, isto é, são aqueles pontos de dados que se encontram mais próximos da superfície de decisão e são, portanto, os mais difíceis de classificar. Estes vetores de suporte têm influência direta na localização ótima da superfície de decisão, por isso o nome Máquina de Vetor de Suporte – SVM.

Assim, o hiperplano ótimo definido pela Equação (4.3) é único no sentido de que o vetor peso w_0 fornece a máxima separação possível entre exemplos positivos e negativos. Esta condição é alcançada minimizando a norma euclidiana do vetor peso $\boldsymbol{w}$. A margem de separação é então definida pela Equação (4.9),

$$\rho = \frac{2}{\|w_0\|} \tag{4.9}$$

4.1.5 Otimização Quadrática para Encontrar o Hiperplano Ótimo

Dada a amostra de treinamento $\Im = \{(\boldsymbol{x}_i, d_i)\}_{i=1}^{N}$, o objetivo agora é desenvolver um procedimento eficiente do ponto de vista computacional para encontrar o hiperplano ótimo sujeito às restrições da Equação (4.8), as quais foram combinadas na Equação (4.10), sendo:

$$d_i\left(\boldsymbol{w}^T\,\boldsymbol{x} + b\ \right) \geq 1 \qquad para\ i = 1, 2, \ldots, N \tag{4.10}$$

Este é um problema de otimização restrita chamado de *problema primordial*. Neste problema primordial, o objetivo é encontrar os valores ótimos do vetor peso $\boldsymbol{w}$ e do bias b, de modo que satisfaçam a restrição imposta pela Equação (4.10), sendo que o vetor peso $\boldsymbol{w}$ deve minimizar a função de custo dada por:

$$\phi(w) = \frac{1}{2} w^T w \tag{4.11}$$

Dado que este problema primordial é caracterizado por uma função $\Phi(\boldsymbol{w})$ convexa de $\boldsymbol{w}$, e as restrições são lineares em relação a $\boldsymbol{w}$, então pode-se resolver este problema usando o método dos multiplicadores de Lagrange [BERTSEKAS, 1995]. Para desenvolver o método de Lagrange, primeiro deve-se construir a função lagrangiana dada pela Equação (4.12) a seguir.

$$J(w,b,\alpha) = \frac{1}{2}w^T w - \sum_{i=1}^{N} \alpha_i \left[d_i \left(\boldsymbol{w}^T \boldsymbol{x_i} + b \right) - 1 \right] \qquad (4.12)$$

onde as variáveis α_i são não-negativas e chamadas de multiplicadores de Lagrange. O ponto de sela da função lagrangiana $J(w,b,\alpha)$ determina a solução do problema de otimização restrito através da sua minimização em relação a $\boldsymbol{w}$ e b, e a maximização em relação a α.

Diferenciando $J(w,b,\alpha)$ na Equação (4.12) em relação $\boldsymbol{w}$ e b e igualando a zero tem-se as seguintes condições de otimização:

$$\begin{aligned} &Condição\ 1: \qquad \frac{dJ(w,b,\alpha)}{dw} = 0 \\ &Condição\ 2: \qquad \frac{dJ(w,b,\alpha)}{db} = 0 \end{aligned} \qquad (4.13)$$

Da condição 1 e remanejando a função lagrangiana da Equação (4.12), produz-se a Equação (4.14). No caso da aplicação da condição 2, o resultado obtido é descrito pela Equação (4.15), sendo:

$$w = \sum_{i=1}^{N} \alpha_i d_i x_i \qquad (4.14)$$

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \qquad (4.15)$$

O vetor solução $\boldsymbol{w}$ é definido em termos de uma expansão que envolve os N exemplos de treinamento. No entanto, embora esta solução seja única em virtude da convexidade da lagrangiana, o mesmo não acontece com os coeficientes de Lagrange α_i.

É importante notar que no ponto de cela, para cada multiplicador de Lagrange, o produto deste multiplicador pela sua restrição correspondente desaparece conforme pode ser visto na Equação (4.16) a seguir.

$$\alpha_i \left[d_i \left(\boldsymbol{w}^T \boldsymbol{x_i} + b \right) - 1 \right] = 0 \qquad para\ i = 1, 2, \ldots N \qquad (4.16)$$

Assim, conclui-se que somente os multiplicadores que satisfaçam a Equação (4.16) podem assumir valores não-nulos.

Dado que o problema primordial trabalha com uma função de custo convexa e com restrições lineares, então é possível construir um outro problema chamado de *problema dual*, o qual possui multiplicadores de Lagrange que fornecem a solução ótima. No teorema da dualidade [BERTSEKAS, 1995], diz-se que, se o problema primordial tem uma solução ótima, então, o problema dual também tem uma solução ótima e os valores ótimos correspondentes são iguais. Além disso, para que $\boldsymbol{w}_0$ seja uma solução primordial ótima e α_0 seja uma solução dual ótima, é necessário e suficiente que $\boldsymbol{w}_0$ seja realizável para o problema primordial, sendo:

$$\phi(w_0) = J(w_0, b_0, \alpha_0) = min_w J(w, b_0, \alpha_0) \tag{4.17}$$

O problema dual pode ser postulado expandindo a Equação 4.12, como:

$$J(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i d_i w^T \boldsymbol{x_i} - b \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i \tag{4.18}$$

Devido à condição imposta pela Equação (4.15), o terceiro termo da Equação (4.18) é nulo. Além disso, da Equação (4.14) pode-se extrair

$$w^T w = \sum_{i=1}^{N} \alpha_i d_i w^T \boldsymbol{x_i} = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, \alpha_j d_i d_j x_i^T x_j \tag{4.19}$$

Fazendo a função objetivo $J(w, b, \alpha) = Q(\alpha)$, pode-se então reformular a Equação (4.18) como:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \, \alpha_j d_i d_j x_i^T x_j \tag{4.20}$$

Deste modo, a Equação (4.20) representa a formulação do problema dual, ou seja, a partir de uma amostra de treinamento $\{(\boldsymbol{x}_i, d_i)\}_{i=1}^{N}$, deve-se encontrar os multiplicadores de Lagrange $\{\alpha_i\}_{i=1}^{N}$ que maximizam a função objetivo (Equação 4.20), sujeito às restrições:

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \quad \text{e} \quad \alpha_i \geq 0 \qquad para\ i = 1, 2, \ldots, N$$

Uma vez calculados os multiplicadores de Lagrange ótimos $\alpha_{o,i}$, pode-se então calcular o vetor de peso $\boldsymbol{w}_0$ usando a Equação (4.14) e o bias ótimo b_0, sendo:

$$w_0 = \sum_{i=1}^{N} \alpha_{0,i} d_i x_i \tag{4.21}$$

$$b_0 = 1 - w_0^T x^{(s)} \qquad para\ d^{(s)} = 1 \tag{4.22}$$

onde $x^{(s)}$ representa um vetor de suporte para o qual sua desejada $d^{(s)}$ é igual a 1.

4.1.6 Hiperplano Ótimo para Padrões Não-Separáveis Linearmente

Considerando agora que os padrões são não-separáveis, então para este novo conjunto de treinamento, não é possível construir um hiperplano de separação sem se defrontar com erros de classificação [HAYKIN, 2001]. Mesmo assim, pode-se encontrar um hiperplano ótimo que minimize a probabilidade de erro de classificação, neste caso a probabilidade é calculada como a média sobre o conjunto de treinamento. Este caso é dito como sendo flexível, pois existirão pontos (exemplos de treinamento) que infringirão as desigualdades da Equação (4.10), reescrita abaixo por simplicidade, sendo:

$$d_i(w^T x_i + b) \geq 1 \qquad para\ i = 1, 2, \cdots, N$$

A Figura 4.7 ilustra as duas formas diferentes de violação da restrição imposta pela Equação (4.10).

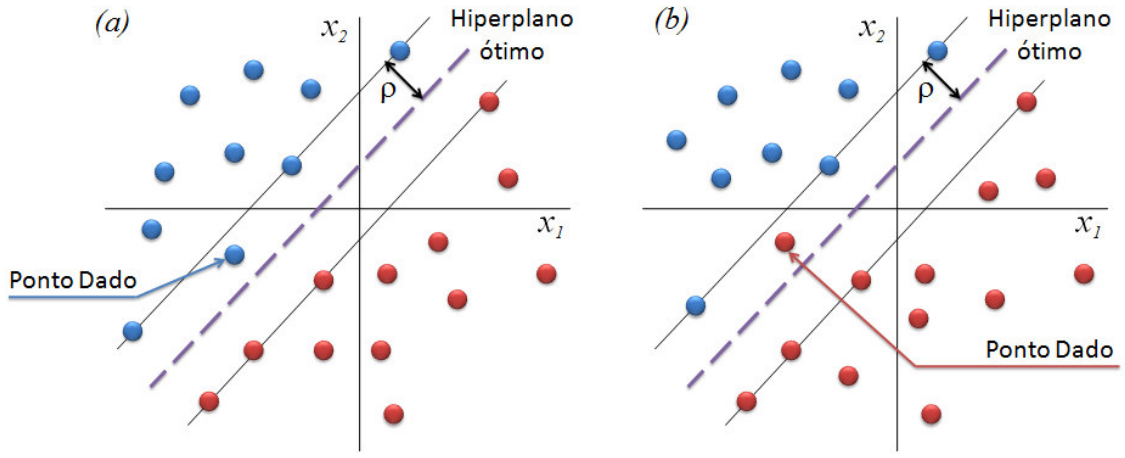

Figura 4.7: (a) Ponto de dado no lado correto e, (b) Ponto de dado do lado errado do Hiperplano.

Pode-se observar na Figura 4.7-a que o ponto de dado (em azul), apesar de violar a restrição, encontra-se no lado correto, ou seja, a classificação é correta. Já na Figura 4.7-b, o ponto de dado (em vermelho) está posicionado na região da outra classe (azul), deste modo

a classificação está incorreta, pois o ponto marcado está localizado do lado incorreto do hiperplano e dentro da margem de separação, indicando a violação da restrição.

Para generalizar a situação descrita acima, é inserida uma variável escalar e não negativa $\xi = (\xi_1, \xi_2, \cdots \xi_l)$ chamada de variável solta que é incluída na equação que define o hiperplano de separação, dado por:

$$d_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad i = 1, 2, \ldots, N \tag{4.23}$$

As variáveis soltas ξ_i medem o desvio de cada amostra de sua condição ideal de separabilidade de padrões. Para $0 < \xi \leq 1$ o ponto de dado se encontra dentro da região de separação, mas no lado correto da superfície de decisão, como observado na Figura 4.7-a. Para ξ_i >1, a amostra está localizada no lado incorreto do hiperplano de separação, Figura 4.7-b.

Os vetores de suporte são as amostras que satisfazem a igualdade presente na Equação (4.23), isto é, são as amostras que estão mais próximas do hiperplano.

É importante ressaltar que, se um exemplo $\xi_i > 0$ for deixado fora do conjunto de treinamento, a superfície de decisão não muda. Deste modo, os vetores de suporte são definidos do mesmo modo tanto para o caso linearmente separável como para o caso não-linearmente separável.

Novamente, o objetivo é encontrar um hiperplano de separação para o qual o erro de classificação, como média sobre o conjunto de treinamento, é minimizado. Isto pode ser feito minimizando o funcional dado pela Equação (4.24),

$$\Phi(\xi) = \sum_{i=1}^{N} I(\xi_i - 1) \tag{4.24}$$

em relação ao vetor peso $\boldsymbol{w}$, a restrição da Equação (4.23) e a restrição em relação a $\|\boldsymbol{w}\|^2$:

$$\|\boldsymbol{w}\|^2 \leq \frac{1}{\rho} \tag{4.25}$$

A função $I(\xi)$ é uma função indicadora e definida por:

$$I(\xi) = \begin{cases} 0 \; se \; \xi \leq 0 \\ 1 \; se \; \xi > 0 \end{cases} \tag{4.26}$$

A minimização de $\Phi(\xi)$ em relação a $\boldsymbol{w}$ é um problema de otimização que pertence a uma classe de problemas NP completo. Para tratar esta questão deve ser feita uma aproximação, a qual é dada pela Equação (4.27), sendo:

$$\Phi(\xi) = \sum_{i=1}^{N} \xi_i \tag{4.27}$$

Além disso, deve-se fazer com que o funcional seja minimizado em relação ao vetor peso **w**, conforme a Equação (4.28), dada por:

$$\Phi(\boldsymbol{w}, \xi) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N} \xi_i \tag{4.28}$$

A minimização de $\boldsymbol{w}$ esta relacionada com a minimização da dimensão VC. Já o segundo termo da Equação 4.28 pode ser pensado como limite superior para o erro de classificação. O parâmetro *C* pode ser considerado como um parâmetro de regularização, isto é, controla o compromisso entre a complexidade da máquina e o número de erros de treinamento. Este parâmetro é escolhido pelo usuário e normalmente é determinado experimentalmente, através do desempenho do algoritmo via dados de validação, ou de forma analítica estimando a dimensão VC.

Portanto, o funcional da Equação (4.28) é otimizado em relação a $\boldsymbol{w}$ e $\{\xi_i\}_{i=1}^{N}$, sujeito à restrição da Equação (4.23) e $\xi_i \geq 0$. Deste modo, o problema de otimização para padrões não-separáveis inclui o problema de otimização para padrões linearmente separáveis como um caso especial.

Em analogia com o que foi feito para o caso linearmente separável, agora pode-se formalizar o problema primordial para o caso de padrões não-separáveis, como: dado uma amostra de treinamento $\{(\boldsymbol{x}_i, d_i)\}_{i=1}^{N}$, deve-se encontrar os valores ótimos do vetor peso **w** e do bias b de modo que satisfaça a restrição da Equação (4.23) reescrita, sendo:

$$d_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad para\ i = 1, 2, \ldots, N$$

$$\text{e} \quad \xi_i \geq 0 \qquad para\ todo\ i.$$

Além disso, o vetor peso **w** e as variáveis soltas $\xi_i$devem minimizar o funcional de custo dada pela Equação 4.28, onde *C* é um parâmetro positivo especificado pelo usuário.

Em seguida, usando o método dos multiplicadores de Lagrange, e novamente de maneira similar a caso de padrões linearmente separáveis, pode-se ter o problema dual formulado como: dada a amostra de treinamento $\{(\boldsymbol{x}_i, d_i)\}_{i=1}^{N}$, deve-se encontrar os multiplicadores de Lagrange $\{\alpha_i\}_{i=1}^{N}$ que maximizem a função objetivo dada por:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j \tag{4.29}$$

sujeito a restrições:

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \text{ e } \quad 0 \leq \alpha_i \leq C \qquad para\ i = 1, 2, \ldots, N$$

A função objetivo *Q(α)* a ser maximizada é a mesma para os casos de padrões linearmente separável e não-separável. O caso não-separável difere do caso separável pelo fato de que a restrição $\alpha_i \geq 0$ é substituída pela restrição mais rigorosa $0 \leq \alpha_i \leq C$. Exceto por esta modificação, a otimização restrita para o caso não-separável e os cálculos dos valores ótimos do vetor peso **w** e do *bias b* procedem do mesmo modo como no caso linearmente separável. A solução ótima para o vetor peso **w** é dada por:

$$\boldsymbol{w}_o = \sum_{i=1}^{N_s} \alpha_{o,i} d_i \boldsymbol{x}_i \tag{4.30}$$

onde N_s é o número de vetores de suporte. Enquanto que b pode ser determinado a partir de α, e pelas novas condições de *Karush-Kuhn-Tucker* [Haykin, 2001]:

$$\alpha_i [d_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 + \xi_i] = 0 \qquad i = 1, 2, \ldots, N \tag{4.31}$$

e

$$\mu_i \xi_i = 0 \qquad i = 1, 2, \ldots, N \tag{4.32}$$

Os μ_i são multiplicadores de Lagrange que foram introduzidos para forçar a não-negatividade das variáveis soltas ξ_i para todo *i*. No cálculo do ponto mínimo (ponto de sela), a derivativa da função Lagrangeana com respeito as variáveis ξ_i é zero, sendo assim:

$$\alpha_i + \mu_i = C \tag{4.33}$$

Combinando as Equações (4.32) e (4.33) pode-se notar que

$$\xi_i = 0 \quad se \quad \alpha_i = C \tag{4.34}$$

Por fim, pode-se determinar o bias ótimo b_0 utilizando qualquer ponto de dado (x_i, d_i) do conjunto de treinamento na Equação (4.31), para o qual tem-se $0 \leq \alpha_i \leq C$ e $\xi_i = 0$.

4.1.7 As Funções *Kernel*

A superfície de decisão da SVM, que no espaço de características é sempre linear, normalmente é não linear no espaço de entrada. Como visto anteriormente, a idéia de uma Máquina de Vetor de Suporte depende de duas operações matemáticas que podem ser resumidas como: Primeiro o mapeamento não-linear de um vetor de entrada para um espaço de características de alta dimensionalidade, que é oculto da entrada e da saída. Em segundo lugar, é necessário construir um hiperplano ótimo para separar as características descobertas no primeiro passo. Para construir este hiperplano ótimo necessita-se de uma função *Kernel*, ou núcleo do produto interno.

Um *Kernel* é uma função que recebe dois pontos x_i e x_j do espaço de entradas e calcula o produto escalar desses dados no espaço de características, dado por:

$$k(x_i, x_j) = \Phi^{\mathrm{T}}(x_i) \cdot \Phi(x_j) \tag{4.35}$$

Para garantir a convexidade do problema de otimização e que o *Kernel* apresente mapeamento no qual seja possível o cálculo de produtos escalares, deve-se utilizar uma função *Kernel* que siga as condições estabelecidas pelo teorema de Mercer [MERCER, 1909]. Os *Kernels* que satisfazem a condição de Mercer são caracterizados por dar origem a matrizes positivas semi-definidas *k*, em que cada elemento k_{ij} é definido por $k_{ij} = k(x_i, x_j)$ para todo $(i, j = 1, \cdots, n)$.

Uma vez que o mapeamento das SVMs é realizado por uma função *Kernel*, e não diretamente por $\Phi(x)$, nem sempre é possível saber exatamente qual mapeamento é efetivamente realizado, pois as funções de *Kernel* realizam um mapeamento implícito dos dados [HAYKIN, 2001].

Desta maneira, nas formulações da SVM os exemplos de treinamento nunca aparecem isolados, mas sempre em forma de um produto interno, que pode ser substituído por uma função de *Kernel* a ser escolhida [VAPNIK, 1995].

A Tabela 4.1 apresenta algumas funções comumente utilizadas como função *Kernel*. A expansão do núcleo do produto interno $K(x_i, x_j)$, na Equação (4.35), permite construir uma superfice de decisão que é não-linear no espaço de entrada, mas cuja imagem no espaço de característica é linear.

Tabela 4.1: Funções típicas usadas como *Kernel*.

Tipo de *Kernel*	Expressão	Parâmetro
Polinomial	$\left(x^T x_i + 1\right)^p$	A potência p é especificada pelo usuário *a priori*
RBF Kernel Gaussiano	$\exp\left(-\frac{1}{2\sigma^2}\left\|x - x_i\right\|^2\right)$	A largura σ^2, comum a todos os núcleos, é especificada *a priori* pelo usuário
Perceptron	$\tanh\left(\beta_0 x^T x_i \beta_1\right)$	O teorema de Mercer é satisfeito apenas para alguns valores de β_0, β_1

Deste modo, substituindo o produto interno $(\boldsymbol{x}_i^T, \boldsymbol{x}_j)$ pelo *Kernel* $K(x_i, x_j)$ na Equação (4.29), a construção do problema dual para a otimização restrita de uma máquina de vetor de suporte é finalizada da seguinte forma: dada a amostra de treinamento $\{(\boldsymbol{x}_i, d_i)\}_{i=1}^N$, deve-se encontrar os multiplicadores de Lagrange $\{\alpha_i\}_{i=1}^N$ que maximizem a função objetivo dada por:

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(x_i, x_j) \qquad (4.36)$$

sujeito a restrições:

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \text{ e } \quad 0 \le \alpha_i \le C \qquad para\ i = 1, 2, \ldots, N$$

onde *C* é um parâmetro positivo especificado pelo usuário.

Independente de como uma máquina de vetor suporte é implementada, ela difere da abordagem convencional do projeto de um *perceptron* de múltiplas camadas de uma forma fundamental. Na abordagem convencional, a complexidade do modelo é controlada mantendo-se o número de características (neurônios ocultos) pequeno. Por outro lado, a máquina de vetor de suporte oferece uma solução para o projeto de uma máquina de aprendizagem controlando a complexidade do modelo independentemente da dimensionalidade [COLLOBERT, 2004].

As SVMs também apresentam uma característica atrativa que é a convexidade do problema de otimização formulado em seu treinamento, implicando na existência de um único mínimo global. O uso de funções *Kernel* para padrões não linearmente separáveis, torna o algoritmo eficiente, pois permite a construção de simples hiperplanos em um espaço de alta dimensão de forma tratável do ponto de vista computacional [HAYKIN, 2001].

4.1.8 SVM para Múltiplas Classes

O SVM é um algoritmo para classificação de padrões baseado em duas classes. No entanto, pode-se construir um classificador de padrões baseado no SVM para múltiplas classes (*multiclass*). Apesar de existirem vários estudos sobre a classificação do tipo *multiclass* com o SVM, ainda não existe um padrão definitivo [CLARKSON, 1999].

Scholkopf *et al* [SCHOLKOPF, 1995] propuseram o modelo de classificador do tipo "um contra todos" (*one vs. all*). Clarkson e Moreno [CLARKSON, 1999] propuseram o modelo de classificador do tipo "um contra um" (*one vs. one*). No entanto, ambos modelos são de fato classificadores de somente duas classes: Classe +1 (C_{+1}) e Classe -1 (C_{-1}).

Como exemplo, pode-se citar o caso de separação de três classes C_X, C_Y *e* C_Z. Para o modelo "um contra todos" são utilizadas três máquinas as quais são treinadas da seguinte forma: A máquina 01 é treinada para [C_X contra Todos], ou seja C_X é considerada a classe +1, enquanto que o conjunto formado por C_Y *e* C_Z (todos) é considerado com sendo a classe -1. A máquina 02 é treinada para [C_Y contra Todos] e a máquina 03 para [C_Z contra Todos], respectivamente. Na Figura 4.8 é apresentado este modelo.

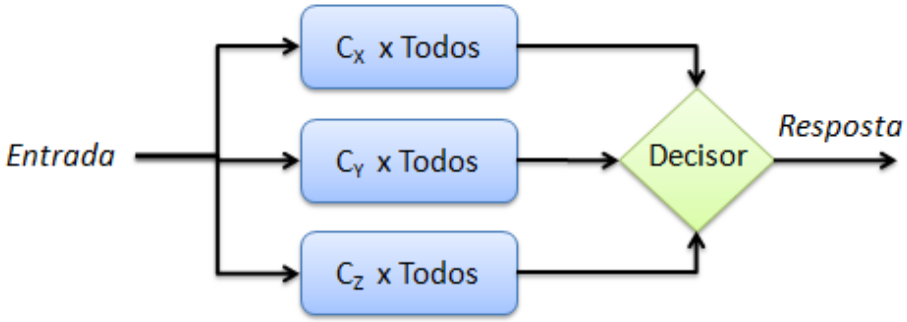

Figura 4.8: SVM *multiclass* para três classes classificadas através do modelo "um contra todos".

Na Figura 4.8, pode-se ver as três máquinas utilizadas para classificação de um padrão desconhecido, que é aplicado na entrada através da estratégia (modelo) "um contra todos".

Cada uma das três máquinas pode ser considerada como um classificador binário, ou seja, elas apresentam como resultado apenas duas hipóteses: (+1) que corresponde à primeira classe e (-1) que corresponde à segunda classe treinada. Se a máquina 01 [C_X contra Todos] apresentar o resultado +1 então o padrão não identificado é classificado como C_X. Por outro lado, se a resposta da máquina for -1, então o padrão desconhecido é classificado como "todos", ou "não C_X", ou ainda diz-se que o padrão desconhecido pode pertencer a C_Y *ou* C_Z. O mesmo raciocínio pode ser aplicado as outras duas máquinas.

Deste modo, se cada uma das 3 máquinas da Figura 4.8 possui duas respostas possíveis, então existem oito combinações ou respostas possíveis. O passo nomeado por "decisor" da Figura 4.8 serve para verificar as respostas encontradas pelas três máquinas e decidir qual é a classificação final do padrão desconhecido. A Tabela 4.2 apresenta os respostas para as três máquinas do exemplo da Figura 4.8 e os respectivos resultados para a análise.

Tabela 4.2: Análise dos resultados da estratégia "um contra todos" para três classes.

Máq. SVM →	01	02	03	Resultado da
↓ Respostas	[C_X x Todos]	[C_Y x Todos]	[C_Z x Todos]	Análise
1	**1**	-1	-1	C_X
2	-1	**1**	-1	C_Y
3	-1	-1	**1**	C_Z
4	1	1	1	Triplo Empate
5	-1	-1	-1	Inconsistente
6	1	1	-1	Empate
7	1	-1	1	Empate
8	-1	1	1	Empate

As combinações das respostas 1, 2 e 3 correspondem à classificação correta para as classes C_X, C_Y e C_Z, respectivamente. Vale salientar que a resposta 5 é considerada como inconsistência pois não é possível encontrar nenhuma resposta. Para as respostas 4, 6, 7 e 8 ocorrem empates os quais podem ser decididos por meio da utilização máquinas treinadas através da estratégia "um contra um".

Para o caso da resposta número 6, por exemplo, pode-se observar um empate entre as classes C_X *e* C_Y, uma máquina treinada com C_X *versus* C_Y resolveria a questão.

No entanto, a solução para os desempates requer a construção de mais máquinas especialistas, aumentado a complexidade do sistema.

Para o mesmo exemplo de separação de três classes C_X, C_Y e C_Z, pode-se também fazer a classificação com somente três máquinas usando a estratégia "um contra um". Neste caso, a máquina 01 é treinada para classificar [C_X contra C_Y], ou seja, C_X é considerada a classe +1, enquanto que o C_Y é considerado com a classe -1. A máquina 02 classifica [C_X contra C_Z] e a máquina 03 classifica [C_Y contra C_Z], respectivamente. A Figura 4.9 apresenta este modelo.

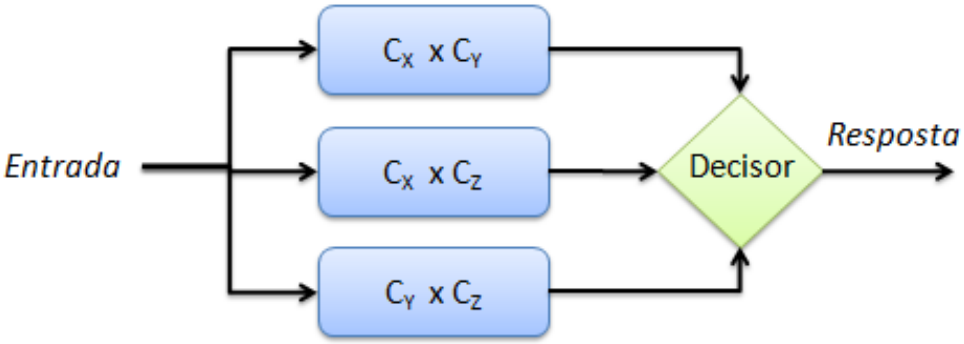

Figura 4.9: SVM *multiclass* para três classes classificadas através do modelo "um contra um".

Apesar de possuir duas respostas inconsistentes, ou seja, duas combinações de respostas sem classificação possível, as outras 6 combinações de respostas oferecem uma decisão conclusiva. A Tabela 4.3 apresenta as respostas possíveis para este caso. Apesar da estratégia "um contra um" ser estatisticamente mais eficiente na separação de três classes, quando o número de classes aumenta a estratégia "um contra todos" torna-se mais atrativa.

Tabela 4.3: Análise dos resultados da estratégia "um contra um" para três classes.

Máq. SVM → ↓ Respostas	01 [C_X x C_Y]	02 [C_X x C_Z]	03 [C_Y x C_Z]	Resultado Análise
1	**1**	**1**	1	C_X
2	**1**	**1**	-1	C_X
3	**-1**	1	**1**	C_Y
4	**-1**	-1	**1**	C_Y
5	1	**-1**	**-1**	C_Z
6	-1	**-1**	**-1**	C_Z
7	1	-1	1	Inconsistente
8	-1	1	-1	Inconsistente

4.2 Máquinas de Comitê

A teoria das Máquinas de Comitê está fortemente ligada ao princípio básico da engenharia "Dividir para Conquistar". De acordo com este princípio, uma tarefa

computacional complexa é resolvida dividindo-a em um número de tarefas computacionais simples, e então combinando as soluções destas tarefas para obtenção do resultado final [HAYKIN, 2001].

Na aprendizagem supervisionada, a simplicidade computacional é alcançada distribuindo a tarefa de aprendizagem entre um número de especialistas, que, por sua vez, divide o espaço de entrada em um conjunto de subespaços. Esta combinação de especialistas é chamada de Máquina de Comitê.

Uma Máquina de Comitê funde o conhecimento adquirido por diversos especialistas para chegar a decisão global que é supostamente superior àquela alcançável por qualquer um deles atuando isoladamente.

A idéia de uma máquina de comitê remonta ao trabalho de Nilsson [NILSSON, 1965]. A estrutura de rede considerada por Nilsson consistia de uma camada de perceptrons elementares seguida de um perceptron de votação na segunda camada da rede.

Basicamente, pode-se considerar que as Máquinas de Comitê são aproximadores universais. Elas podem ser classificadas em duas grandes categorias: Estruturas Estáticas e Estruturas Dinâmicas.

Nas estruturas estáticas as respostas das várias máquinas especialistas são combinadas por meio de um mecanismo que não envolve o sinal de entrada, elas possuem dois métodos:

- Média de Ensemble: Saídas de diferentes previsores são combinadas linearmente para produzir uma saída global.
- Reforço: Um algoritmo de aprendizagem com baixa precisão é convertido em um algoritmo que alcança um precisão arbitrariamente alta.

Já nas estruturas dinâmicas, o sinal de entrada está envolvido na atuação de todos os especialistas adiante na estrutura. Estas estruturas também possuem dois métodos de aplicação:

- Mistura de Especialistas: As respostas individuais dos especialistas são combinadas não linearmente por meio de uma única rede de passagem.
- Mistura Hierárquica de Especialistas: neste método, as respostas individuais dos especialistas são combinadas não linearmente por meio de várias redes de passagem arranjadas em uma forma hierárquica.

Estes dois métodos de estruturas dinâmicas podem ser vistos como exemplos de uma rede modular [Haykin, 2001]. Uma definição formal da noção de modularidade pode ser encontrada em [OSHERSON, 1990]. Basicamente, uma rede neural pode ser chamada de modular se a computação realizada pela rede pode ser decomposta em dois ou mais módulos (subsistemas) que operam sobre entradas distintas sem comunicação entre eles e sem realimentação.

4.2.1 Estruturas Estáticas

Na Figura 4.10, é apresentado um exemplo de Máquina de Comitê na forma de estrutura estática baseada na Média de *Ensemble*. Cada máquina especialista é treinada de forma diferente e todas as máquinas compartilham uma entrada comum. As saídas individuais de cada máquina são combinadas de alguma forma para produzir uma saída global y, dai o nome Média de *Ensemble* ou média do conjunto.

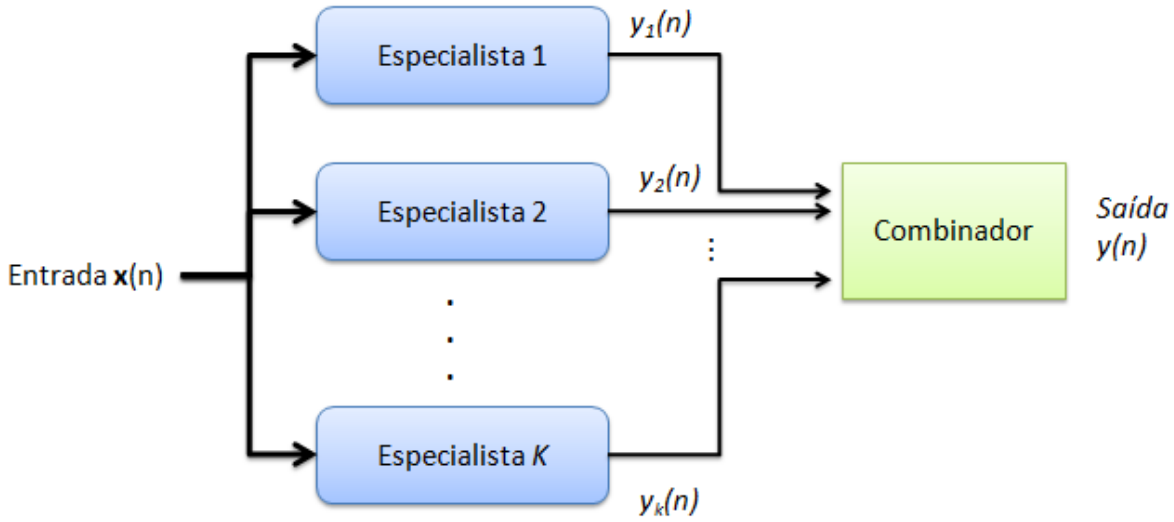

Figura 4.10: Diagrama de uma máquina de comitê baseada na média de *ensemble,* [HAYKIN, 2001].

A expectativa no uso desta estrutura é que os especialistas, diferentemente treinados, convirjam para diferentes mínimos locais na superfície de erro, e deste modo, o desempenho global possa ser melhorado com a combinação das saídas individuais.

É importante notar que, se a combinação de especialistas da Figura 4.10 fosse substituída por uma única rede neural, um grande números de parâmetros ajustáveis seriam necessários para o treinamento desta única rede. O tempo de treinamento para uma rede

grande assim seria provavelmente muito maior do que para o caso do conjunto de especialistas treinados em paralelo.

O reforço é outro método que pertence à classe estática das máquinas de comitê. O reforço é bastante diferente da média de *ensemble*. Em uma máquina de comitê baseada na média de *ensemble*, todos os especialistas da máquina são treinados com o mesmo conjunto de dados e podem diferir entre si na escolha das condições iniciais usadas no treinamento da rede. Em uma máquina por reforço, ao contrário, os especialistas são treinados com conjuntos de dados com distribuições inteiramente diferentes. O método do reforço visa melhorar o desempenho de qualquer algoritmo de aprendizagem. Basicamente, o reforço pode ser implementado de três formas diferentes:

- Reforço por Filtragem: Esta abordagem envolve a filtragem dos exemplos de treinamento por diferentes versões de um algoritmo de aprendizagem fraca. Este método necessita a disponibilidade de uma grande fonte de exemplos. Uma vantagem desta abordagem é que ela requer pouca memória comparada com as outras duas abordagens.
- Reforço por Sub-amostragem: Esta segunda abordagem trabalha com uma amostra de treinamento de tamanho fixo. Os exemplos são amostrados novamente durante o treinamento, de acordo com uma determinada distribuição de probabilidade.
- Reforço por Ponderação: A exemplo do reforço por sub-amostragem, esta abordagem também trabalha com uma amostra de treinamento fixa, mas assume que o algoritmo de aprendizagem fraca pode receber exemplos ponderados.

4.2.2 Estruturas Dinâmicas

Uma estrutura dinâmica denominada de mistura de especialistas, é uma rede constituída por *K* módulos supervisionados chamados de redes especialistas ou simplesmente especialistas, e por uma unidade integradora chamada de rede de passagem que desempenha a função de um mediador entre as redes de especialistas. Assume-se aqui que os diferentes especialistas funcionem melhor em regiões diferentes dos espaço de entrada de acordo com o modelo probabilístico de geração, por isso, a necessidade da rede de passagem.

A Figura 4.11 apresenta o diagrama de blocos do modelo de Mistura de Especialistas com a rede de passagem. Uma interpretação probabilística do papel da rede de passagem, seria vê-la como um "classificador" que mapeia o vetor de entrada **x** em probabilidades multinominais de modo que os diferentes especialistas serão capazes de encontrar a resposta desejada.

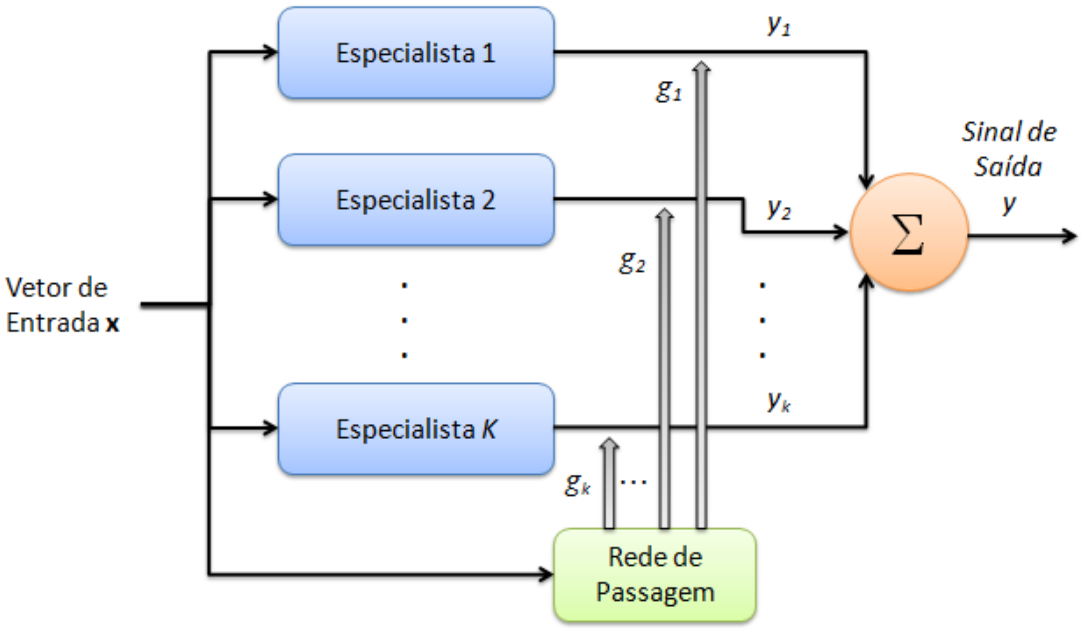

Figura 4.11: Diagrama de blocos de modelo de mistura de especialistas, [HAYKIN, 2001].

Resumindo, o modelo de mistura de especialistas - ME da Figura 4.11 funciona dividindo o espaço de entrada em diferentes subespaços, com uma única rede de passagem responsável pela distribuição da informação (extraída dos dados de treinamento) para os vários especialistas.

O modelo de mistura hierárquica de especialistas, que também pertence ao grupo de estruturas dinâmicas, é uma extensão natural do modelo ME. A arquitetura do modelo de mistura hierárquica de especialistas – MHE é similar a uma árvore, na qual as redes de passagem estão em vários pontos não terminais e os especialistas se encontram nas folhas da árvore. A Figura 4.12 apresenta esta estrutura com dois níveis.

O modelo MHE se diferencia do modelo ME na medida em que o espaço de entrada é dividido em conjuntos aninhados de subespaços, com a informação sendo combinada e redistribuída entre os especialistas sob o controle de várias redes de passagem arranjadas em uma forma hierárquica.

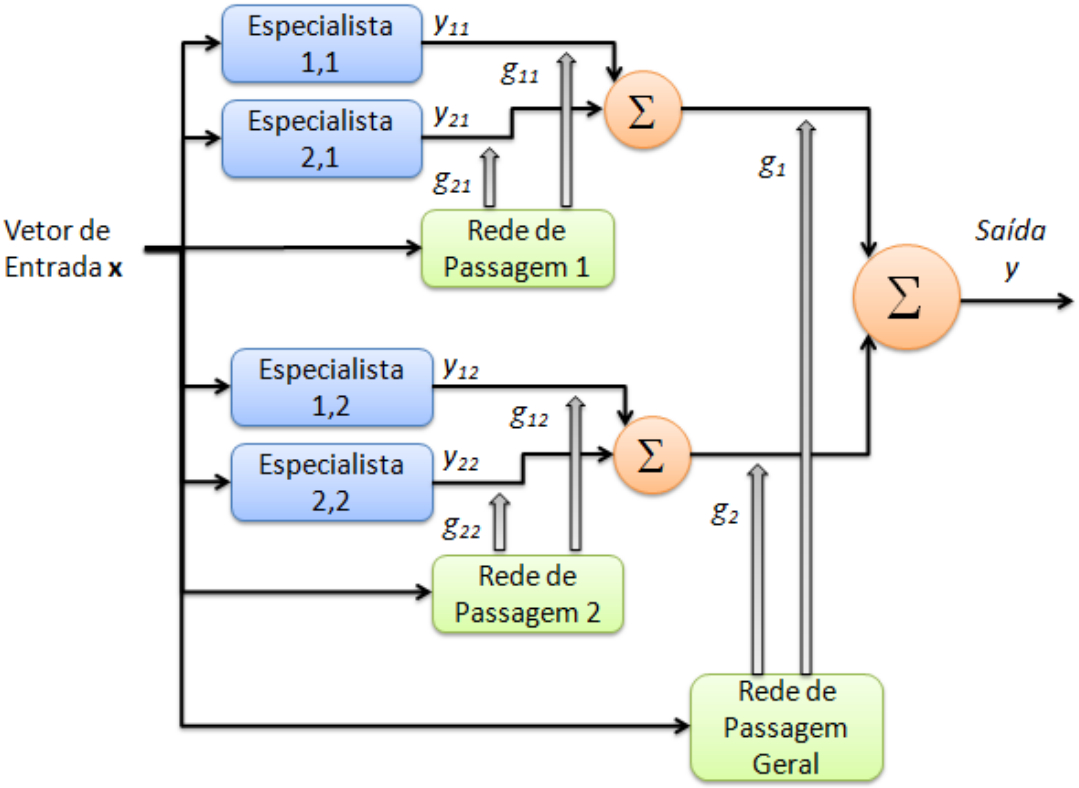

Figura 4.12: Mistura hierárquica de especialistas com dois níveis, [HAYKIN, 2001].

4.3 Sistema Proposto Utilizando SVM e Máquinas de Comitê

A Figura 4.13 apresenta o diagrama em blocos simplificado do sistema de reconhecimento de voz proposto neste trabalho. Com objetivo de validar este sistema, o foco deste trabalho foi direcionado ao reconhecimento de palavras (comandos) dentro do universo do modo dependente do locutor.

Inicialmente, o sistema hierárquico da Figura 4.13 faz o reconhecimento das vogais através do reconhecimento fonético *frame a frame* (Capítulo II, Figura 2.7). Em seguida ao reconhecimento da vogal o sistema faz o reconhecimento das consoantes que existem antes e depois da vogal através da utilização de difones.

Devido a pouca duração das consoantes e a conseqüente influência da vogal sobre o sinal consonantal [MAIA, 1998], o reconhecimento consonantal é feito através da utilização de descritores que representam um difone (consoante + vogal ou vogal + consoante). Como existem doze vogais (cinco nasais e sete orais) diferentes na língua Portuguesa, doze sistemas iguais de reconhecimento de consoantes são necessários, teoricamente.

Todos os doze sistemas de reconhecimento consonantal são treinados da mesma forma e com as mesmas consoantes (Capítulo II, Figura 2.5), mas com difones formados por diferentes vogais. Exemplificando, se o bloco de reconhecimento de vogais detectar que a vogal é um /a/, então o decisor da vogal irá direcionar o sistema para o bloco de reconhecimento de consoantes composto pela vogal /a/.

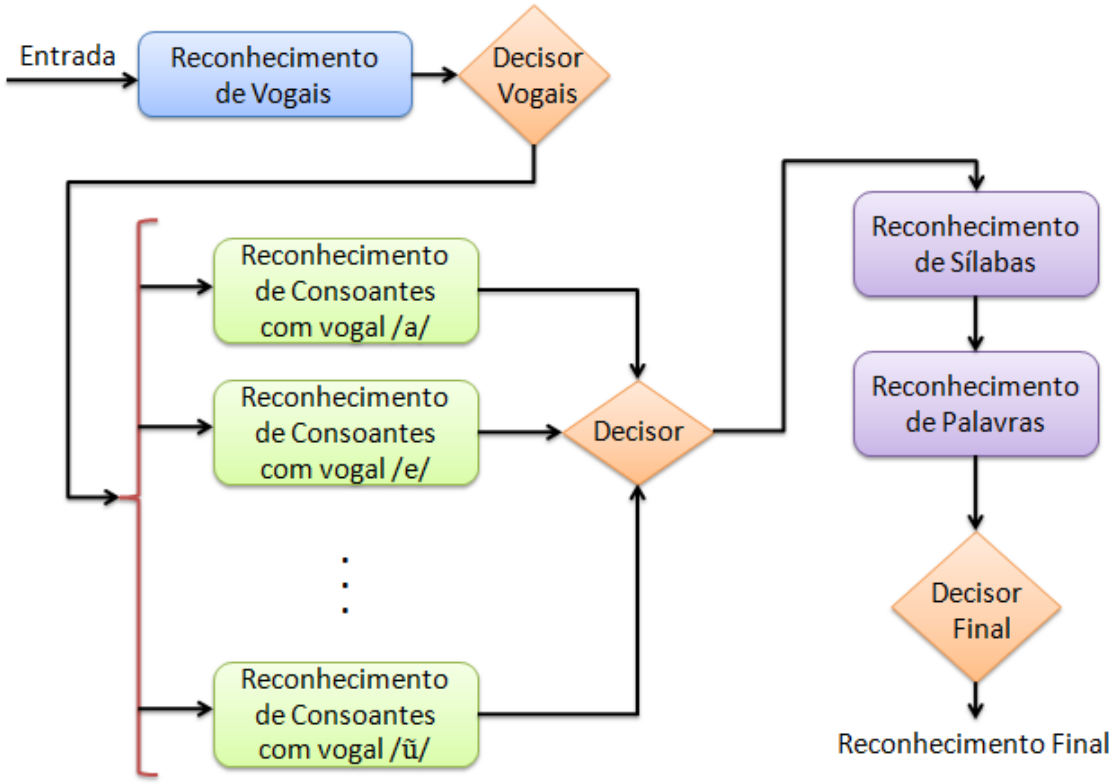

Figura 4.13: Diagrama de blocos simplificado do sistema de reconhecimento de voz proposto.

Uma vez que a vogal e as consoantes posterior e anterior sejam identificadas, então o próximo passo é o reconhecimento da sílaba. Os passos anteriores são repetidos para todas as sílabas existentes na palavra, deste modo o bloco de reconhecimento de palavras é responsável pela junção das sílabas reconhecidas anteriormente. Por fim, um decisor final será responsável pela identificação correta da palavra. No Capítulo V é apresentado o diagrama lógico do sistema desenvolvido com todos os detalhes deste sistema.

4.4 Análise Estatística da Decisão

A análise de decisão não é uma teoria descritiva ou explicativa, uma vez que não faz parte de seus objetivos descrever ou explicar como e por que as pessoas ou sistemas agem

de determinada forma ou tomam certas decisões. Pelo contrário, trata-se de uma teoria prescritiva ou normativa no sentido de pretender ajudar os sistemas a tomar decisões melhores face as suas preferências básicas [BEKMAN, 1980].

Para que um problema de decisão possa ser formulado, é necessária uma descrição completa do problema, que compreende as seguintes informações:

- A relação de todas a opções possíveis, seja como referência aos possíveis cursos de ação, seja a respeito da coleta ou aquisição de novas informações.
- A lista de todos os eventos que podem ocorrer como resultado das possíveis decisões.
- A cronologia em que as informações chegam ao conhecimento do decisor e em que as decisões devem ser tomadas.
- A quantificação das preferências do decisor em relação às consequências que podem resultar dos possíveis cursos de ação.
- Um julgamento probabilístico a respeito da ocorrência dos possíveis eventos.

A maior parte das informações necessárias à formulação do problema é de natureza objetiva, mas algumas, tais como a estrutura básica de preferências do decisor e seus julgamentos probabilísticos, são essencialmente subjetivas.

4.4.1 Probabilidade Condicionada

O cálculo das probabilidades é um ferramental matemático que se presta ao estudo de fenômenos aleatórios ou probabilísticos. Nestes fenômenos, o resultado de um experimento não pode ser previsto com certeza, mas, é em geral, possível relacionar todos os resultados possíveis de ocorrer.

Chama-se de espaço amostral (S) ou espaço de probabilidades ao conjunto de todos os possíveis resultados de um experimento aleatório. Qualquer resultado referente a um experimento aleatório pode ser descrito com um subconjunto do espaço amostral S, estes subconjuntos são chamados de eventos. Por exemplo, para o caso do reconhecimento das vogais do Português, descrito pela Figura 2.7, tem-se um espaço amostral composto por 12 possíveis eventos ou fonemas /a/, /ɛ/, /i/ /ɔ/, /u/, /e/, /o/, /ã/, /ẽ/, /ĩ/, /õ/ e /ũ/. Intuitivamente, a probabilidade de um evento é uma medida da certeza a respeito de sua

ocorrência, ou seja, representa o grau de crença no resultado, podendo ser de natureza objetiva ou subjetiva. A definição básica de probabilidade é feita através do quociente entre o número de resultados para os quais o evento em questão se verifica e o número de todos os resultados possíveis conforme a Equação (4.37) a seguir,

$$P(A) = \frac{m}{n} \tag{4.37}$$

onde $\boldsymbol{m}$ é o número de resultados favoráveis ao evento $\boldsymbol{A}$, e $\boldsymbol{n}$ é o número de resultados possíveis. O estabelecimento de uma probabilidade está, em geral, diretamente relacionado como o estado da informação disponível. Sendo $P(A)$ a probabilidade da ocorrência de um evento A, atribuída apenas ao conhecimento da mecânica do experimento, se houver a informação de que outro evento B ocorreu, então a probabilidade do evento A deve ser reavaliada por $P(A|B)$, ou seja, a probabilidade de A condicionada a B dada por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad para\ P(B) \neq 0 \tag{4.38}$$

Analogamente, é claro que pode-se ter,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad para\ P(A) \neq 0 \tag{4.39}$$

Das expressões acima, (4.38) e (4.39), pode-se facilmente resultar na regra do produto que permite o cálculo da probabilidade do evento de interseção, dada por:

$$P(A \cap B) = P(A).P(B|A) = P(B).P(A|B) \tag{4.40}$$

Se $P(A|B) = P(A)$, o evento A é dito estatisticamente independente do evento B. Isto implica que B também será estatisticamente independente de A. Para o caso de independência estatística, a regra do produto se simplifica conforme Equação (4.41),

$$P(A \cap B) = P(A).P(B) \tag{4.41}$$

4.4.2 Teorema de Bayes

Sejam $A_1, A_2, \ldots, A_n$ eventos mutuamente exclusivos e exaustivos (constituindo, pois, uma partição) e B um evento qualquer do espaço amostral S, então esses eventos podem ser simbolicamente representados num diagrama de Venn, no qual supõe-se que a área

correspondente a cada evento é numericamente igual à sua probabilidade conforme mostra a Figura 4.14.

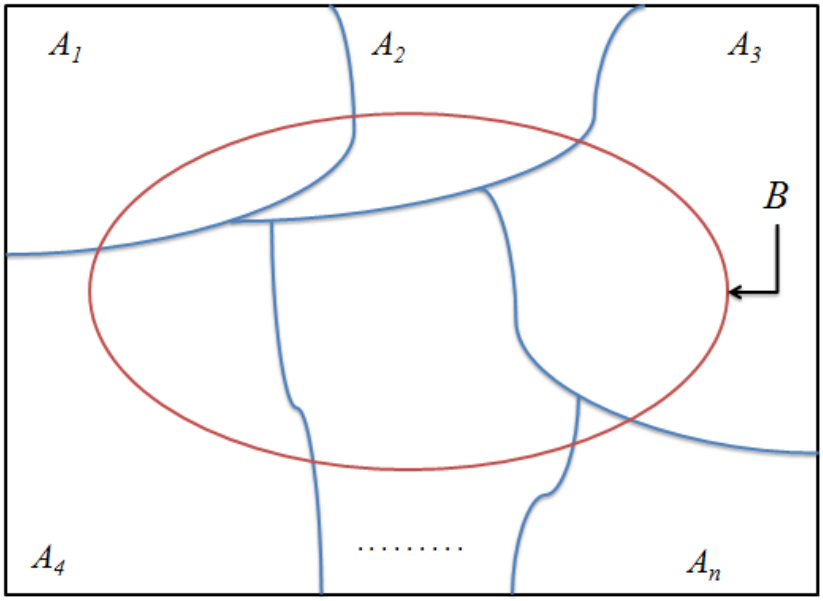

Figura 4.14: Diagrama de Venn

Pode-se obter a probabilidade de um evento particular A_k dada por $P(A_k|B)$ pela aplicação direta da Equação 4.38, ou seja,

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} \tag{4.42}$$

A Equação acima pode ser reescrita da seguinte forma,

$$P(A_k|B) = \frac{P(A_k \cap B)}{\sum_{i=1}^{n} P(A_i \cap B)} \tag{4.43}$$

ou usando a Equação 4.39 pode-se escrever,

$$P(A_k|B) = \frac{P(A_k).P(B|A_k)}{\sum_{i=1}^{n} P(A_i).P(B|A_i)} \tag{4.44}$$

As três expressões descritas acima são equivalentes. Entretanto, a forma apresentada na Equação (4.44) define o Teorema de Bayes.

A importância do Teorema de Bayes se revela quando as probabilidades $P(A_i)$ são consideradas como sendo representativas de certo estado inicial de informação (*priori*), que se modifica (*posteriori)* tão logo chegue ao conhecimento do decisor a ocorrência de um evento B.

Neste trabalho, sobre reconhecimento de voz, pode-se afirmar que os blocos de decisão tanto para as vogais como para as consoantes (Figura 4.13) possuem números fixos de eventos e portanto sua probabilidade pode ser calculada de forma simples através da Equação 4.37. No entanto, quando o sistema é direcionado ao decisor final que fará a junção das respostas, tanto do subsistema de vogais como dos subsistemas de consoantes, o problema de decisão torna-se dependente e portanto o Teorema de Bayes pode ser utilizado para o cálculo das probabilidades deste decisor.

Outra forma de ver o problema de decisão é a utilização das árvores de probabilidades. Neste caso os dados originais são convenientemente apresentados por meio de grafos que formam uma árvore de decisão conforme a hierarquia do sistema proposto [BEKMAN, 1980]. Como as decisões das máquinas especialistas (SVM) são sempre binárias a árvore de decisão do sistema hierárquico proposto pode ser considerada como uma árvore de decisão binária, conforme mostra a Figura 4.15.

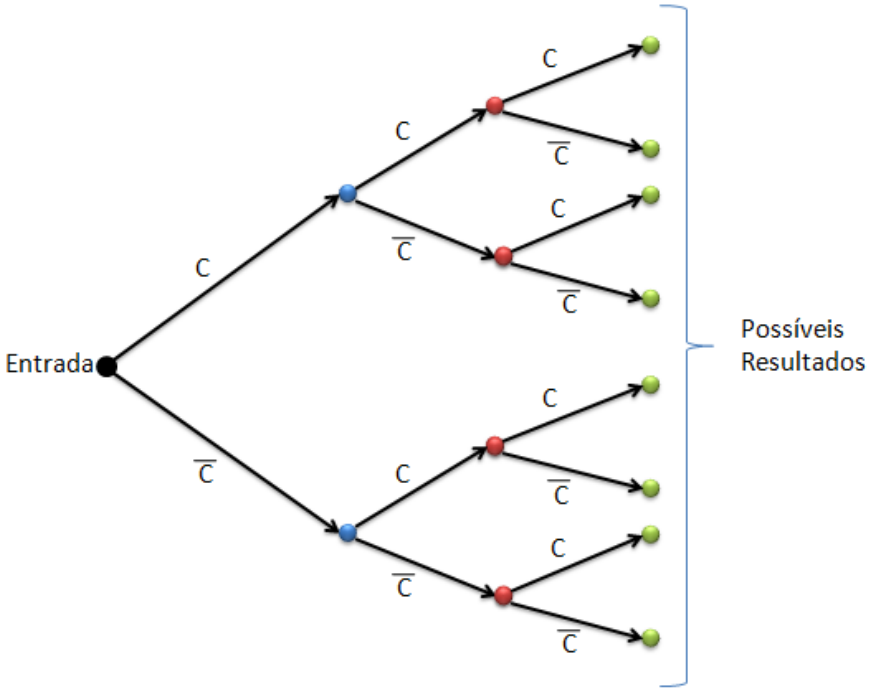

Figura 4.15: Árvore de decisão binária. Adaptada de [BEKMAN, 1980].

Esta árvore binária representa a cronologia real dos acontecimentos ou das decisões, isto é, a cada passo ou em cada especialista é possível apenas se ter duas opções de resposta (+1) representado por C na Figura 4.15, ou (-1) representado por C “barrado” ou o complemento de C.

4.5 Estado da Arte

A exemplo dos capítulos anteriores, esta seção visa descrever alguns trabalhos anteriores que justificam a utilização das ferramentas descritas neste capítulo.

Clarkson e Moreno [CLARKSON, 1999] usaram as Máquinas de Vetor de Suporte - SVM na classificação de fonemas. Os autores trabalharam com as vogais da língua inglesa com bons resultados.

Abe e Inoue [ABE, 2002] utilizaram a lógica Fuzzy como regra de decisão de um classificador SVM. Em todos os casos testados, os autores obtiveram um bom desempenho do classificador.

Deshmukh *et al* [DESHMUKH, 2002] utilizaram os parâmetros acústico-fonéticos no reconhecimento de voz. Os resultados apresentaram até 99,88% de acerto no reconhecimento de certos fonemas, utilizando o classificador SVM.

Picone [PICONE, 2002] usou um sistema híbrido com SVM no reconhecimento de voz. Este método apresentou uma melhoria significativa do desempenho, em comparação com o classificador base HMM.

Juneja e Espy-Wilson [JUNEJA, 2003] demonstraram a utilidade do SVM (*Support Vector Machine*) na classificação de fonemas, sendo que os resultados apresentaram desempenho superior ao modelo mais utilizado, o HMM.

Abdulla *et al* [ABDULLA, 2003] investigaram o uso do SVM como base para separar fonemas a partir de um discurso contínuo. A segmentação e extração de unidades menores do que a palavra dentro do discurso são um passo essencial do pré-processamento em muitos sistemas de reconhecimento de fala e de codificação. O SVM foi comparado com os modelos de Markov (HMM) e mostrou uma grande vantagem em termos de deteção de fonemas intra-palavra, bem como os períodos de silêncio.

Rafael Marinho *et al* [MARINHO,2004] usaram o classificador SVM na classificação de fonemas. Os autores utilizaram o banco de dados com corpus *TIMIT*. Os resultados indicaram que o SVM é eficiente, quando o espaço de entrada possui dimensão elevada, mesmo que os parâmetros sejam altamente redundantes.

Iosif Mporas *et al* [MPORAS,2006] fizeram um estudo da aplicabilidade do SVM no reconhecimento de fonemas na língua grega. Os autores compararam diferentes métodos

de classificação. O classificador SVM demonstrou um desempenho superior aos outros métodos testados.

Capítulo V

5. Sistema de Reconhecimento de Voz Hierárquico

Como citado no capítulo introdutório, a proposta deste trabalho é utilizar unidades menores do que a palavra tais como: fonemas, difones e sílabas, como unidades base para o reconhecimento da voz, utilizando a *Wavelet Packet* como principal descritor do sinal de voz e as Máquinas de Vetor de Suporte (SVM) como classificadores agrupadas em um sistema de decisão hierárquico.

O reconhecimento de voz de modo contínuo e independente do locutor é o objetivo principal dos sistemas de reconhecimento de voz. No entanto, neste trabalho limitou-se o foco ao reconhecimento de palavras isoladas e no modo dependente do locutor. O objetivo principal desta limitação é a validação desta nova proposta. Um estudo futuro poderá incluir a utilização deste novo sistema no reconhecimento contínuo da fala e no modo independente do locutor.

A seção inicial deste capítulo é dedicada as características físicas do som as quais influenciam diretamente nos parâmetros e definições do sistema como um todo. As demais seções são dedicadas a explanação do sistema desenvolvido incluindo: o pré-processamento do sinal, a extração dos descritores, o treinamento da máquinas, os experimentos realizados e os resultados obtidos.

5.1 Características Físicas do Som

Inúmeros são os parâmetros que devem ser regulados em um sistema complexo como o de reconhecimento de voz. Muitos destes parâmetros estão diretamente relacionados com as características físicas do som, tais como:

- Freqüência do sinal;
- Taxa de amostragem;
- Amplitude ou intensidade do sinal;
- Freqüência fundamental e formantes;
- Característica quase-estacionária do sinal de voz.

Além destes, outros fatores como a resolução e precisão da aquisição do sinal, equipamentos utilizados (*hardware* e *software*) afetam a qualidade da aquisição do sinal de voz e, conseqüentemente, o desempenho do sistema.

5.1.1 O Som

O som pode ser entendido como uma variação de pressão muito rápida, a qual se propaga na forma de ondas em um meio elástico. Qualquer corpo elástico capaz de vibrar rapidamente e conseqüentemente produzir som, recebe o nome de fonte sonora. Para o caso de um sistema de reconhecimento de voz a fonte sonora é o trato vocal do ser humano.

O som pode ser percebido através das variações de pressão do ar que atingem os ouvidos. Se essas variações ocorrem entre 20 e 20.000 vezes por segundo, então esse som é potencialmente audível mesmo que a variação de pressão seja de alguns milionésimos de Pascal. Assim, da definição da acústica, *Som* é toda perturbação do ar capaz de estimular o aparelho auditivo [MAIA, 1985].

5.1.2 Freqüência do Sinal e Taxa de Amostragem

A freqüência do sinal de voz é o principal determinante da sensação de altura do som, ou seja, das variações entre o grave e o agudo que nosso ouvido distingue. No entanto, a relação entre a freqüência e a altura do som não é das mais simples. Efeitos como o ruído de ambiente afetam significativamente a percepção e conseqüente distinção entre sons graves e agudos.

Além disso, a relação com a freqüência é linear para freqüências abaixo de 1000 Hz e logarítmica para freqüências superiores. Isso quer dizer que abaixo de 1000 Hz há uma correspondência termo a termo entre as diferenças de freqüência e as diferenças de altura, de tal forma que um tom de 600 Hz difere de um tom de 700 Hz tanto quanto de outro de 500 Hz. Em contrapartida, acima de 1000 Hz o intervalo entre dois tons depende da razão entre as suas freqüências, isto que dizer que um tom de 4000 Hz difere de um de 2000 Hz, da mesma forma que difere de outro tom na faixa de 8000 Hz [MAIA, 1985].

A taxa de amostragem do sinal de voz deve ser condizente com a faixa de freqüência do sinal de voz, ou seja, deve satisfazer o Teorema da Amostragem [LATHI, 1987]. O Teorema de Amostragem diz que a freqüência de amostragem F_s deve ser maior ou igual ao

dobro da freqüência máxima do sinal ($F_{máx}$) conforme a Equação (5.1). Se este limite não for respeitado haverá o entrelaçamento do espectro do sinal e conseqüente perda de características importantes.

$$F_s = 2.F_{máx} \tag{5.1}$$

5.1.3 Amplitude do Sinal de Voz

Outro fator importante a ser considerado é a relação da amplitude do sinal de voz com a intensidade. Na verdade a intensidade auditiva, isto é, a energia sonora detectada pelo receptor (ouvido) é função da intensidade física. O sistema auditivo tem dois limites de audibilidade: Limiar de audibilidade (mínima intensidade audível) e Limite de dor (máximo nível de intensidade audível sem danos fisiológicos ou dor).

A gama entre este dois limites é muito grande. Por exemplo: para uma freqüência pura de 1000 Hz, esses limites vão de 10^{-12} watt/m^2 à 1 watt/m^2, ou seja, uma razão de 1 trilhão para 1. Numericamente, a referência em watt/m^2 não é confortável. Para resolver este problema foi introduzida uma razão de compressão logarítmica, o decibel (dB). O decibel é uma relação logarítmica entre duas potências ou intensidades, dada pela Equação (5.2):

$$dB = 10 \log_{10} \frac{I_1}{I_2} \tag{5.2}$$

onde I_1 representa o valor da intensidade a ser transformada e I_2 o valor de referência utilizado (10^{-12} watt/m^2). Assim, o nível de intensidade sonora é melhor representado em uma escala em decibéis do que em watt/m^2. A Tabela 5.1 define os limiares de audibilidade e do limite de dor na escala em decibéis.

Tabela 5.1: Limiares de audibilidade do som.

Limiar	Valores em dB
Limiar de Audibilidade	10 log (10^{-12}/10^{-12}) = 10 log 1 =0 dB
Limiar de dor	10 log (1/10^{-12}) = 10 log 10^{12} = 120dB

Vale salientar ainda que, a cada 3dB a intensidade dobra. Graças a escala em decibéis pode-se reduzir a faixa de audição do ser humano a números inteligíveis. A Figura 5.1 mostra a relação entre a freqüência do som que o ser humano pode ouvir, com amplitude em decibéis. Nota-se que a região que compreende a área da fala está situada

aproximadamente entre as freqüências de 90 a 8500 Hz dentro uma escala variável entre as faixas de 40 a 90 dB.

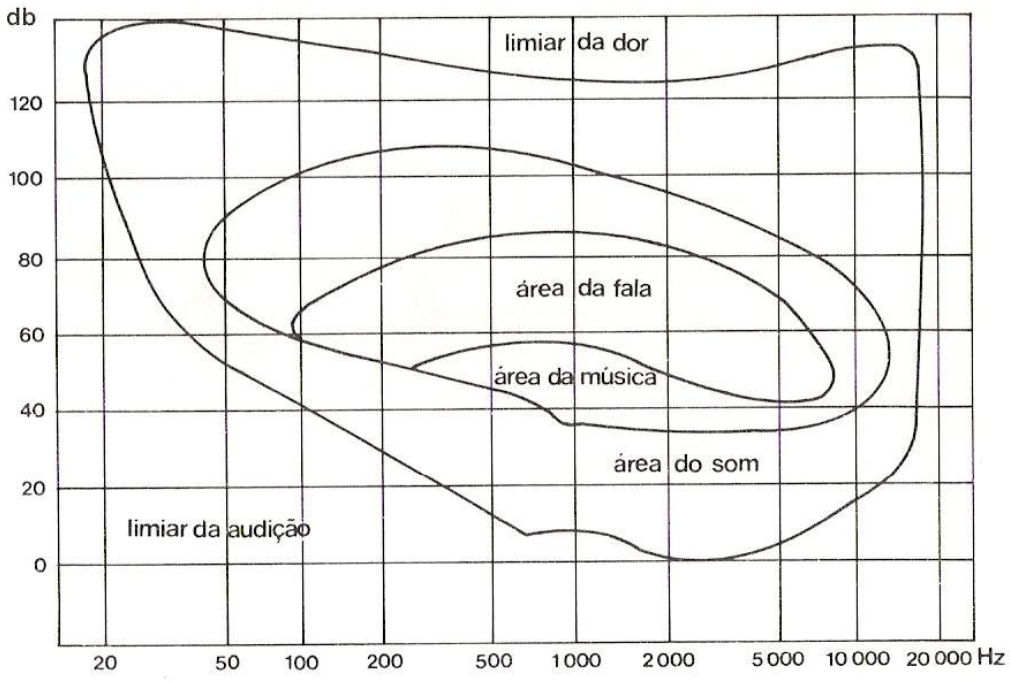

Figura 5.1: Limites de freqüência e intensidade para a audição e a fala humana. (Maia, 1985).

5.1.4 Freqüência Fundamental, Formantes e Estacionaridade do Sinal

Da física elementar, pode-se afirmar que quando uma fonte sonora vibra, tal vibração pode ser decomposta numa série de movimentos harmônicos simples. Esta decomposição é classicamente demonstrada pelas séries de Fourier [LATHI, 1987]. As componentes de freqüência do sinal de voz são múltiplos inteiros da freqüência com maior energia, a qual é chamada de fundamental (f_0). A freqüência fundamental, também chamada de *pitch*, é a componente de freqüência com maior energia do pulso de ar gerado no sistema laríngeo [MAIA, 1985].

Os falantes masculinos adultos apresentam, em média, valores de freqüência fundamental em torno de 120 Hz. Já os falantes femininos adultos tem a média de 220 Hz e para as crianças a média fica em aproximadamente 300 Hz. A relação entre essa freqüência e a fisiologia das cordas vocais leva a concluir que a freqüência fundamental é um importante parâmetro comparativo na individualização do falante [RUSSO, 1993].

Já as freqüências formantes podem ser consideradas como efeitos ressonantes no trato vocal relacionados à amplificação da energia do som no subsistema supralaríngeo. As

freqüências formantes estão relacionadas à anatomia e às configurações específicas do aparelho fonador de cada indivíduo. A freqüência do primeiro formante (F_1) está relacionada à posição da língua no plano vertical e é influenciado pelo grau de abertura da boca, enquanto a freqüência do segundo formante (F_2) está relacionada à posição da língua no plano horizontal (grau de anterioridade). As demais freqüências dos formantes estão relacionadas à geometria do trato vocal e são, dentro de um conjunto de convergências, fortes elementos para a individualização do falante [MAIA, 1985].

Os formantes são particularmente importantes na determinação da fala. De certo modo, a formação das vogais se dá praticamente pela alteração das regiões formânticas do aparelho fonador. As três freqüências formantes importantes são chamadas de F_1, F_2 e F_3. Estas formantes situam-se em freqüências abaixo de 3500 Hz [RABINER, 1993].

Por fim, mas não menos importante, deve-se citar um conceito fundamental no reconhecimento de voz que a característica quase-estacionária do sinal de voz. Este conceito diz que dentro do limite de uma janela com tamanho de no máximo 30ms o sinal de voz pode ser considerado quase-estacionário [RABINER, 1975].

5.2 Diagrama Geral do Sistema

A utilização de unidades menores do que a palavra (fonemas, difones e sílabas) está baseada na metodologia básica da engenharia: *dividir para conquistar*. Portanto, uma estrutura parecida com as Máquinas de Comitê do tipo mistura hierárquica de especialistas é utilizada para construir o sistema de reconhecimento de voz apresentado neste trabalho.

A lógica geral do sistema desenvolvido baseia-se no reconhecimento da palavra. Para fazer este reconhecimento, o sinal da palavra é dividido em sinais menores que representam as sílabas. Para cada sinal que representa uma sílaba, é feita a identificação da vogal ou vogais presentes no sinal e, em seguida, é identificada a existência ou não de consoantes antes e depois das vogais. Uma vez que cada sílaba seja identificada, faz-se a junção destas informações para a identificação final da palavra.

Para desenvolver esta lógica, foi necessário construir um sistema de decisão hierárquico o qual é apresentado na Figura 5.2. A formulação deste sistema hierárquico de decisão obedece às características fonéticas da produção da voz, tais como modo e lugar de

articulação para as consoantes, e as posições vertical e horizontal da língua para o caso das vogais e semivogais (Capítulo II).

Os principais descritores do sinal de voz são obtidos através da transformada *Wavelet Packet* com as bandas selecionadas com o auxílio da escala Mel. A *Wavelet* mãe utilizada em todos os casos foi a *Daubechies* 5. Os descritores MFCC são utilizados nos casos de reconhecimento de padrões com baixa energia, tais como as consoantes fricativas (Capítulo III, item 3.4.1).

Cada decisão do sistema é feita através de uma ou mais máquinas especialistas. Estas máquinas são agrupadas em sistema com características aproximadas do sistema de Mistura Hierárquica de Especialistas – MHE (Figura 4.12). Além disso, estas máquinas são constituídas por uma rede neural (SVM) a qual é responsável pelo reconhecimento de um padrão ou de vários padrões específicos (Capítulo IV). A Figura 5.2 apresenta o diagrama lógico do sistema desenvolvido.

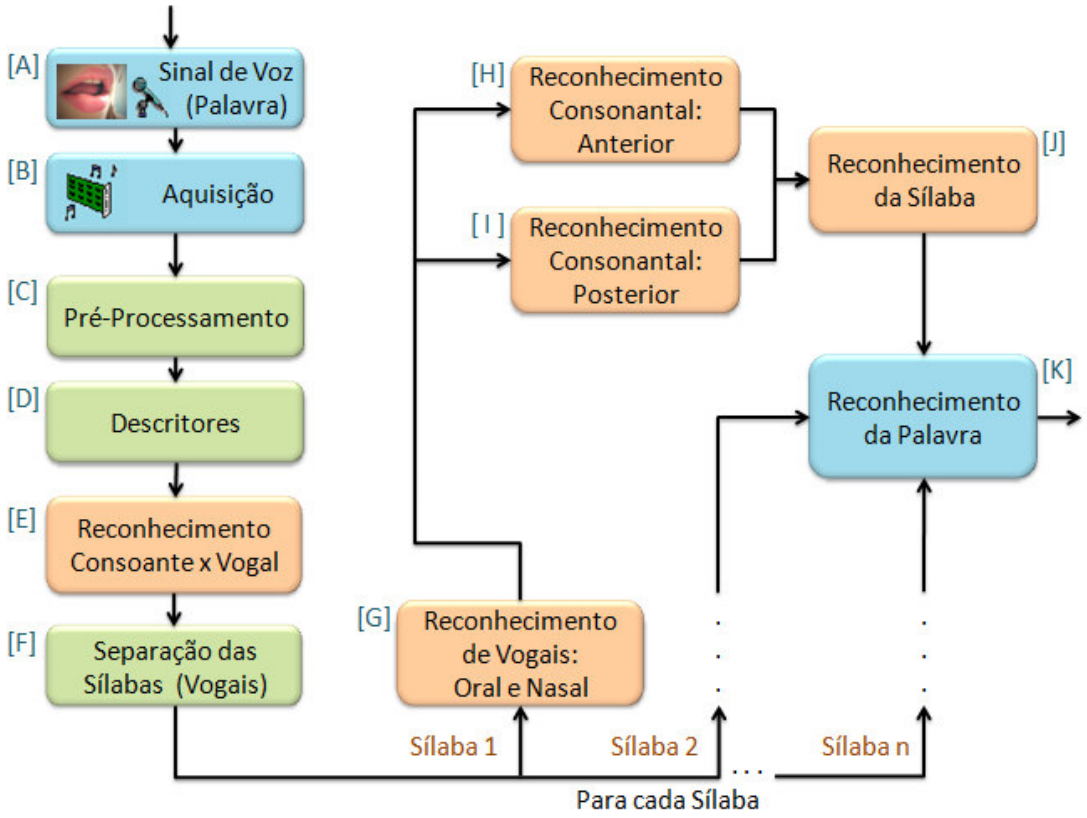

Figura 5.2: Diagrama de blocos lógico do sistema de reconhecimento proposto.

Novamente, vale salientar que o sistema apresentado tem as seguintes restrições ou limitações: Reconhecimento de palavras isoladas e modo dependente do locutor. Além

disso, todo o sistema de decisão foi baseado na fonologia e fonética da língua Portuguesa do Brasil. Estas restrições foram impostas com objetivo de validar o sistema desenvolvido.

Os blocos da Figura 5.2 (A, B, …, K) serão detalhados nas secções a seguir. O objetivo é explicar a seqüência lógica do sistema. Alguns destes blocos representam um sub-sistema interno, como é o caso dos blocos de reconhecimento de Vogais e Consoantes. Outros blocos representam apenas o processamento do sinal, como é o caso dos blocos de Aquisição, Pré-processamento e Descritores.

Para melhor descrever cada passo (ou bloco) o sinal da palavra "pare" será utilizada como exemplo no decorrer das próximas seções. Os experimentos realizados como o objetivo de validar cada um dos blocos de reconhecimento, bem como os resultados correspondente, serão apresentados no decorrer de cada seção. Todos os sistemas foram desenvolvidos na plataforma do Matlab® 6.0.

5.3 Aquisição e Gravação do Sinal de Voz

Com o objetivo de facilitar a tarefa de comparação do sinal original com o sinal reconhecido, cada palavra gravada foi rotulada com um código numérico por sílaba.

Os doze fonemas vocálicos (vogais) foram rotulados com uma numeração de 11 a 22 sendo a sequência respectiva dada por: /a/, /ɛ/, /i/, /ɔ/, /u/, /e/, /o/, /ã/, /ẽ/, /ĩ/, /õ/ e /ũ/.

Já os 21 fonemas consonantais (consoantes) foram rotulados com códigos de 101 a 121 representando respectivamente as seguintes consoantes: /p/, /b/, /t/, /d/, /k/, /g/, /f/, /v/, /s/, /z/, /ʃ/, /ʒ/, /h/, /ɾ/, /L/, /ʎ/, /m/, /n/, /ɲ/, /tʃ/ e /dʒ/. Esta seqüência foi propositalmente estipulada com o objetivo de juntar os conjuntos de consoantes sendo: oclusivas, fricativas, erres, laterais e nasais.

Vale lembrar que os símbolos fonéticos /ʃ/, /ʒ/, /h/, /ʎ/, /ɲ/ representam os sons das letras ou encontro de letras [ch ou x], [J], [rr], [lh] e [nh] respectivamente. Além disso, por efeito de simplificação os fonemas /tʃ/ e /dʒ/ não foram utilizados nos testes, devido a sua característica dialética de representar somente as sílabas [di] e [ti] respectivamente em algumas partes do País.

Portanto, 19 consoantes são utilizadas nos testes de reconhecimento. Deste modo, uma sílaba formada por um conjunto CVC (Consoante + Vogal + Consoante), como por

exemplo [par] ou /paɾ/, seria representada pelo código numérico dado por [101 11 114] representando os fonemas /p/, /a/ e /ɾ/ respectivamente.

5.3.1 Microfone e parâmetros de gravação utilizados

O mundo dos microfones é extenso devido a uma grande variedade de tipos, modelos e marcas existentes hoje em dia no mercado. No entanto, a especificação de todos eles gira em torno das mesmas características: ganho em freqüência (+ ou - 3 dB), sensibilidade (+ ou - captação segundo a intensidade), qualidade e resposta em freqüência. Microfone é o termo genérico utilizado para falar dos elementos que transformam a energia acústica (som ou voz) em energia elétrica (um sinal de áudio digital ou analógico).

Com o objetivo de padronizar a aquisição do sinal de voz, em todas as gravações foi utilizado o mesmo microfone omnidirecional acoplado a um fone de ouvido com as seguintes características:

- Ação Diretiva: Omnidirecional;
- Impedância: 2,2 ohms;
- Sensibilidade: 60 dB +- 3 dB em 1 kHz;
- Freqüência de resposta: 50 até 13 kHz.

A Figura 5.3. apresenta um exemplo do microfone utilizado na gravação do sinal de voz. Uma vez que o sinal de voz apresenta uma freqüência máxima em torno de 8 kHz podendo chegar a alguns casos a 8,5 kHz, a freqüência de amostragem do sinal foi fixada em 22050 Hz. Deste modo, o limite imposto pelo teorema da Amostragem é respeitado. Assim, para cada segundo de gravação, 22050 amostras são adquiridas. A resolução do conversor A/D (analógico/digital) é 16 bits.

Figura 5 3: Exemplo de microfone utilizado para gravação da voz.

5.3.2 Amostragem

O primeiro passo na aquisição do sinal de voz (Figura 5.2-B) é a amostragem. Neste procedimento o sinal analógico é convertido em números binários que o computador ou processador digital possa entender. O processo de transformar valores da amplitude de um sinal contínuo em valores discretos tomados de um conjunto finito de amplitudes é chamado de quantização [LATHI, 1987]. Durante a quantização, o conversor A/D usa um número finito de valores uniformemente espaçados para representar o sinal analógico. O número de valores diferentes é determinado pelo número de bits usados para a conversão. A maioria dos conversores A/D trabalha com 12 ou 16 bits. Tipicamente, o conversor A/D seleciona o valor digital que é mais próximo do valor amostrado. O número de *bits* que representam um sinal analógico determina a precisão (resolução) do dispositivo. Quanto mais *bits* tiver a placa de aquisição de dados mais precisa será a medida, assim:

$$Precisão = \frac{Faixa\ Dinamica\ [V]}{2^{Nbits}} \tag{5.3}$$

Por exemplo, usando um conversor A/D de 12 bits, para uma faixa dinâmica de 12 volts, a precisão será de 2,44 mV. Isto significa que o conversor pode detectar corretamente diferentes tensões até o nível de 0.00244 Volts.

A Figura 5.4 apresenta o sinal de voz da palavra “pare” gravado com duração de 2 s.

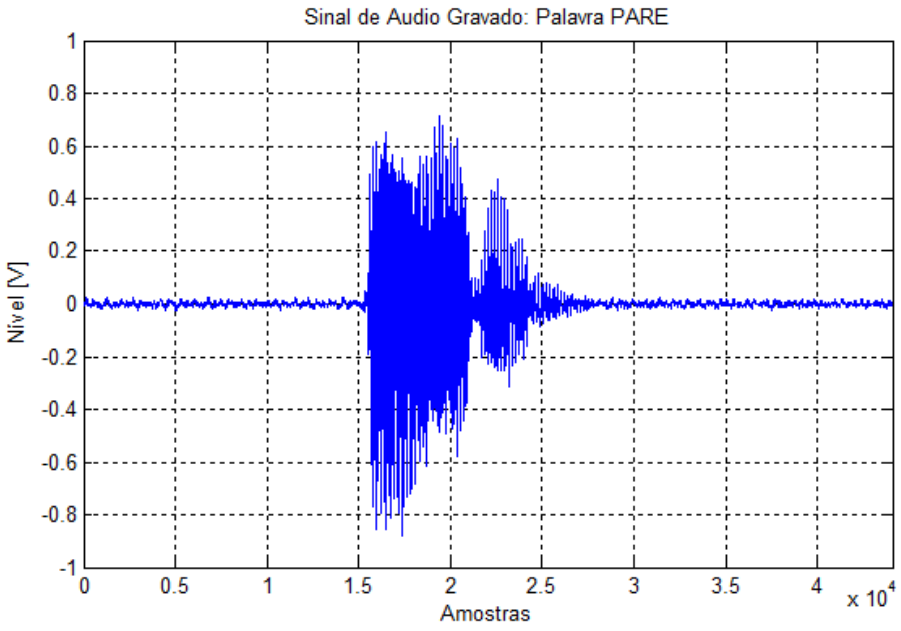

Figura 5.4: Sinal de voz da palavra “pare” adquirida com freqüência de 22050Hz.

Como cada segundo de amostragem gera 22050 amostras, então em 2 segundos teremos 44100 amostras adquiridas.

Pode-se notar na Figura 5.4, que existe um sinal de fundo (parte sem voz: silêncio) com amplitude de aproximadamente 0,04V. Esta amplitude é devida principalmente ao ruído da rede elétrica de 60Hz. Vale salientar que este ruído não aparecerá se a gravação for efetuada em um sistema desconectado da rede elétrica, como por exemplo, um computador notebook funcionando apenas através da bateria.

5.4 Pré-processamento do Sinal

Após a aquisição do sinal, o próximo passo na lógica do sistema é o pré-processamento (Figura 5.2-C). Esta etapa é fundamental tanto no treinamento quanto na classificação do sinal de voz. O sinal gravado na etapa de aquisição (Figura 5.2-A e B) é pré-processado através dos seguintes procedimentos: filtragem em freqüência, filtragem no tempo (pré-ênfase), normalização, separação de *background* (separação do som da voz do som de fundo) e janelamento do sinal.

5.4.1 Filtragem, Pré-ênfase, Normalização e Truncagem do Sinal

O sinal de voz foi adquirido no modo *stereo*, mas somente um canal foi utilizado. A amostragem é feita a uma freqüência de 22050 Hz, sendo portanto a largura de banda de 11025 Hz, conforme o teorema da amostragem. Um determinado sinal de voz $\boldsymbol{x}_{\mathrm{i}}$ pode ser então representado pela Equação (5.4) dada por:

$$x_i = [x_0, x_1, x_2, \dots, x_N]^T \tag{5.4}$$

onde $\boldsymbol{x}_{\mathrm{i}}$ é o vetor que representa o sinal de voz adquirido e *N* é o tamanho da amostra. Para o exemplo da palavra “pare” da Figura 5.4 o valor de *N* é igual a 44100.

Inicialmente, o sinal de voz é passado por um filtro passa-banda com freqüência de corte de 80Hz e 9 kHz, o objetivo deste filtro é extrair do sinal freqüências acima de 9 kHz e eliminar o ruído da rede elétrica (60 Hz). Em seguida, é aplicada ao sinal a filtragem de pré-enfase. Este filtro é utilizado para equalizar o espectro do sinal de voz e melhorar o desempenho da análise espectral [YANG, 2001]. A filtragem de pré-ênfase, feita no domínio do tempo, é dada pela Equação (5.5), sendo:

$$x_i = x_i - b.x_{i-1} \qquad (0{,}9 \leq b \leq 1) \qquad (5.5)$$

o valor empírico estipulado de b neste trabalho foi b= 0,97.

O próximo passo é a normalização. O sinal de voz é normalizado pelo máximo valor da sua própria amplitude, conforme a Equação (5.6).

$$xnorm_i = \frac{x_i}{max|x_i|} \qquad (5.6)$$

onde $xnorm_i$ representa o sinal de voz normalizado.

A Figura 5.5 apresenta as operações do pré-processamento sobre o sinal da palavra "pare" utilizada aqui como exemplo. O sinal original é repetido por conveniência em (A), o sinal filtrado é apresentado em (B), o sinal após a pré-ênfase em (C) e o sinal normalizado e truncado em (D).

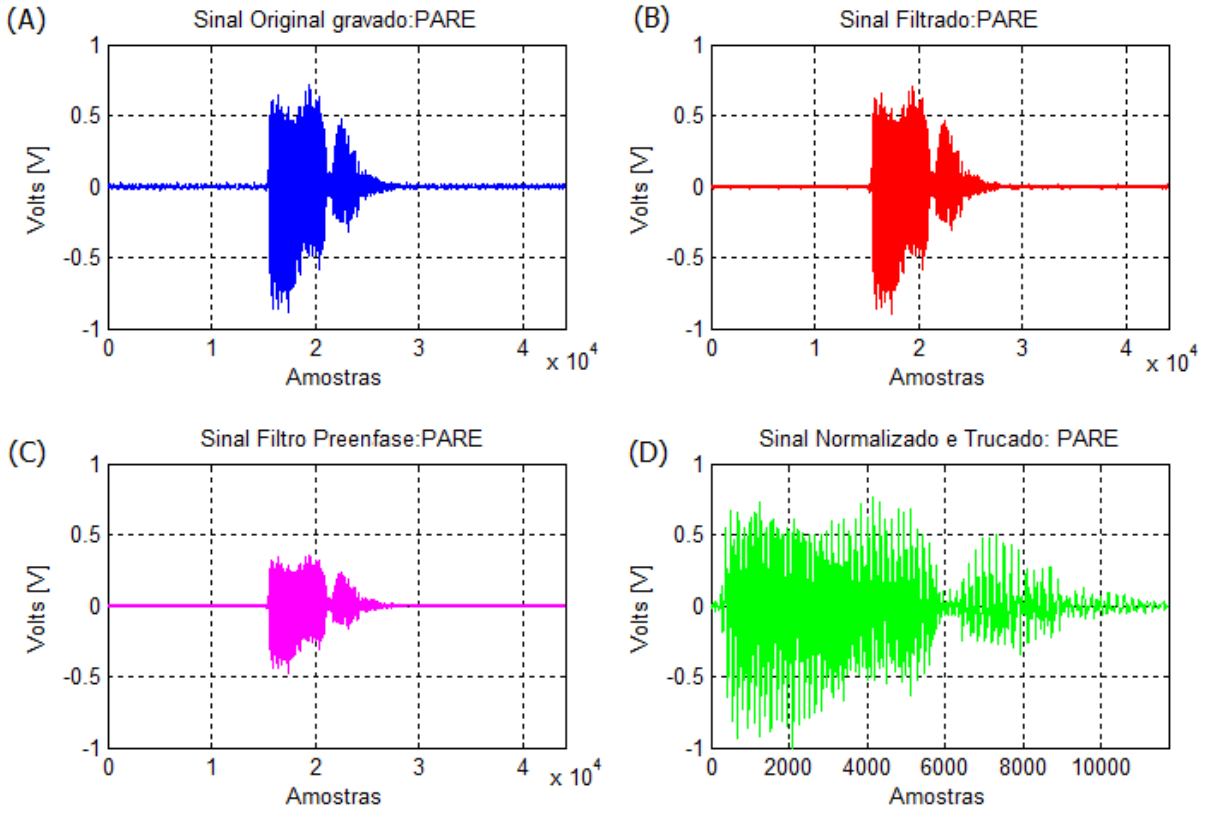

Figura 5.5: (A) Sinal de Voz gravado, (B) filtragem, (C) Pré-ênfase e (D) Normalização e truncagem do sinal.

Pode-se notar, na Figura 5.5-C, que o ruído de fundo praticamente desapareceu após as filtragens de passa-faixa e pré-ênfase. Por fim, a Figura 5.5-D apresenta o sinal de voz

normalizado e truncado. A truncagem do sinal tem por objetivo separar o sinal de voz do sinal de fundo (ruído ou silêncio).

A Figura 5.6-A apresenta a sinal de voz da palavra “pare” após a filtragem de pré-ênfase sendo limitado por duas linhas pontilhadas verticais em vermelho. Estas linhas representam o limite de truncagem do sinal, ou seja, as fronteiras entre o sinal de voz e o sinal de fundo (silêncio). A Figura 5.6-B mostra um *zoom* do sinal exatamente no ponto de fronteira inicial (truncagem).

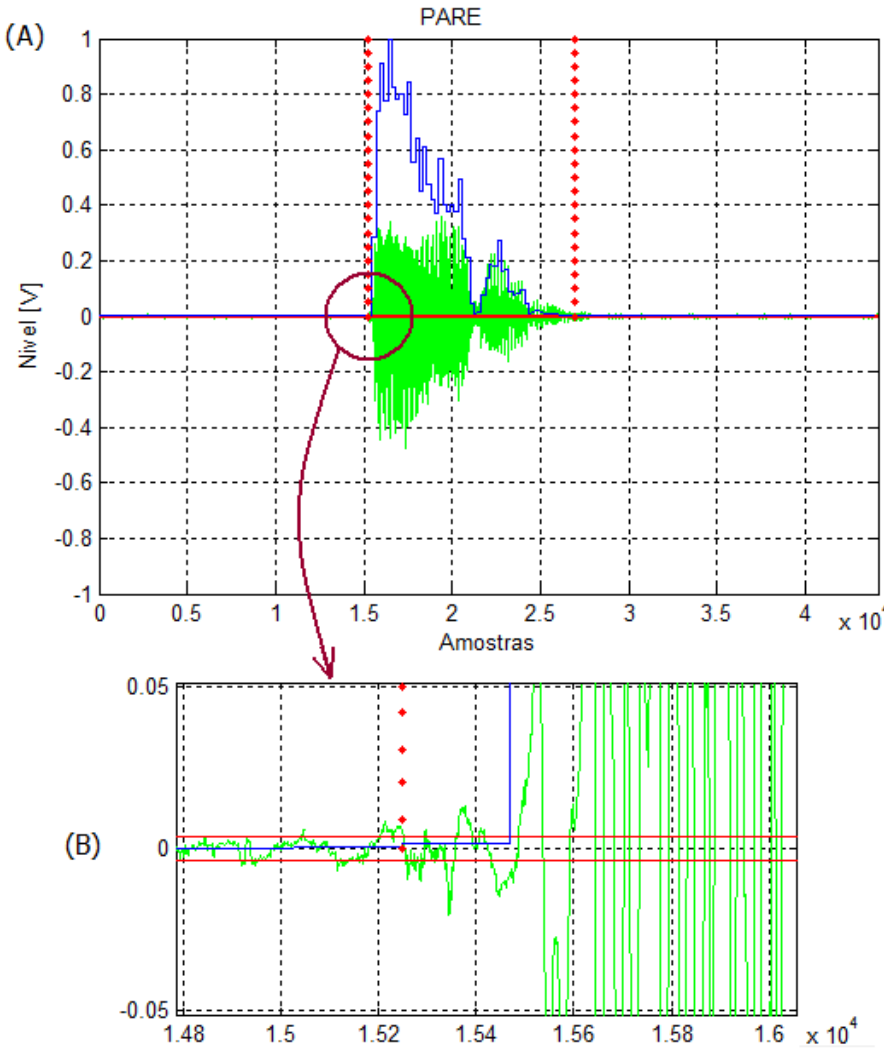

Figura 5.6: (A) Truncagem do sinal através da energia, (B) zoom fronteira da trucagem.

O método de truncagem do sinal de voz é simples, basicamente, após a filtragem de pré-ênfase divide-se o todo o sinal em janelas de 5ms e calcula-se a energia de cada janela. Na Figura 5.6 (A e B), este vetor energia é representado pela linha em azul. Como existem muitas janelas (cerca de 400 janelas em 2s) e grande parte destas janelas representam o

sinal de fundo (ou silêncio), então a mediana do vetor de energia representará aproximadamente, a energia do sinal de fundo.

Assim, um valor de três vezes a mediana do vetor energia é definido como um limiar empírico que possibilita a separação automática do sinal de voz do sinal de fundo. Na Figura 5.6-B (*zoom*), podem se vistas duas linhas contínuas horizontais vermelhas que marcam este limiar. Deste modo, todas janelas que possuírem energia (linha azul) maior que este limiar são consideradas janelas do sinal de voz. O tamanho da janela foi estipulado em 5ms para evitar a perda de partes importantes do sinal de voz, como as consoantes oclusivas que possuem pouco tempo de duração. Além disso, os limites de truncagem são fixados, uma janela antes no início e uma janela depois no final, com este mesmo objetivo.

O uso da energia do sinal para truncagem e a conseqüente separação do sinal de voz do sinal de fundo (ou a detecção do início e fim de uma expressão) em uma gravação foi inspirado nos trabalhos de [RABINER, 1975], [BISHNU, 1976], [ZHU, 2000] e [YANG, 2001].

5.4.2 Janelamento do Sinal

Após a truncagem do sinal, somente a parte correspondente à voz propriamente dita é utilizada (Figura 5.5 D). A partir deste ponto o sinal de voz está preparado para extração dos descritores. Considera-se que a estacionaridade do sinal de voz esteja na faixa de 10 a 30 ms [RABINER, 1975], daí a necessidade de "janelamento" (*frames*) do sinal. Com objetivo de uniformização e redução da quantidade de processamento e armazenamento, a largura da janela foi estipulada em 30ms com superposição de 1/3. Segundo Stephen Levison [LEVISON, 2005] a superposição de ¼ ou 25% é matematicamente a ideal. No entanto, nos experimentos deste trabalho a superposição de 1/3 (33,3%) apresentou um custo benefício melhor devido à redução do números de janelas e a conseqüente diminuição do tempo de processamento sem perda significativa de desempenho. Para efeito de arredondamento, o tamanho de cada janela foi fixado em 660 amostras (aproximadamente 30 ms). Deste modo, o passo de superposição foi então definido a cada 220 amostras (aproximadamente 10 ms). Ao final do processo de janelamento tem-se uma matriz *K* x *M,* onde *K* representa o número de janelas e *M* é o tamanho da janela. Cada janela do sinal $y_{k,j}$ é então representada por:

$$y_{k,j} = x_{V.k+j} \qquad k = 0,1,\dots,K-1 \quad j = 0,1,\dots,M-1 \qquad (5.7)$$

onde "v" é o passo de superposição das janelas.

Com objetivo de reduzir as descontinuidades do sinal, cada *frame* foi multiplicado pela Janela de *Hamming* (w_j), que é representada pela Equação 5.8. A Equação (5.9) representa esta multiplicação.

$$w_j = 0{,}54 - 0{,}64.\cos\left(\frac{2.\pi.j}{M}\right) \qquad j = 1{,}2, \ldots, M \tag{5.8}$$

$$y_{k,j} = y_{k,j}.w_{j,1} \tag{5.9}$$

A partir deste ponto, para cada janela do sinal $y_{k,j}$ serão calculados os descritores do sinal tanto da transformada *Wavelet Packet* como dos coeficientes MFCC.

5.5 Extração dos Descritores

A extração dos descritores, bem como a sua utilização no treinamento e classificação, serão explanadas nesta seção. Este passo refere-se ao bloco "D" da Figura 5.2. Basicamente, três conjuntos de descritores do sinal de voz são extraídos.

O primeiro conjunto refere-se ao descritor individualizado por janela, ou seja, para cada janela $y_{k,j}$ um descritor com 26 elementos é extraído, tanto através da transformada *Wavelet Packet,* quanto pelos coeficientes MFCC (Capítulo III). Estes descritores serão utilizados em duas etapas distintas: identificação de consoantes versus vogal (Figura 5.2-E) e na classificação da vogal propriamente dita (Figura 5.2-G).

O segundo conjunto de descritores refere-se à concatenação dos descritores obtidos nas 15 primeiras janelas do sinal. Como, para cada janela, o descritor extraído possui um tamanho de 26 elementos, então a dimensão deste descritor será de 390. Este descritor será utilizado exclusivamente na identificação da consoante que ocorre antes da vogal (Figura 5.2-H). Por exemplo: na sílaba "par" (/paɾ/) este descritor será utilizado na identificação da consoante anterior a vogal, ou seja, na tentativa de identificação do fonema consonantal /p/. Esta consoante será rotulada como consoante "anterior", pois antecede a vogal.

Devido ao fato de que as consoantes possuem uma grande variação no tempo de duração (de 10 a 100ms para as oclusivas e até 200ms para o caso das fricativas [LUKASIK, 2000]), subconjuntos com 5 e 10 janelas serão também utilizados no reconhecimento destas

consoantes em algumas máquinas especialistas. A Figura 5.7 ilustra a variação no tamanho do tempo de duração das consoantes marcadas por círculos.

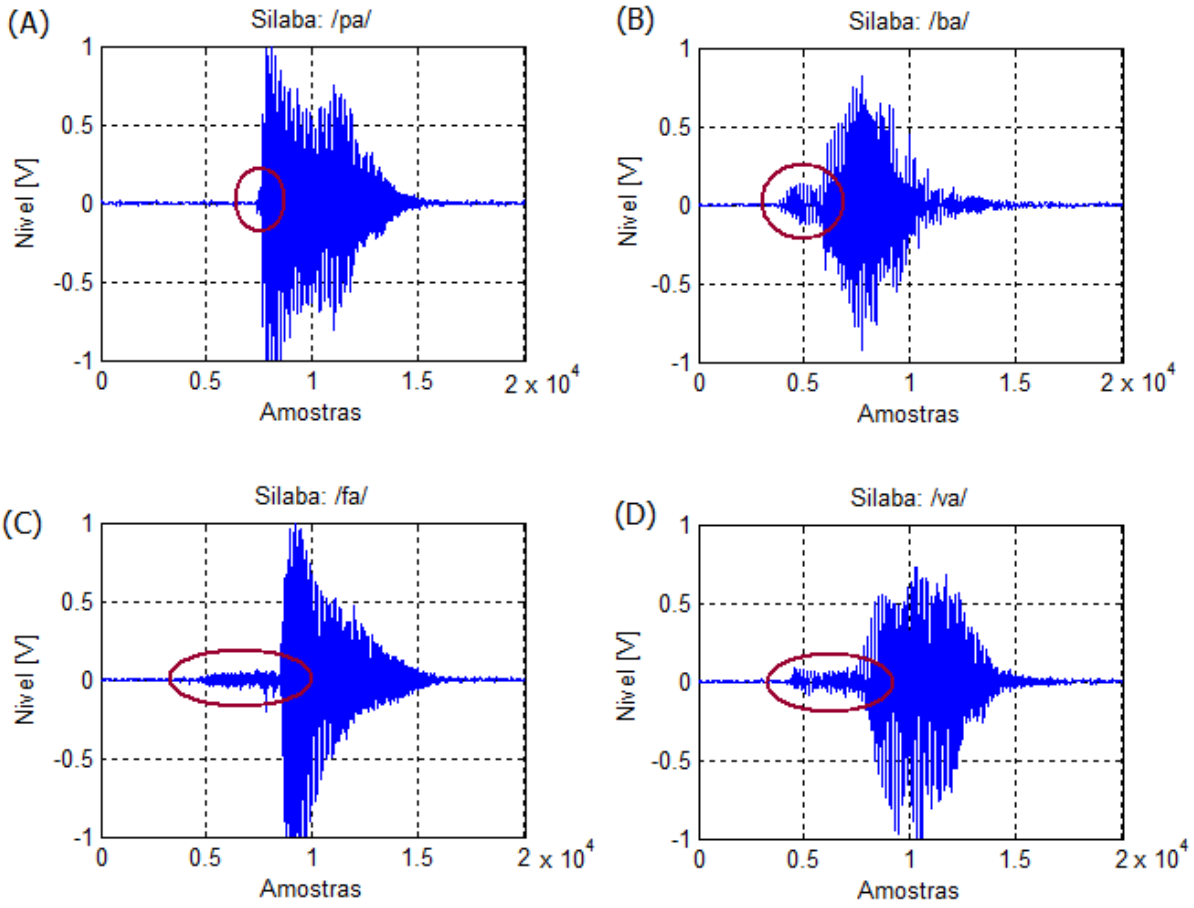

Figura 5.7: Duração das algumas consoantes: (A) /p/, (B) /b/, (C) /f/ e (D) /v/.

A Figura 5.7-A apresenta o sinal da sílaba /pa/ onde a consoante /p/ (oclusiva, bilabial e não vozeada) possuí uma duração de aproximadamente 15ms. A Figura 5.7-B apresenta o sinal da sílaba /ba/ na qual a consoante /b/ (oclusiva, bilabial e vozeada) possuí uma duração aproximada de 90ms. As Figuras 5.7-C e D apresentam os sinais das sílabas /fa/ e /va/ respectivamente. Nesta caso as consoantes /f/ (fricativas, labiodental e não vozeada) e /v/ (fricativa, labiodental e vozeada) apresentam uma duração que varia de 150 a 200ms aproximadamente.

O terceiro conjunto de descritores é obtido através da concatenação dos descritores extraídos das últimas 10 janelas do sinal. Neste caso, a dimensão deste descritor será de 260. Este descritor será utilizado exclusivamente na identificação da consoante que aparece após a vogal (Figura 5.2-I). Novamente, como exemplo cita-se a sílaba “par” (/paɾ/) mas

agora o descritor será utilizado na identificação da consoante posterior a vogal, ou seja, na tentativa de identificação do fonema consonantal /ɾ/. Esta consoante será rotulada como consoante "posterior", pois aparece após a vogal. Vale salientar que para este caso somente dois fonemas são considerados como "posteriores": o /ɾ/ e o /s/ (Capítulo II).

A Figura 5.8 apresenta os sinais das sílabas /paɾ/ e /pas/ com o objetivo de salientar o tamanho da consoante final conforme a sua finalização em /ɾ/ ou /s/. Neste caso, o tamanho do descritor é único (260 ou as 10 últimas janelas concatendas), pois os dois padrões são bem distintos.

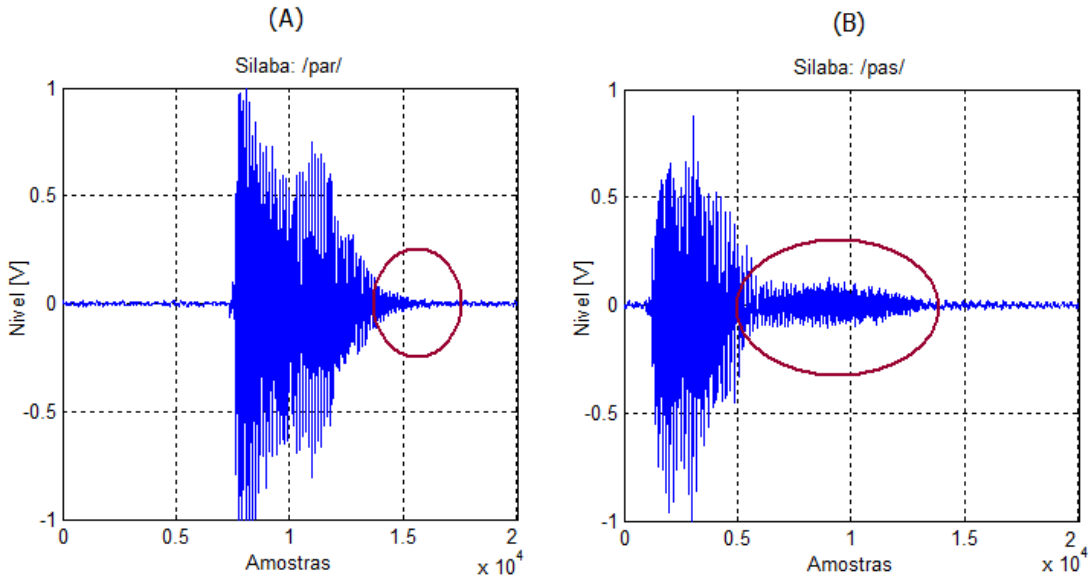

Figura 5.8: Tamanho das consoantes posteriores. (A) final em /ɾ/, (B) final em /s/.

A metodologia empregada na extração dos descritores através da transformada *Wavelet Packet* e dos coeficientes MFCC foi apresentada no Capítulo III. No entanto, vale salientar novamente que, para o caso da *Wavelet Packet* as 26 dimensões (por janela) foram obtidas através do cálculo da energia das 26 bandas selecionadas através da escala Mel (Tabela 3.1) com a *Wavelet* mãe *Daubechies* 5. Esta escolha foi baseada nos trabalhos de Long e Datta (LONG, 1996) e Sarikaya e Hansen (SARIKAYA, 2000) que testaram diversas *Waveletes* mãe no reconhecimento de fonemas. Já para os coeficientes MFCC, as 26 dimensões foram escolhidas propositalmente com o mesmo tamanho da *WP* com o objetivo de padronizar o tamanho dos descritores. Estes dois tipos de descritores, derivados da transformada *Wavelet* e dos coeficientes MFCC ,são utilizados distintamente nas diversas fases do reconhecimento.

5.6 Reconhecimento: Consoante x Vogal

Após a extração dos descritores, segue-se a primeira etapa de identificação propriamente dita, ou seja, o reconhecimento de cada janela como consoante ou vogal (Figura 5.2-E). Esta é a primeira decisão lógica do sistema. O objetivo deste reconhecimento é identificar as janelas que são vogais (sem identificar especificamente a vogal) para, através delas, fazer a separação silábica da palavra ou detectar quantas sílabas existem na palavra. Qualquer etapa que represente um determinado reconhecimento, através de uma máquina ou de um conjunto de máquinas especialistas, possui duas fases distintas: o treinamento e a classificação.

5.6.1 SVM Especialista C x V: Treinamento

A etapa que realiza a identificação das janelas em consoante ou vogal é composta por duas máquinas de vetor de suporte, conforme mostra a Figura 5.9.

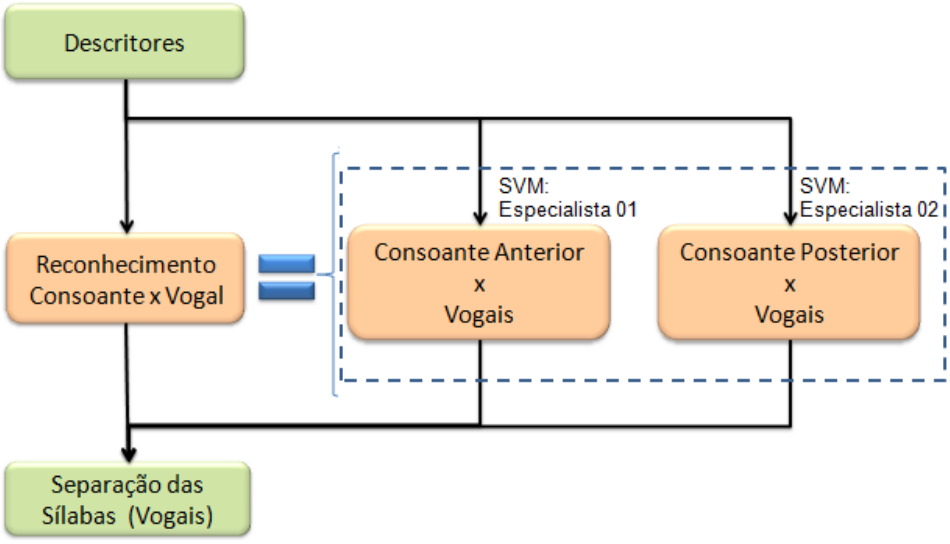

Figura 5.9: Máquinas especialistas para fazer o reconhecimento de Consoantes x Vogais.

A máquina SVM especialista 01 foi treinada com duas classes: a primeira representa as consoantes anteriores e a segunda representa as vogais. A primeira classe é formada por um conjunto de descritores com dimensão igual a 26 (referente a uma janela). De cada sílaba gravada para o treinamento, foram extraídos os descritores das duas primeiras janelas. O motivo de utilizar apenas as duas janelas iniciais de cada sílaba reside no fato mostrado na Figura 5.7-A. Esta restrição é imposta pelas consoantes oclusivas não vozeadas (/p/, /t/ e

/k/), que possuem pouco tempo de duração e portanto, a partir da terceira janela o predomínio é praticamente todo da vogal. A segunda classe é formada por um conjunto de descritores extraídos dos sinais gravados das 12 vogais orais e nasais, janela por janela.

A máquina SVM especialista 02 também foi treinada com duas classes: a primeira representa as consoantes posteriores e a segunda representa as vogais. A primeira classe é formada por um conjunto de descritores com dimensão igual a 26 (referente a uma janela). De cada sílaba gravada para o treinamento, foram extraídos os descritores das duas últimas janelas. A segunda classe desta máquina é representada pelo mesmo conjunto utiliizado na máquina especialista 01.

Para obter os descritores das consoantes anteriores foram utilizadas 1284 gravações de sílabas formadas pelo conjunto CV (consoante + vogal) e CVC (consoante + vogal + consoante). Todas as 19 consoantes foram gravadas em conjunto com a vogal /a/ formando sílabas tais como: /pa/, /paɾ/, /pas/, /ta/, /taɾ/, tas/ etc. Como, para cada sílaba, duas janelas foram extraídas, então o conjunto de treinamento das consoantes anteriores possui 2568 padrões. Já para as consoantes finais, 1342 gravações de sílabas CVC e VC (final em /ɾ/ ou /s/) foram utilizadas, totalizando 2684 padrões.

Para o caso das vogais, 20 gravações de cada vogal foram efetuadas totalizando 240 gravações. O método de extração das vogais, apresentado na Figura 5.10, foi aplicado em cada gravação gerando assim diferentes quantidades de janelas selecionadas em cada sinal. Ao total foram obtidas 3470 padrões das 12 vogais. Este método de extração dos descritores das vogais é um pouco mais complexo e será explanado a seguir.

Para extrair os descritores das janelas que representam um sinal de uma determinada vogal um novo método foi proposto, testado exaustivamente e faz parte da primeira publicação gerada por este trabalho, a qual é citada no Anexo III.

Ao invés de utilizar todas as janelas do sinal da vogal, foi realizada uma segmentação das janelas utilizando o algoritmo *Kmeans* de Duda e Hart [DUDA, 1973] com duas classes. Este procedimento proporciona a escolha das janelas que melhor representam a vogal, diminuindo sensivelmente o tempo de treinamento e melhorando o desempenho do sistema.

A Figura 5.10 apresenta este procedimento onde a linha em vermelho marca as janelas selecionadas automaticamente pelo algoritmo *Kmeans* para o sinal da vogal “a”.

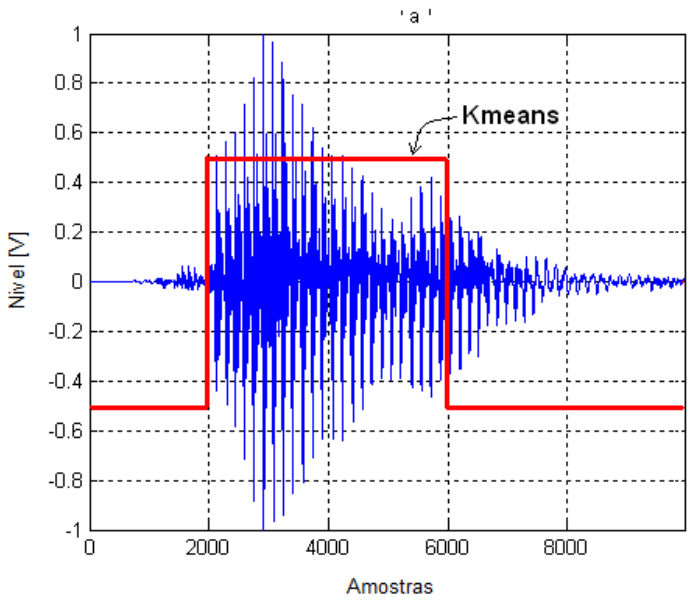

Figura 5.10: Sinal da vogal /a/ com janelas separadas pelo algoritmo *Kmeans*.

Pode-se ver na Figura 5.10 que as janelas selecionadas estão situadas mais ao centro do sinal, evitando a utilização de janelas que possam representar variações de pronúncia ou ruído, tanto no início quanto no final da locução.

As duas máquinas especialistas (SVM) foram treinadas com os descritores MFCC pois as *Wavelets,* para este caso específico, não apresentam bom desempenho devido a pouca energia existente nas consoantes [MALBOS, 1994]. O *kernel* utilizado foi o RBF (função de base radial - Gaussiano) com *C*=20. Devido a grande quantidade de padrões, cada máquina levou cerca de 2 horas para ser treinada. No entanto, os resultados do treinamento foram muito promissores. A máquina 01 gerou uma superfície de separação com apenas 2% dos padrões (das duas classes), ou seja, apenas 2% dos padrões foram utilizados para criar a superfície de separação das duas classes. Já a máquina 02 teve um desempenho melhor ainda utilizando apenas 1% dos padrões como vetores de suporte. Pode-se concluir que, quanto menor for a quantidade de padrões utilizados como vetores de suporte, melhor será a superfice de separação encontrada e o desempenho na identificação das classes.

5.6.2 SVM Especialista C x V: Classificação

Na fase de classificação, as máquinas especialistas apresentadas na Figura 5.9 (previamente treinadas) são utilizadas para fazer o reconhecimento ou identificação do sinal janela a janela. O objetivo é descobrir quais janelas são vogais, e assim, utilizá-las para

separar ou identificar o posicionamento das sílabas dentro da palavra. Novamente, o sinal da palavra “pare”, apresentado na Figura 5.4, é utilizado como exemplo.

Após o sinal da palavra “pare” ter passado por todas as etapas do pré-processamento, janelamento e extração dos descritores, cada janela (descritor) é apresentado ao sistema composto pelas duas máquinas especialistas (previamente treinadas).

Os resultados da classificação podem ser vistos na Figura 5.11 onde as linhas vermelhas, azuis e magentas (parte inferior da figura) delimitam as janelas conforme a superposição. Cada máquina especialista apresenta duas respostas possíveis: 1 para Consoante e -1 para Vogal. Assim, a janela será considerada vogal sempre que uma das máquinas apresentar o resultado -1. A janela será considerada consoante somente quando as duas máquinas apresentarem o resultado igual 1. Pode-se ver que dois conjuntos de vogais (-1) aparecem no sinal. Deste modo, o ponto final do primeiro conjunto e o ponto inicial do segundo conjunto delimitam o intervalo de separação das sílabas. A metodologia utilizada na separação das sílabas utilizando as vogais é explanada na seção a seguir.

5.7 Separação Silábica através das Janelas de Vogais

A Figura 5.11 apresenta também quatro linhas verticais pontilhadas nas cor preta, que servem para delimitar o espaço das vogais determinado pelas máquinas especialistas.

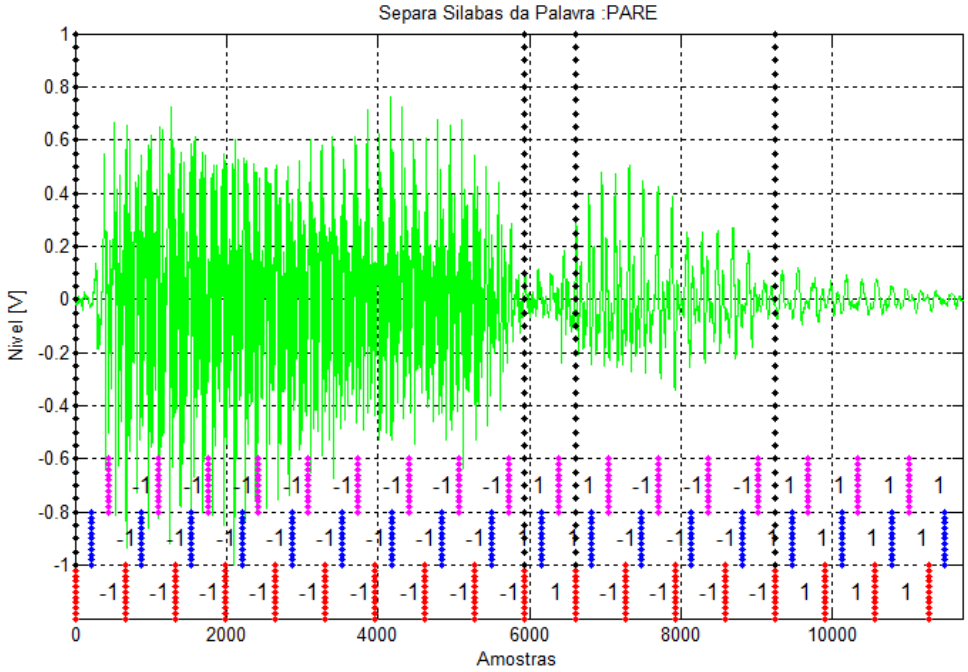

Figura 5.11: Sinal da palavra “pare” com separação das sílabas feitas através das vogais.

O espaço entre a segunda e terceira linha (pontilhada vertical na cor preta) delimita o local onde será feita a divisão silábica. Este processo de separação é simples, uma vez encontrado o espaço entre as vogais, divide-se este espaço em pequenas janelas de 5ms e calcula-se a energia do sinal. O ponto onde houver a menor energia corresponderá ao ponto da divisão silábica.

A Figura 5.12 apresenta os sinais das duas sílabas da palavra “pare” após a separação silábica. Vale salientar que este processo é feito automaticamente e corresponde ao passo lógico do sistema representado pelo bloco “F” da Figura 5.2.

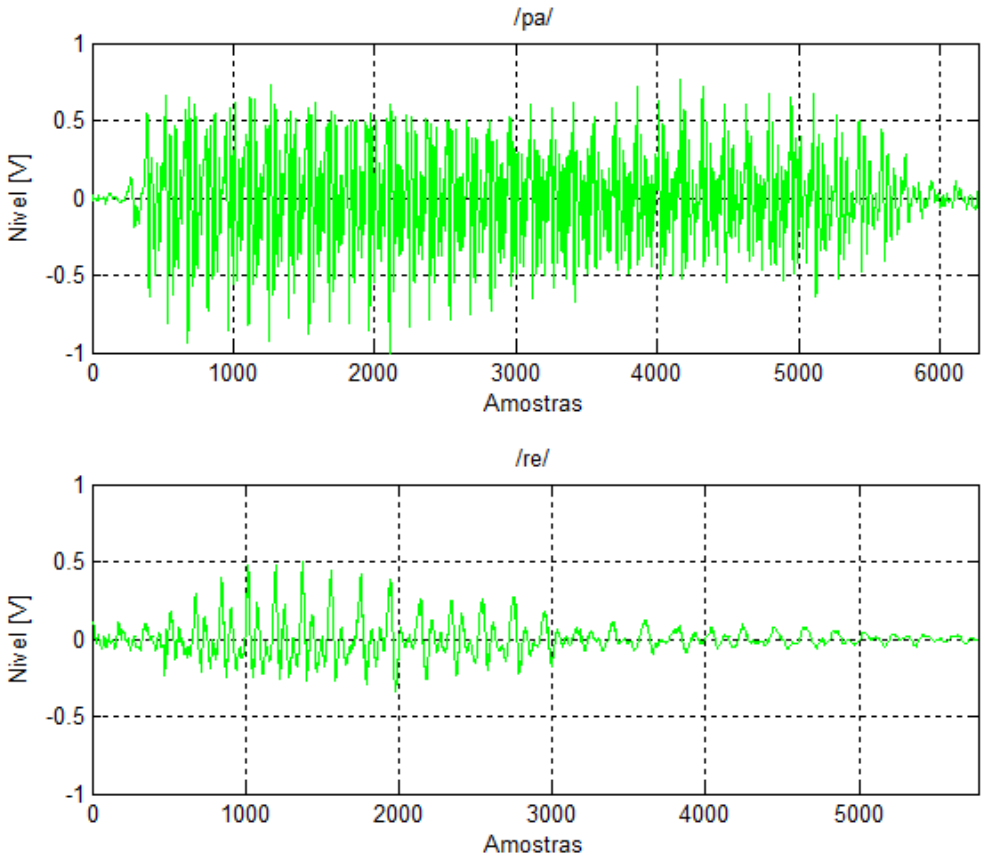

Figura 5.12: Sílabas /pa/ e /ɾe/ separadas da palavra “pare” através das vogais e da energia.

Um ponto interessante a ser explanado é o tempo gasto desde a aquisição do sinal até a separação das sílabas. Obviamente, este tempo depende do tamanho da palavra, pois quanto maior a palavra maior será a quantidade de janelas a serem processadas. Para este exemplo, o tempo até o cálculo dos descritores MFCC é de os aproximadamente 0,05 segundos. O maior tempo é gasto na análise das janelas (51 janelas para este exemplo), correspondendo a 13,35 segundos para este exemplo. O tempo total gasto foi de 15,53 s.

Com o intuito de reduzir este tempo, utilizou-se apenas as janelas subsequentes do sinal, ou seja, não se utilizou as janelas de superposição. Para o caso citado da palavra "pare", ao invés de 51 janelas, foram usadas apenas 17.

A Figura 5.13 apresenta novamente o sinal da palavra "pare", mas agora com análise somente das janelas subsequentes. Pode-se ver que o resultado é o mesmo mas o tempo de análise das janelas diminui para 4,58 segundos e o tempo total gasto diminui para apenas 5,53 segundos. Pode-se concluir, que a redução do tempo é praticamente proporcional à redução da quantidade de janelas a serem analisadas. As linhas verticais pontilhadas na cor magenta marcam o intervalo exato da divisão silábica.

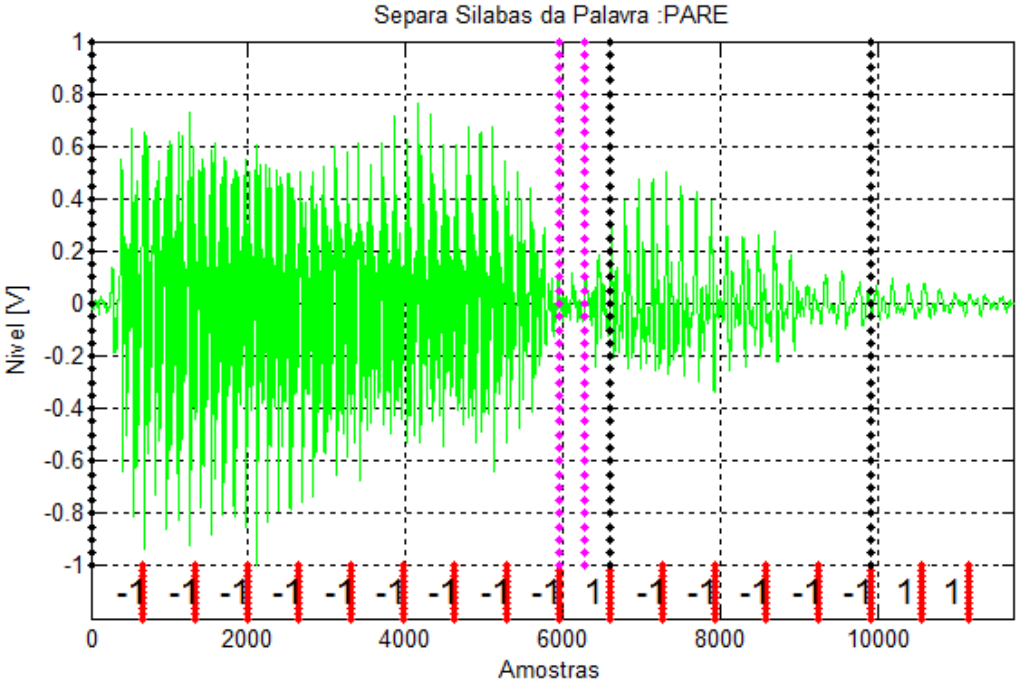

Figura 5.13: Análise somente nas janelas subsequentes para o sinal da palavra "pare".

Este método de separação de sílabas apesar de ser relativamente eficiente é totalmente dependente da velocidade com que o locutor pronuncia a palavra. Nos experimentos realizados neste trabalho, as palavras gravadas foram pronunciadas com a velocidade moderada pois o objetivo maior é a validação do reconhecimento das sílabas.

5.8 Reconhecimento de Vogais

Após a separação das sílabas, os passos do reconhecimento das vogais e consoantes são executados para cada sílaba individualmente.

Com o objetivo de facilitar o entendimento destas etapas, parte da Figura 5.2 é reapresentada na Figura 5.14. Novamente, vale salientar que as etapas que correspondem ao reconhecimento de um padrão ou de um conjunto de padrões possuem duas fases: treinamento e classificação.

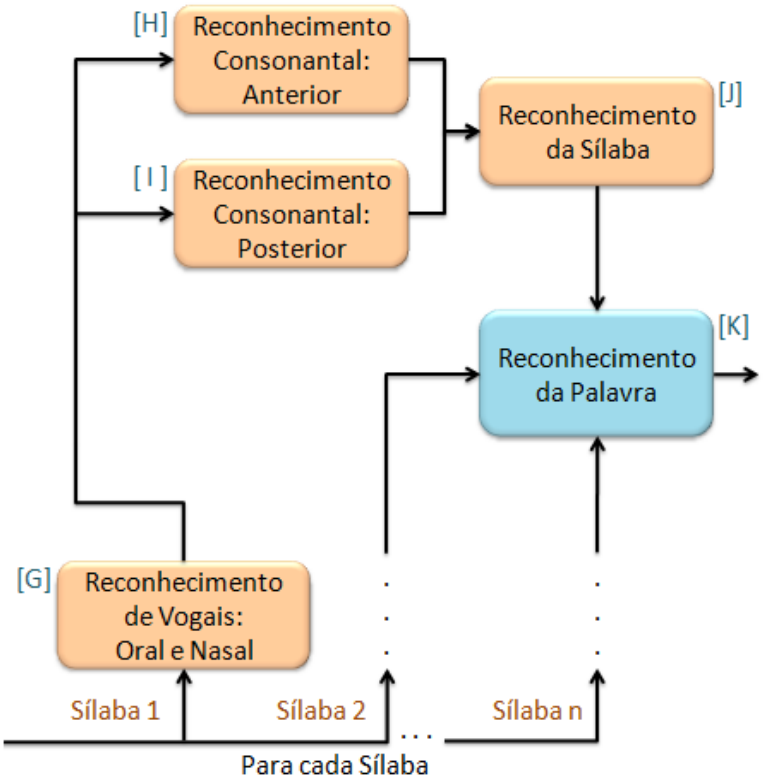

Figura 5.14: Etapas do reconhecimento das vogais e consoantes.

5.8.1 Reconhecimento de Vogais: Treinamento

A Figura 5.15 reapresenta o diagrama de blocos do sistema hierárquico de reconhecimento das vogais referente ao Bloco "G" da Figura 5.14.

O sistema de reconhecimento das vogais construído nesta etapa obedece à hierarquia ditada pelas regras fonéticas da produção das vogais (Capítulo II). Os mesmos arquivos utilizados anteriormente no reconhecimento das janelas (consoantes x vogais) são utilizados para treinar todas as 12 máquinas deste subsistema. Todas as máquinas são constituídas por redes neurais SVMs que separam sempre duas classes. O *kernel* utilizado em todas as máquinas é o RBF (função de base radial) com o parâmetro *C*=20. Para este subsistema os descritores utilizados são os obtidos através da transformada *Wavelet Packet* (Capítulo III) com a *Wavelet* mãe *Daubechies* 5.

A máquina 01 foi treinada com objetivo de separar as vogais orais das nasais. A classe 1 foi selecionada como vogal oral e a classe -1 para as vogais nasais.

A máquina 02 foi treinada com objetivo de separar a vogal /a/ de todas as outras vogais orais. Isto se dá pelo fato da vogal /a/ possuir uma característica de produção fonética totalmente diferente das outras vogais orais (Tabela 2.4). A classe 1 representa a vogal /a/ e a classe -1 representa as outras vogais orais.

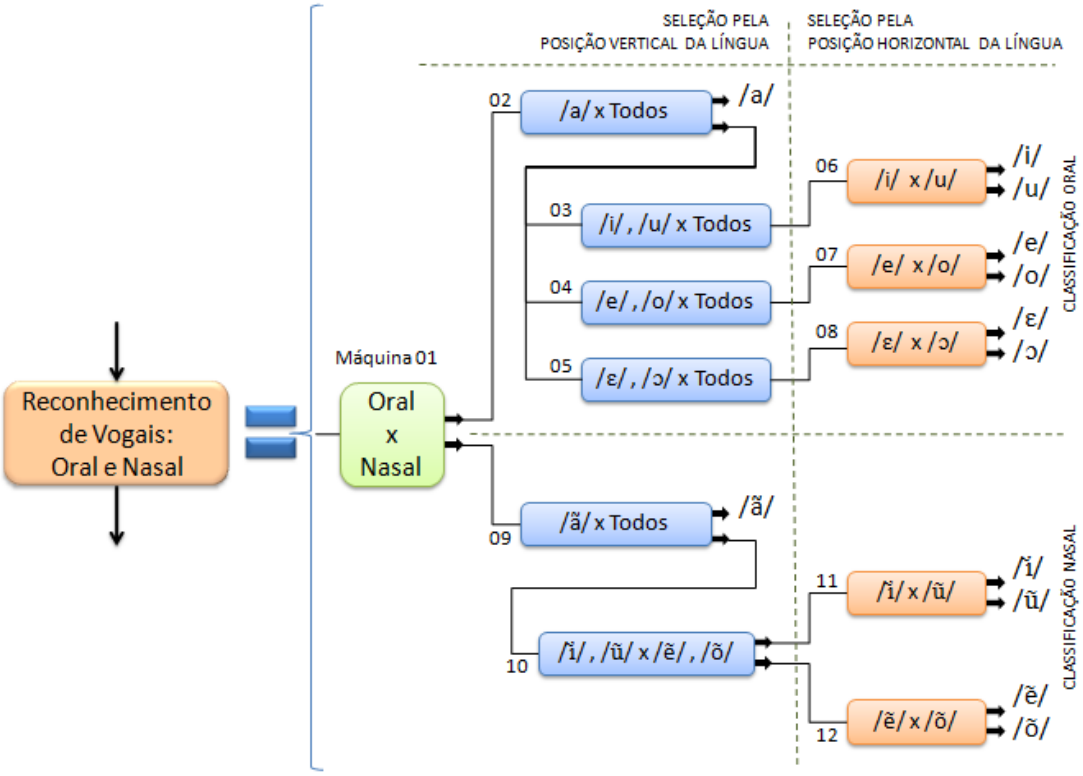

Figura 5.15: Diagrama de blocos do subsistema de reconhecimento de vogais.

As máquinas 03, 04 e 05 foram treinadas com o objetivo de separar as vogais pela posição vertical da língua. Na máquina 03, a classe 1 representa as vogais que têm a mesma posição vertical da língua (alta), neste caso /i/ e /u/. A classe -1 representa as outras vogais orais restantes. O mesmo raciocínio de treinamento foi efetuado para as máquinas 04 e 05, sendo a classe 1 representada pelo conjunto /ɛ/, /ɔ/ e /e/, /o/, respectivamente. Novamente, a classe -1 representa as outras vogais orais restantes.

As máquinas 06, 07 e 08 servem para classificar as vogais conforme a posição horizontal da língua. A máquinas 06 é treinada com duas classes /i/ e /u/ e será acionada somente se a máquina 03 for a vencedora na etapa anterior. O mesmo raciocínio vale para as máquinas 07 e 08 que visam separar as vogais /ɛ/, /ɔ/ e /e/, /o/, respectivamente.

As máquinas 09, 10, 11 e 12 são treinadas com o objetivo de classificar as vogais nasais. A máquina 09 tem o objetivo de separar a vogal nasal /ã/ das demais vogais nasais. A máquina 10 é treinada com o objetivo de separar as vogais nasais pela posição vertical da língua. A máquina 11 é treinada para separar as vogais nasais /ĩ/ e /ũ/ e somente será acionada se a classe 1 for vencedora na máquina 10. Já a máquina 12 é treinada para separar as vogais nasais /ẽ/ e /õ/ e somente será acionada se a classe -1 for vencedora na máquina especialista 10. A Tabela 5.2 apresenta as 12 máquinas (M1, …, M12) do sistema de reconhecimento de vogais com as respectivas quantidades de padrões de treinamento.

Tabela 5.2: Padrões de treinamento para cada classe do sistema de reconhecimento de vogais.

Máquinas	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Classe 1	1178	198	318	355	307	168	170	172	187	349	213	197
Classe -1	942	980	662	625	673	150	185	135	755	406	136	209

5.8.2 Reconhecimento de Vogais: Classificação

Na fase da classificação, o subsistema hierárquico de reconhecimento da Figura 5.15 (previamente treinado) é utilizado para reconhecer a vogal (ou vogais) presente em cada sílaba. A classificação é feita em cada sílaba através da análise das janelas do sinal.

Com o objetivo de validar este subsistema, dois experimentos foram realizados: o primeiro com os descritores MFCC e o segundo com os descritores *Wavelet Packet*.

Ambos os experimentos foram feitos com o mesmo banco de dados. Este banco de dados de validação é composto por 800 gravações de cada uma das 12 vogais.

A Tabela 5.3 apresenta os resultados obtidos no reconhecimento exclusivo de vogais na forma de matriz de confusão para os descritores obtidos através dos coeficientes MFCC. Pode-se notar que apesar do bom desempenho no reconhecimento das vogais orais, os descritores MFCC apresentaram um desempenho não muito satisfatório para as vogais nasais, com exceção da vogal /ĩ/. Na média, a taxa de reconhecimento geral das vogais ficou em 96,87% com os descritores MFCC.

A Tabela 5.4 apresenta os resultados obtidos no reconhecimento de vogais na forma de matriz de confusão para os descritores obtidos através da *Wavelet Packet*. Vale salientar, que as mesmas características do classificador (SVM) foram utilizadas nestes dois experimentos sendo: *kernel* RBF Gaussiano e constante *C*=20.

Tabela 5 3: Resultados do reconhecimento de vogais (MFCC) na forma de validação cruzada.

Vogal	a	ɛ	i	ɔ	u	e	o	/ã/	/ẽ/	/ĩ/	/õ/	/ũ/	% Acerto
a	800	0	0	0	0	0	0	0	0	0	0	0	100,00
ɛ	0	785	0	0	0	0	6	0	3	6	0	0	98,12
i	0	0	800	0	0	0	0	0	0	0	0	0	100,00
ɔ	0	20	0	775	0	0	0	0	0	0	3	2	96,87
u	0	0	11	0	782	0	0	0	0	0	0	7	97,75
e	0	0	0	0	0	786	8	0	2	0	4	0	98,25
o	0	13	0	0	0	10	777	0	0	0	0	0	97,12
/ã/	41	0	0	0	0	0	0	759	0	0	0	0	94,87
/ẽ/	0	0	0	0	0	22	6	0	743	1	27	1	92,87
/ĩ/	0	0	0	0	0	0	0	0	0	795	0	5	99,37
/õ/	0	0	0	0	0	7	29	0	31	0	733	0	91,62
/ũ/	0	0	3	0	15	0	0	0	0	17	0	765	95,62

Tabela 5.4: Resultados do reconhecimento de vogais (WP) na forma de validação cruzada.

Vogal	a	ɛ	i	ɔ	u	e	o	/ã/	/ẽ/	/ĩ/	/õ/	/ũ/	% Acerto
a	800	0	0	0	0	0	0	0	0	0	0	0	100,00
ɛ	0	781	0	0	0	0	19	0	0	0	0	0	97,62
i	0	0	800	0	0	0	0	0	0	0	0	0	100,00
ɔ	0	10	0	790	0	0	0	0	0	0	0	0	98,75
u	0	0	8	0	789	0	0	0	0	0	0	3	98,62
e	0	0	0	0	0	776	18	0	2	0	4	0	97,00
o	0	5	0	0	0	15	776	0	0	0	4	0	97,00
/ã/	0	0	0	0	0	0	0	800	0	0	0	0	100,00
/ẽ/	0	0	0	0	0	15	6	0	766	1	11	1	95,75
/ĩ/	0	0	0	0	0	0	0	0	0	800	0	0	100,00
/õ/	0	0	0	0	0	4	11	0	14	0	771	0	96,37
/ũ/	0	0	1	0	3	0	0	0	0	11	0	785	98,12

Apesar de em alguns casos (vogal ɛ por exemplo) a taxa de reconhecimento ser menor que o obtido com a MFCC, em geral o desempenho da *Wavelet* é melhor pois o reconhecimento global foi de 98,26%, com ênfase na significativa melhoria do reconhecimento das vogais nasais. Além disso, o desvio padrão, para os valores de acerto de cada vogal, é menor para o caso da *Wavelet*. Isto demonstra também uma maior uniformidade nos resultados obtidos.

O gráfico da Figura 5.16 apresenta uma comparação da dos resultados obtidos no reconhecimento de vogais, entre os descritores MFCC e *Wavelet Packet* (Tabelas 5.3 e 5.4). Pode-se ver que para as vogais orais, o desempenho de ambos os descritores é muito

parecido. No entanto, quando trata-se das vogais nasais a *Wavelet Packet* apresenta um desempenho melhor.

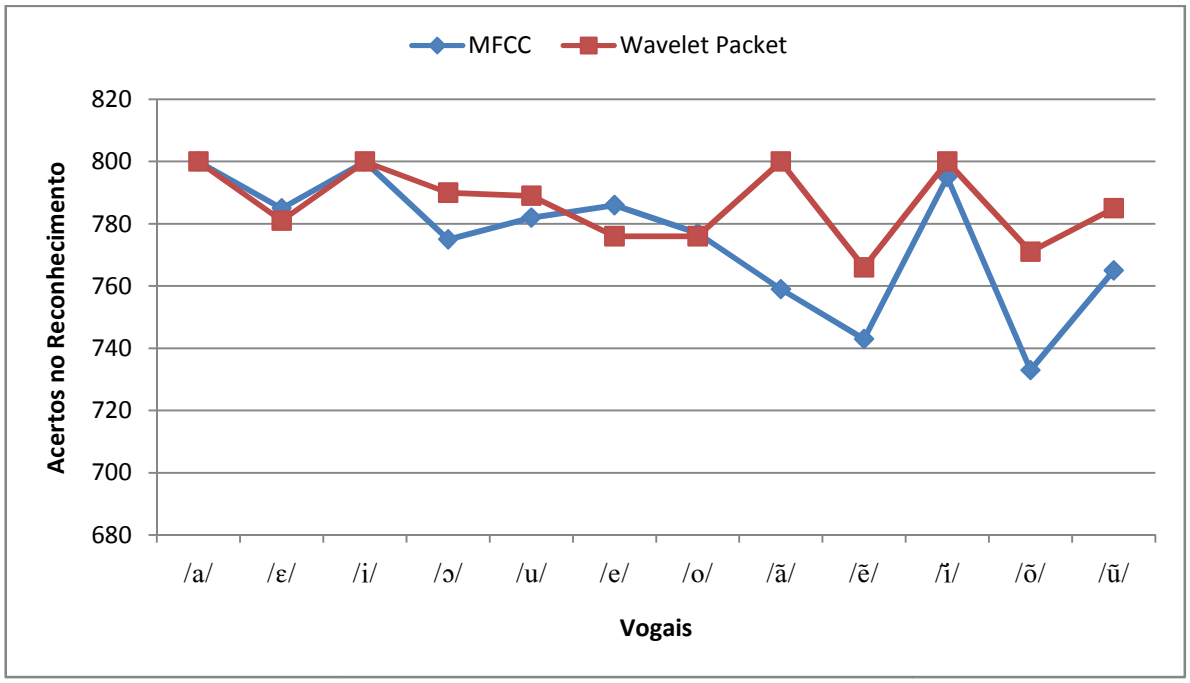

Figura 5.16: Comparação entre MFCC e Wavelet Packet no reconhecimento de vogais.

Por fim, vale salientar que com este subsistema de reconhecimento, é possível encontrar todas as vogais presentes no sinal, inclusive encontros vocalicos tais como os ditongos.

No entanto, devido a limitação imposta neste trabalho, com objetivo de validar a proposta, o reconhecimento de encontros vocálicos não foi considerado.

5.9 Reconhecimento de Consoantes Anteriores

Após o reconhecimento da vogal existente na sílaba, o próximo passo é tentar reconhecer as consoantes que aparecem (ou não) no início e no fim da sílaba. No entanto, uma determinada sílaba pode ser formada somente por uma vogal sem a existência de consoantes antes ou depois da vogal, como por exemplo na palavra formada por apenas uma sílaba, “um”.

A Figura 5.17 apresenta o subsistema de reconhecimento para a consoante anterior. Este sistema foi exaustivamente testado e gerou a última publicação deste trabalho, a qual encontra-se citada no Anexo III.

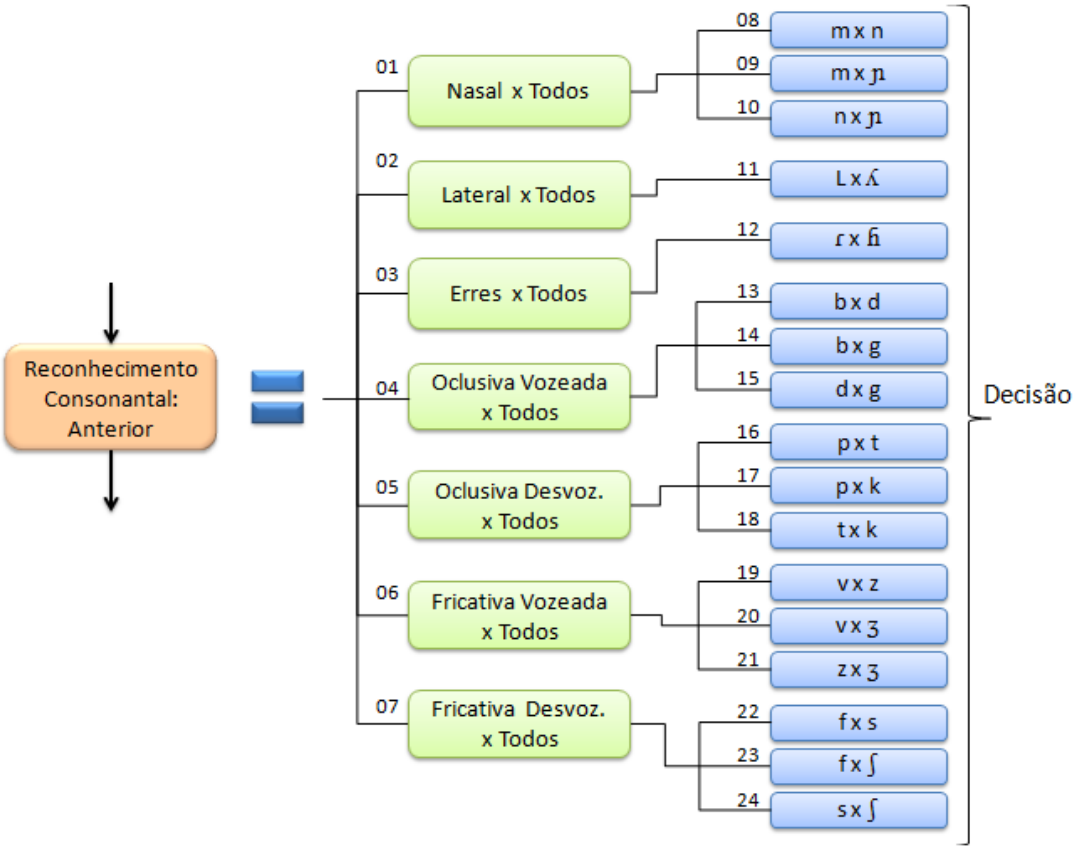

Figura 5.17: Subsistema hierárquico de reconhecimento das consoantes anteriores.

O subsistema apresentado na Figura 5.17 é composto por duas etapas: reconhecimento pela maneira de produção e reconhecimento pelo modo de articulação das consoantes. Novamente, vale salientar que como este subsistema é responsável pelo reconhecimento de padrões, portanto o mesmo é utilizado em duas etapas distintas: treinamento e classificação.

5.9.1 Reconhecimento de Consoantes Anteriores: Treinamento

O sistema de reconhecimento das consoantes obedece à hierarquia ditada pelas regras fonéticas da produção das consoantes (Capítulo II). Todas as máquinas são constituídas por redes neurais SVMs que separam sempre duas classes. O parâmetro *C*=20 foi utilizado para todas as máquinas. Devido á pouca energia existente nos sinais consonantais [MALBOS, 1994], para este subsistema os descritores utilizados são aqueles obtidos através dos coeficientes MFCC (Capítulo III).

Para este subsistema específico, os resultados encontrados nos experimentos apontaram para a utilização de diferentes *kernels* para os classificadores e diferentes

dimensões dos descritores. Este fato só é possível devido à total independência de cada máquina especialista.

Basicamente, o subsistema da Figura 5.17 possui duas etapas distintas: a primeira é composta pelas máquinas 01 a 07 e é responsável pela classificação das consoantes através do modo de articulação, tendo sete padrões definidos como: Nasal, Lateral, Erres, Oclusivas Vozeadas, Oclusivas Não vozeadas, Fricativas Vozeadas e Fricativas Não vozeadas. A segunda etapa é composta pelas máquinas 08 a 24 e é responsável pelo reconhecimento final.

As máquinas 01 a 07 foram treinadas através da estratégia “um contra todos”. Por exemplo, a máquina 01 foi treinada com duas classes: a classe 1 é formada pelo conjunto das consoantes nasais /m/, /n/ e /ɲ/ (nh), e a classe -1 é formada por todas as outras consoantes. O mesmo raciocínio foi utilizado para o treinamento das máquinas 02 a 07. Este grupo de máquinas foi treinado com os descritores MFCC com dimensão 390, ou seja, obtidos da concatenação da 15 primeiras janelas do sinal. O *kernel* utilizado para todo este conjunto foi o RBF Gaussiano.

As máquinas de 08 a 12 foram treinadas com descritores MFCC de tamanho 260 (10 janelas iniciais). Já as máquinas responsáveis pelo reconhecimento de consoantes vozeadas (máquinas 13, 14, 15, 19, 20 e 21) foram treinadas com descritores MFCC de tamanho 130 (5 janelas iniciais). Por fim, as máquinas restantes (16, 17, 18, 22, 23, e 24) foram treinadas com os descritores MFCC com dimensão 390. Para todas estas máquinas (08 a 24), o *kernel* utilizado o *Polinomial.*

As escolhas diferenciadas de tamanho do descritor e do tipo de *kernel* foram feitas empiricamente, ou seja, após exaustivos testes de reconhecimento. Diversos *kernels* foram testados neste trabalho, no entanto, somente os *kernels* polinomial e RBF-Gaussiano apresentaram resultados significativos.

As Figuras 5.18 e 5.19 apresentam a os gráficos com as quantidades de vetores de suporte (para cada *kernel*), utilizados pelas máquinas SVM no treinamento, da primeira e segunda etapa do reconhecimento de consoantes anteriores, respectivamente (conforme Figura 5.17).

Pode-se ver que na Figura 5.18, a qual se refere a quantidade de vetores de suporte utilizados pela máquinas 01 a 07, que o *kernel* RBF-Gaussiano utilizou menos vetores de suporte que o polinomial no treinamento. Já no gráfico da Figura 5.19, o qual se refere a

quantidade de vetores de suporte utilizados pela máquinas 08 a 24, o *kernel* polinomial utilizou menos vetores de suporte que o RBF-Gaussiano.

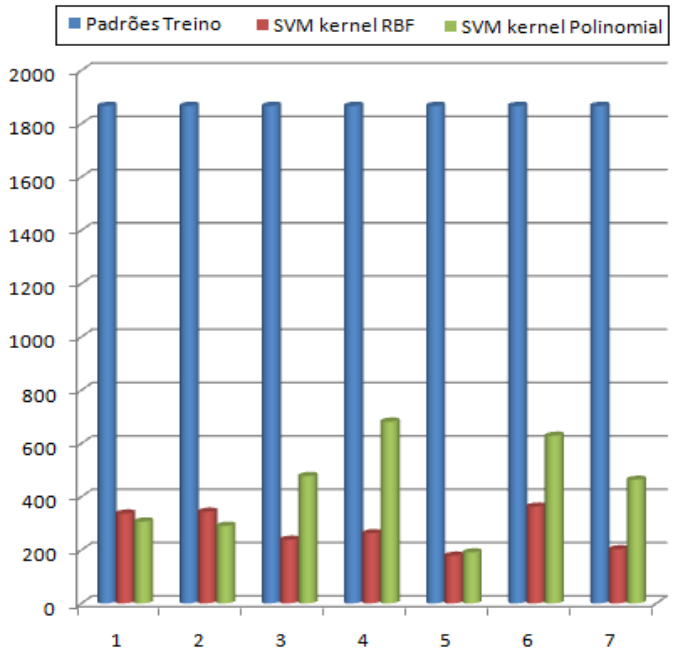

Figura 5.18: Quantidade de vetores de suporte utilizados no treino das máquinas 01 a 07.

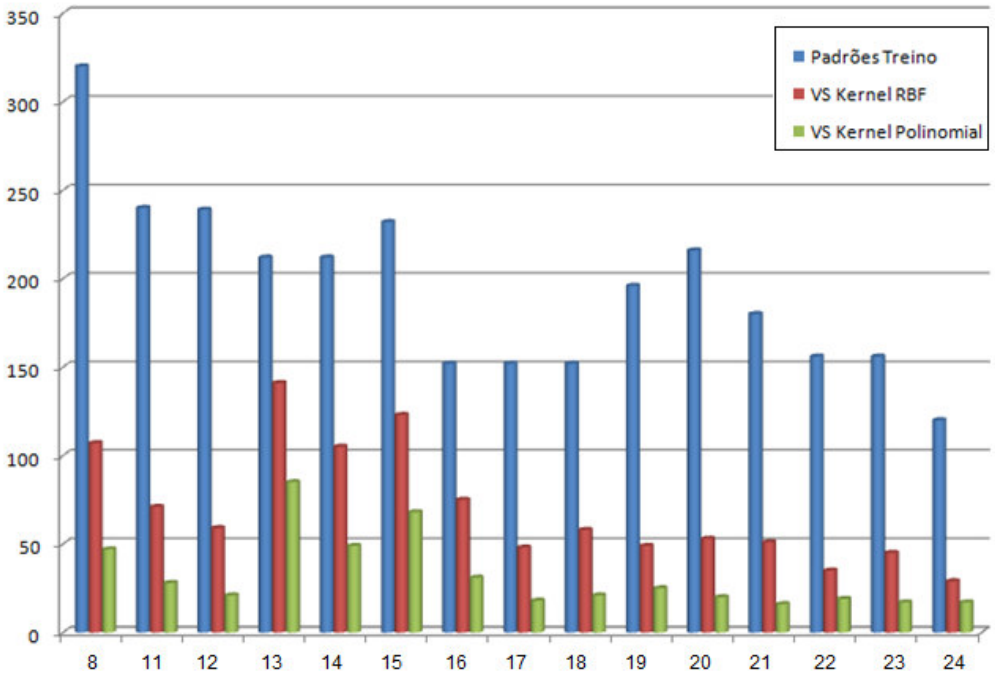

Figura 5.19: Quantidade de vetores de suporte utilizados no treino das máquinas 08 a 24.

Vale salientar que, na Figura 5.19, as máquinas 09 e 10 não aparecem pois não foram utilizadas nos treinos.

A partir dos resultados apresentados nos gráficos das Figuras 5.18 e 5.19, pode-se concluir que quanto menor for a quantidade de vetores de suporte utilizados no treinamento de uma máquina SVM, melhor será a superfície de separação das classes, e conseqüentemente, melhor será o desempenho na classificação.

O banco de dados de treinamento foi formado pela gravação de sílabas individuais de cada uma das 19 consoantes na configuração CV (consoante + vogal /a/). Ao total foram utilizados 100 padrões de treinamento para cada consoante.

5.9.2 Reconhecimento de Consoantes Anteriores: Classificação

Para validar o subsistema hierárquico de reconhecimento das consoantes anteriores, um banco de vozes foi previamente gravado com todas as possíveis palavras formadas pela junção CV (consoante + vogal) com duas sílabas. Para efeito de simplificação somente palavras com a vogal /a/ foram gravadas. A Tabela 5.5 apresenta todas as 123 palavras utilizadas neste experimento de reconhecimento de consoantes.

Tabela 5.5: Palavras utilizadas no reconhecimento de consoantes.

/p/	/t/	/k/	/b/	/d/	/g/	/f/	/s/	/ʃ/	/v/	/z/	/ʒ/	/L/	/h, ɾ/	/m/	/n/
papa	tapa	capa	bata	data	gata	faca	sapa	chapa	vaga	zata	japa	lapa	rapa	mapa	napa
pata	tata	cata	baba	dada	gava	fada	saca	chata	vaca	zaza	jaca	lata	rata	mata	nata
paca	taca	caça	baja	daga	gaza	fafa	saga	chaga	vaza	zaga	jaba	lara	rafa	maca	naba
paga	taga	cada	bala	dava	gaja	faça	sasa	xara	vala	zara	java	lada	raça	massa	nada
passa	taxa	casa	barra	dalha	gala	fala	saxa	xaxa	valha		jarra	lava	rasa	mala	nasa
pajá	tala	cava			garra	falha	sala		vara		jaja	lala	rala	malha	naça
pala	talha	caja			galha	fara	sara		vava			lara	ralha	mara	naja
palha	tara	cala				farra			varra			laça	racha	marra	nara
para	taça	calha				fava							rará	mama	narra
paxa		cara				fata									nana
paza															

Este experimento foi realizado somente com 18 fonemas consonantais pois o fonema /ɲ/ (nh) não possui formação com vogais orais. O banco de dados foi gravado utilizando somente um locutor (dependente do locutor) e é composto por 1854 gravações de palavras individuais, ou 3708 sílabas.

Como existem duas etapas distintas dentro deste subsistema, duas tabelas com resultados são apresentadas.

A Tabela 5.6 apresenta os resultados na forma de matriz de confusão referente a classificação dos fonemas consonantais conforme a maneira de articulação (máquinas 01 a 08). A Tabela 5.7 apresenta os resultados em forma de matriz de confusão para cada fonema consonantal.

Tabela 5.6: Resultado do reconhecimento de consoantes (etapa: modo de articulação).

Grupos de Consoantes	N	L	E	OV	OD	FV	FS	Sílabas	% Acerto
Nasal (N)	438	2	0	4	0	0	0	444	98,65
Lateral (L)	0	380	0	0	0	0	0	380	100,00
Erres (E)	0	0	437	1	2	0	0	440	99,32
Oclusiva Vozeada (OV)	0	0	0	590	0	0	0	590	100,00
Oclusiva Não vozeada (OD)	0	0	0	0	710	0	0	710	100,00
Fricativa Vozeada (FV)	0	5	0	3	0	600	2	610	98,36
Fricativa Não vozeada (FD)	0	0	0	2	0	0	532	534	99,63

Tabela 5.7: Resultado do reconhecimento de consoantes.

/	m	n	L	ʎ	ɾ	h	b	d	g	p	t	k	v	z	ʒ	f	s	ʃ	% Acerto
m	215	7	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	95,98
n	4	212	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	96,36
L	0	0	230	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100,00
ʎ	0	0	0	150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100,00
ɾ	0	0	0	0	237	0	0	1	0	2	0	0	0	0	0	0	0	0	98,75
h	0	0	0	0	0	200	0	0	0	0	0	0	0	0	0	0	0	0	100,00
b	0	0	0	0	0	0	158	0	2	0	0	0	0	0	0	0	0	0	98,75
d	0	0	0	0	0	0	11	175	4	0	0	0	0	0	0	0	0	0	92,11
g	0	0	0	0	0	0	2	5	233	0	0	0	0	0	0	0	0	0	97,08
p	0	0	0	0	0	0	0	0	0	240	0	0	0	0	0	0	0	0	100,00
t	0	0	0	0	0	0	0	0	0	1	259	0	0	0	0	0	0	0	99,62
k	0	0	0	0	0	0	0	0	0	0	0	210	0	0	0	0	0	0	100,00
v	0	0	1	0	0	0	0	3	0	0	0	0	216	0	0	0	0	0	98,18
z	0	0	0	0	0	0	0	0	0	0	0	0	0	198	0	0	2	0	99,00
ʒ	0	0	0	4	0	0	0	0	0	0	0	0	2	0	184	0	0	0	96,84
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150	0	0	100,00
s	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	222	0	99,11
ʃ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	160	100,00

Os resultados apresentados nesta seção, representam a melhor combinação dos parâmetros de treinamento, os quais foram obtidos após o teste de diversos experimentos. Estes parâmetros de treinamento foram especificados na seção anterior.

Como o sistema é hierárquico, os erros apresentados na primeira etapa (Tabela 5.6) são refletidos na segunda etapa (Tabela 5.7). Além disso, outros erros internos na segunda etapa ocorrem diminuindo o desempenho do reconhecimento final das consoantes. No

entanto, como pode ser visto na Tabela 5.7, diversas consoantes obtiveram 100% de acerto e na média o acerto foi de 98,41%.

No que diz respeito as palavras, o desempenho geral é menor, pois como cada palavra é composta por duas sílabas, se apenas uma das sílabas for reconhecida erroneamente então toda a palavra estará errada. O desempenho final do reconhecimento de palavras foi de 96,82%. Por fim, vale salientar que os resultados obtidos e apresentados nas Tabelas 5.6 e 5.7 foram realizados com sílabas formadas somente com a vogal /a/. Esta simplificação se fez necessária para a validação do sistema.

No entanto, conforme descrito no Capítulo IV, mais especificamente na Figura 4.13, teoricamente, para fazer o reconhecimento de quaisquer sílabas, ou seja, o reconhecimento de qualquer sílaba com as 12 vogais existentes, se faz necessário a implantação de 12 subsistemas consonantais, cada qual com sua própria vogal. Este fato está embasado nos trabalhos de Eleonora Maia [MAIA, 1985] e Thais C. Silva [SILVA, 2003], que enfatizam a influência da vogal sobre o sinal da consoante, devido a maior energia e maior tempo de duração da vogal. Os resultados apresentados nesta seção validam esta suposição, pois os descritores utilizados no reconhecimento consonantal possuem janelas pertencentes à consoante, a transição entre a consoante e a vogal, bem como as janelas iniciais do sinal da vogal. Este fato é válido também para o reconhecimento das consoantes posteriores.

5.10 Reconhecimento de Consoantes Posteriores

O reconhecimento das consoantes posteriores é feito com apenas uma máquina especialista (Figura 5.2-I). Diversas consoantes (letras: “z”, “s”, “r”, “l”, “m”, “n”) aparecem após a vogal, formando sílabas do tipo CVC (consoante + vogal + consoante) ou VC (vogal + consoante).

No entanto, para a fonética, somente dois fonemas consonantais são considerados posteriores: o fonema /s/, como na palavra “paz” (/pas/) e o fonema /ɾ/, como na palavra “par” (/paɾ/).

Nas sílabas terminadas em “L”, o fonema é considerado como uma semi-vogal representada pelo símbolo /w/. Já as sílabas terminadas em “m” e “n” não geram um novo fonema, mas sim a nasalização da vogal anterior tornando-a uma vogal nasal.

A máquina especialista (Figura 5.2-I), responsável pela classificação do fonema consonantal posterior, é apresentada na Figura 5.20 a seguir.

Figura 5.20: Máquina especialista SVM para reconhecimento da consoante posterior.

5.11 Reconhecimento da Sílaba e Palavra

O reconhecimento final da sílaba é feito através da junção dos reconhecimentos da vogal, consoante anterior e posterior. Os resultados foram apresentados nas Tabelas 5.6 e 5.7. Após o reconhecimento de cada sílaba, o reconhecimento final da palavra se dá pela junção das sílabas reconhecidas.

Vale salientar que, alguns erros de reconhecimento podem ser corrigidos através de uma etapa simples de correção ortográfica, por exemplo: Em muitos casos a palavra “pare” (/paɾe/) foi reconhecida como /paɾi/, /paɾɛ/, /baɾe/ e /baɾi/, em todos estes casos as palavras identificadas não existem, e portanto, podem ser corrigidas por um corretor ortográfico para palavra correta.

Este fato também é valido para o reconhecimento de palavras isoladas. Em um estudo posterior, cada palavra reconhecida pode ser concatenada formando o reconhecimento de frases, teoricamente.

O sistema apresentado neste capítulo tem o objetivo do reconhecimento de palavras isoladas, e portanto, o mesmo somente estará completo se todos os subsistemas consonantais (uma para cada uma das 12 vogais) forem treinados. Para fazer este treinamento e uma posterior validação, é necessária a utilização de um banco de dados de vozes diversificado e com grande número de padrões, o qual ainda não existe para a língua Portuguesa do Brasil.

Outro fator que limita a utilização deste sistema é o tempo de execução, pois devido as diversas etapas o sistema gasta em torno de 20 segundos para realizar o reconhecimento de

uma palavra com duas sílabas. A utilização do processamento paralelo em algumas etapas do sistema poderia vir a reduzir sensivelmente este tempo de execução.

5.12 Comparação de Resultados

As Figuras 5.21 e 5.22 apresentam os resultados obtidos neste trabalho (Capítulo V) em comparação com outros resultados apresentados na literatura, para o reconhecimento de vogais e de palavras, respectivamente.

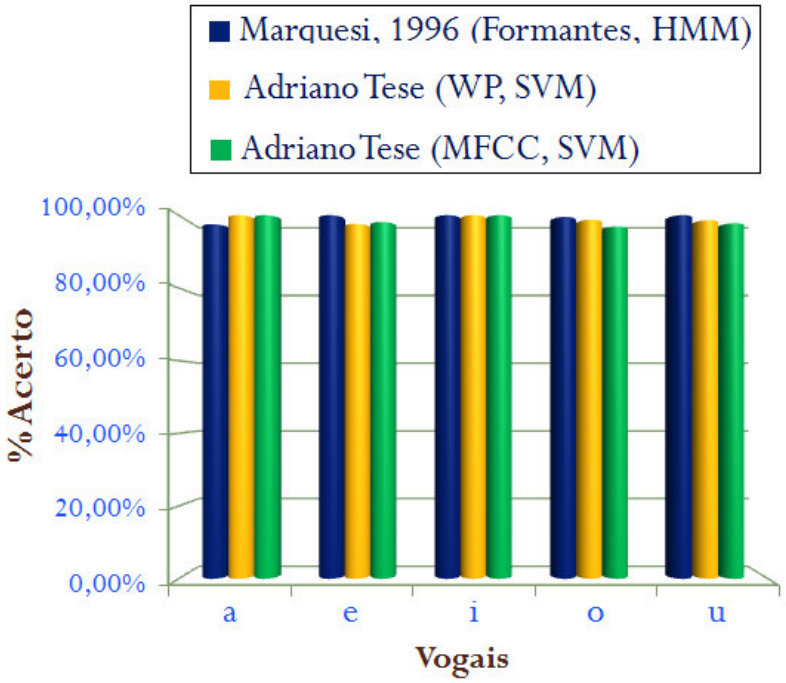

Figura 5.21: Comparação de resultados obtidos no reconhecimento de Vogais.

Pode-se ver na Figura 5.21, que os resultados obtidos neste trabalho, em ambos os casos, são próximos aos obtidos no trabalho de Marquesi [MARQUESI, 1996]. O Trabalho de Marquesi utilizou a formantes do sinal de voz para o reconhecimento de cinco vogais orais (a, e, i, o, u). Vale salientar que neste trabalho foram utilizadas 12 vogais no reconhecimento, sendo portanto uma quantidade de classes maior, ou seja, a complexidade deste trabalho foi maior.

Na Figura 5.22 é apresentada uma comparação dos resultados obtidos no reconhecimento de palavras deste trabalho, em comparação com o trabalho de Rocha *et al* [ROCHA, 2005].

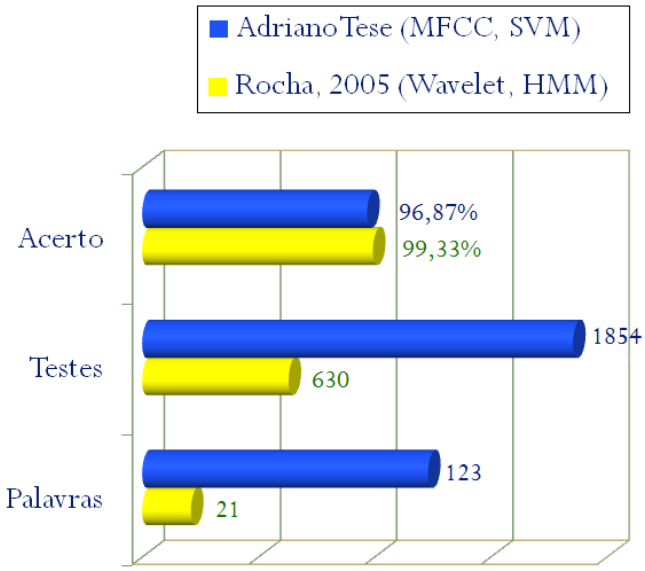

Figura 5.22: Comparação de resultados obtidos no reconhecimento de Palavras.

O trabalho de Rocha *et al* [ROCHA, 2005] utilizou a *Wavelet* com seis níveis de decomposição para o reconhecimento de 21 comandos diferentes para acionar um vídeo cassete. O classificador utilizado foi o HMM. Por fim, vale salientar, que neste trabalho foram utilizados mais padrões, mais classes e os resultados obtidos foram próximos.

Capítulo VI

6. Considerações Finais, Contribuições e Trabalhos Futuros

O reconhecimento automático da voz por máquinas tem sido a meta de pesquisadores por quase cinco décadas. Neste período, inúmeros foram os avanços, no entanto ainda se está longe de desenvolver um sistema que possa ter um desempenho parecido com o ser humano [LIPPMANN, 1997]. Apesar disso, uma grande quantidade de pesquisas foram desenvolvidas e tornaram padrão algumas fases do reconhecimento de voz. A amostragem e o pré-processamento do sinal são hoje praticamente uniformizados, diferenciando apenas algunas detalhes especificos em cada sistema conforme a utilização desejada.

Atualmente, algumas versões de sistemas de reconhecimento de voz estão sendo disponibilizadas comercialmente para o uso em computadores pessoais, celulares, automação embarcada etc. Entretanto, seus desempenhos não são de confiança e deixam muito a desejar [LEVISON, 2005].

O grande objetivo de todas as pesquisas sobre o reconhecimento de voz, é a busca do desenvolvimento de um sistema de reconhecimento de voz que opere de modo contínuo e independente do locutor. Diversas contribuições são apresentadas periódicamente em congressos e revistas especializadas da área. Este trabalho tem como objetivo tornar-se uma destas contribuições, ou seja, contribuir ao grande universo de pesquisas que estão em desenvolvimento em todas as partes do mundo.

Diversos são os desafios no desenvolvimento de pesquisas sobre o reconhecimento de voz, dentre eles, pode-se citar a grande quantidade de padrões existentes, pois qualquer linguagem moderna tais como: Inglês, Francês, Espanhol e Português possuem aproximadamente 500.000 palavras ou padrões a serem identificados.

Em resumo, a proposta deste trabalho foi utilizar unidades menores do que a palavra tais como: fonemas, difones, e sílabas como unidades base para o reconhecimento da voz, com o objetivo de reconhecer quaisquer palavras sem necessariamente utilizá-las. O sistema foi desenvolvido com uma lógica de reconhecimento hierárquica baseada nas características de produção dos fonemas da língua Portuguesa do Brasil.

Apesar dos descritores obtidos através dos coeficientes MFCC, serem praticamente um padrão junto à comunidade científica, as *Wavelets* demonstram que possuem um potencial muito bom e, em muitos casos, apresentam melhores resultados do que os descritores MFCC. No entanto, o tempo de processamento ainda é um empecilho a utilização das *Wavelets* em tempo real.

Os classificadores baseados nos modelos de Markov (HMM) também são o padrão nas pesquisas de reconhecimento de voz. No entanto, em seu editorial, Russell e Bilmes [RUSSEL, 2003] declaram que nos últimos anos reascendeu o interesse em classificadores que possam ir além da desempenho dos sistema baseados nos modelos de HMM. As Máquinas de Vetor de Suporte (SVM) têm provado em inúmeras pesquisas que podem fazer isto. A idéia principal do SVM é construir um hiperplano como superfície de decisão de tal forma que a margem de separação entre exemplos positivos e negativos seja máxima. Para isso a máquina faz uso de duas teorias: Dimensão VC e a Minimização Estrutural do Risco. Resumindo, as SVMs possuem características importantes que justificam a sua utilização:

- Boa capacidade de generalização;
- Robustez diante de objetos de dimensões elevadas;
- Convexidade da função objetivo, ou seja, possui apenas um mínimo global;
- Capacidade de lidar com dados ruidosos;
- Uma base teórica bem estabelecida na Matemática e Estatística;
- Capacidade de agrupamento na forma de Sistemas hierárquicos de Máquinas Especialistas.

Vale salientar que, o foco deste trabalho foi limitado ao reconhecimento de palavras isoladas e ao modo dependente do locutor. O objetivo desta limitação foi a validação desta nova proposta. Portanto, pode-se concluir que o método proposto neste trabalho apresentou bons resultados nas etapas de reconhecimento de vogais e consoantes, se comparado com outros métodos existentes na literatura (Capítulo V).

O trabalho de Lippmann que comparava o desempenho das máquinas de reconhecimento de voz com o desempenho do ser humano [LIPPMANN, 1997] apresentou alguns problemas fundamentais existentes a serem enfrentados pelos novos sistemas (Capítulo I, pg. 08), os quais são válidos ainda hoje. Dentre eles está a afirmação de que, até momento, os algoritmos não são capazes de reconhecer palavras novas sem ter que refazer o treinamento. No sistema apresentado neste trabalho este problema é praticamente

eliminado pois não seria preciso retreinar o sistema se uma nova palavra surgisse. Isto se deve ao seguinte fato, se uma nova palavra fosse criada, a mesma seria formada por sílabas, que são por sua vez compostas por vogais e consoantes anteriores e posteriores, as quais já teriam sido treinadas pelo sistema.

6.1 Contribuições

A principal contribuição deste trabalho está na elaboração de um novo Sistema de Decisão Hierárquica. Esta nova proposta tem como base as regras de classificação da fonética articulatória da língua portuguesa do Brasil. Basicamente, os fonemas são agrupados conforme suas classificações, ou seja, conforme a maneira de produção de cada fonema.

Nas fases de treinamento e classificação dos padrões, este trabalho apresenta outra contribuição através da utilização de fonemas, difones e sílabas como partes separadas (conjuntos), tendo como objetivo final o reconhecimento da palavra. Cada conjunto de reconhecimento pode ser considerado como uma máquina especialista ou um conjunto de máquinas especialistas, pois é treinado com propriedades exclusivas e com objetivo específico. Estas máquinas são constituídas basicamente por redes neurais do tipo Máquinas de Vetor de Suporte (SVM), as quais são agrupadas em uma estrutura similar a uma Máquina de Comitê do tipo Mistura Hierárquica de Especialistas [HAYKIN, 2001].

A utilização da transformada *Wavelet Packet* (WPT) com a escala Mel, em conjunto com os coeficientes MFCC, como descritores do sinal de voz também pode ser considerada como uma contribuição deste trabalho, uma vez que o uso dos dois descritores em conjunto, apresentou bons resultados.

Por fim, a estratégia de dividir para conquistar, aplicada diretamente a uma determinada sílaba, buscando primeiramente identificar a vogal como parte central e em seguida identificar as consoantes anterior e posterior, também é considerada como uma contribuição deste trabalho.

6.2 Trabalhos Futuros

O próximo passo sugerido é a implementação e teste com as outras 11 vogais. Assim, pode-se ter o sistema completo de reconhecimento para todas as vogais.

Outro passo importante seria acrescentar as classes (grupos) de consoantes anteriores, os encontros consonantais tais como: PRa, PLa, TRa, BRa, DLa, … etc. Estes encontros consonantais ocorrem somente com as consoantes Oclusivas e Fricativas.

Outra sugestão seria para o desenvolvimento de um estudo mais aprofundado para a melhoria da fase de separação de sílabas.

Como sugestão para melhoria do tempo de processamento, pode-se pesquisar a implementação do sistema em processamento paralelo e/ou em DSP específicos ou através de hardware dedicado FPGA (*field-programmable gate array*).

O Anexo I apresenta os experimentos com o reconhecimento de palavras isoladas, os quais foram importantes na escolha das melhores *Wavelet* mãe para o reconhecimento de voz. O Anexo II apresenta o diagrama de blocos completo do sistema desenvolvido. O Anexo III contém a lista das publicações aceitas em congressos e revistas internacionais que foram geradas durante a elaboração deste trabalho. Dentre eles encontra-se listado o trabalho sobre reconhecimento de voz audio-visual, o qual foi resultado do estágio de doutorado no exterior realizado no ano de 2007 na FEUP (Faculdade de Engenharia da Universidade do Porto) em Portugal.

7. Referências:

[ABDELATTY, 1998] Abdelatty Ali, A. M.; Spiegel, J. V. D.; e Mueller, Paul. **"An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants".** *IEEE Int. Conference on Acoustics, Speech, and Signal Processing*, ICASSP '98. Vol. 02, p. 961-964. May 1998.

[ABDELATTY, 1999] Abdelatty Ali, A.M.; Spiegel, J. V. D.; Mueller, Paul.; e Berman, Jeffrey. **"An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech"**. *IEEE International Symposium on Circuits and Systems*, ISCAS '99. Vol. 03, p. 118-121. May 1999.

[ABDULLA, 2003] W. Abdulla, V. Kecman, N. Kasabov**. "Speech-background classification by using SVM technique".** *in Proc. of Artificial Neural Networks and Neural Information Processing* ICANN/ICONIP 2003 International Conference, Istanbul, Turkey, (310-315), June 2003 .

[ABE, 2002] Abe, Shigeo e Inoue, Takuya. "**Fuzzy Support Vector Machines for Multiclass Problems".** *European Symposium on Artificial Neural Networks*, ESANN'02- Bruges (Belgium). p. 113-118, April 2002.

[ALCAIM, 2001] Alcaim, Abraham; e Santos, Sidney. C. B. dos. **"Sílabas como unidades fonéticas para o reconhecimento de voz em português".** *SBA Controle & Automação*. vol. 12, nº 01. 2001.

[ANDERSON, 1991] Anderson, Timothy R. **"Speaker independent phoneme recognition with an auditory model and a neural network: a comparison with traditional techniques".** *Acoustics, Speech, and Signal Processing*, 1991. ICASSP-91. vol.1, p.149-152. April 1991.

[BAUM, 1968] Baum, L. E.; e Petrie, T.; **"Statistical inference for probabilistic functions of finite state Markov chains".** *Ann. Math. Stat*., 37:1559-1563. 1966.

[BEKMAN, 1980] Bekman, Otto R.; Costa Neto, Pedro Luiz O. **Análise Estatística da Decisão.** Editora Edgard Blücher Ltda. São Paulo – SP. 1980.

[BERTSEKAS, 1995] Bertsekas, D. P. **"Dinamyc Programming and Optimal Control".** *Athenas Scientific.* vol. I and II, Belmont, MA 1995.

[BISHNU, 1976] Bishnu, Atal S. **"Automatic Recognition of Speakers From Their Voices".** *Proceedings of the IEEE*, vol. 64, nº 04, p. 460-475, April 1976.

[BLINN, 1993] Blinn, J. F. **"What's that deal with the DCT?."** *Computer Graphics and Applications*, IEEE.Vol.13, Issue 4, p.78-83. July 1993.

[BRIGHAM, 1998] Brigham, E. Oran. **The Fast Fourier Transform and its applications.** New Jersey, USA, *Prentice Hall - Signal Processing Series*. 1998.

[BURGES, 1998] Burges, Christopher J. C. **"A Tutorial on Support Vector Machines for Pattern Recognition".** Data Mining and Knowledge Discovery 2:121 - 167, 1998.

[BURKE, 1994] Burke, Barbara. **"The Mathematical Microscope: waves, wavelets and beyond".** In M. Bartusiak, et al, editor, *A Positron Named Priscilla, Scientific Discovering at the Frontier.* Chapter 7, pages 196-235. National Academy Press, Washington, DC – USA, 1994.

[BURRUS, 1998] Burrus, Sidney C.; Gopinath, Ramesh. A.; and Guo, H., **Introduction to Wavelets and Wavelet Transforms**. Prentice Hall, New Jersey. 1998.

[CAGLIARI, 1981] Cagliari, Luiz Carlos. **"Elementos da Fonética do Português Brasileiro"**. Tese de livre docência. Unicamp. Campinas – SP, Brasil. 1981.

[CHAN, 2001] Chan, C. P. Ching, P. C. e Lee, Tan**. "Noisy speech recognition using de-noised multiresolution analysis acoustic features".** *Journal Acoustical Society of America*. vol. 110., nº 05, pt. 01, p. 2567-2574. Nov. 2001.

[CLARKSON, 1999] P. Clarkson and P.J. Moreno, **"On the use of support vector machines for phonetic classification,"** in *Proc. ICASSP '99*, Phoenix, AZ USA, vol. 2, pp. 585–588, Mar. 1999.

[COIFMAN, 1992] Coifman, R. R.; Meyer, Y. e Wickerhauser, M. V. **Wavelet Analysis and signal processing.** In M. B. Ruskai et al. Editor *Wavelets and their applications,* Jones and Bartlett, Boston, 1992.

[COLLOBERT, 2004] R. Collobert and S. Bengio**; "Links Between Perceptrons, MLPs and SVMs".** *International Conference on Machine Learning*, ICML, 2004.

[COMBRINCK, 1996] Combrinck, H. P.; e Botha E. C., **"On The Mel-scaled Cepstrum".** *Proceedings of the Seventh Annual South African Workshop on Pattern Recognition*, University of Cape Town, Nov. 1996.

[COOLEY, 1965] Cooley, James W., and Tukey, John W. "**An algorithm for the machine calculation of complex**

Fourier series". Math. Comput. 19: 297–301. 1965.

[CRISTIANINI, 2003] Cristianini, Nello; e Shawe-Taylor, Jonh. **A Introdution to Support Vector Machine, and other kernel-based learning methods.** *Cambridge University Press,* United Kingdom, Ed. 3, 2003

[DAUBECHIES, 1988] Daubechies, Ingrid. **"Orthonormal bases of compactly supported wavelets"**. Communications on Pure and Applied Mathematics, 41: 909-996, November 1988.

[DAVENPORT, 1991] Davenport, Michel R.; e Garudadri, H. **"Neural net acoustic phonetic feature extractor based on Wavelets".** *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, p. 449-452, May 1991.

[DAVIS, 1952] Davis, K. H.; Biddulph, R.; e Balashek, S. **"Automatic Recognition of Spoken Digits"**. *Journal on Acoustics Soc. Am.*, 24 (6). p. 637-642, 1952.

[DAVIS, 1980] Davis, Steven B;. e Mermelstein, Paul. **"Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences".** *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, nº 04, p. 357-366, August 1980.

[DESHMUKH, 2002] Deshmukh, O.; Espy-Wilson, Carol Y.; e Juneja, A. **"Acoustic Phonetic Speech Parameters for Speaker-Independent Speech Recognition".** *ICASSP-2002*. Nº 2162. Speech Processing. May 2002.

[DUDA, 1973] Duda R.O. e Hart, P.E. **Pattern classification and scene analysis.** *John Wiley & Sons, New York*, 1973.

[FAROOQ, 2003] Farooq O. e Datta S. **"Phoneme recognition using wavelet based features".** *Elsevier, Information Sciences*, vol. 150, Issues 1-2, p. 5-15, March 2003.

[FECHINE, 2000] Fechine, Joseana M. **"Reconhecimento Automático da Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística".** *Tese de Doutorado.* Universidade Federal de Campina Grande -PB. 2000.

[FORGIE, 1959] Forgie, J. W.; e Forgie, C. D. **"Results Obtained From a Vowel Recognition Computer Program".** *Journal on Acoustics Soc. Am*., 31(11): 1480-1489, 1959.

[FRY, 1959] Fry, D. B. **"Theoretical Aspects of Mechanical Speech Recognition"**; e Denes P. **"The Design and Operation of the Mechanical Speech Recognizer at University College London"**. *Journal British Inst. Radio Engr.*, 19:4, p. 211-229, 1959.

[FUGIMURA, 1975] Fugimura, Osamu. **"Syllable as a unit of speech recognition".** *IEEE Trans. on Acoustics, Speech, and Signal Processing,* vol. 23, nº 01, p. 82-87. 1975.

[FUJISAKI, 1982] Fujisaki, H.; e Tominaga, M. **"Automatic recognition of voiced stop consonants in CV and VCV utterances".** *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*. vol.07, p. 1996-1999. May 1982.

[GOWDY, 2000] Gowdy, J.N. e Tufekci Z. **"Mel-scaled discrete wavelet coefficients for speech recognition".** *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 1351-1354. 2000.

[HAAR, 1910] A. Haar," **Zur Theorie der Orthogonalen Funktionen-Systeme".** Math. Ann, 69, pp. 331-371, 1910.

[HAYKIN, 2001] Haykin, Simon. **Redes Neurais: Princípios e prática**. 2º Edição, Porto Alegre, *Editora Bookman*, 2001.

[HOSOM, 2002] Hosom, John P. **"Automatic Phoneme Alignment Based on Acoustic-Phonetic Modeling".** *International Conference on Spoken Language Processing-ICSLP'02*, Boulder, Co., vol. I, p. 357-360, Sep. 2002.

[HOSSAIN, 1999] Hossain, M. I. Liu, James. e Lee, Raymond. **"A study of multilingual speech features: perceptive scalogram based on wavelet analysis".** *IEEE Int. Conference on Systems, Man, and Cybernetics*, SMC '99. Conference Proceedings. vol.02, p. 178-183, Oct. 1999.

[IPA, 2008] http://www2.arts.gla.ac.uk/IPA/ipachart.html.

[ITAKURA, 1975] Itakura, F. **"Minimum Prediction Residual Applied to Speech Recognition"**. *IEEE Trans. Acoustics, Speech, Signal Proc.,* ASSP-23 (1): 67-72, February 1975.

[JIANG, 2003] Jiang, Hai. Er, Meng Joo. e Gao, Yang. **"Feature extraction using wavelet packets strategy".** *Proceedings 42º IEEE Conference on Decision and Control*. vol.5, p. 4517-4520. Dec. 2003.

[JUNEJA, 2003] Juneja, A.; e Espy-Wilson, C. **"Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines".** *Proceedings of International Joint Conference on*

Neural Networks, Portland, Oregon, 2003.

[KEPUSKA, 1989] Kepuska, Veton Z.; e Gowdy, John N**. "Investigation of phonemic context in speech using self-organizing feature maps".** *ICASSP-89*. vol.01, p. 504-507, May 1989.

[KIM, 2000] Kim, Kidae. Youn, Dae Hee e Lee Chulhee. **"Evaluation of wavelet filters for speech recognition".** *IEEE International Conference on Systems, Man, and Cybernetics*. vol.4 p.2891 – 2894. Oct. 2000.

[LATHI, 1987] Lathi, B. P. **Sistemas de Comunicação**. Rio de Janeiro - RJ, Ed. Guanabara, 1987.

[LEVISON, 2005] Levison, Stephen C. **Mathematical Models for Speech Technology.** West Sussex, England. Ed. John Wiley & Sons Ltd. 2005.

[LIPPMANN, 1997] Lippmann, Richard P. **"Speech recognition by machines and humans".** *Speech Communication*, 22-01, p. 01-15. 1997.

[LONG, 1996] Long, C. J.; e Datta, S. **"Wavelet Based Feature Extraction for Phoneme Recognition".** *Proceedings' ICSLP '96*. 1996.

[LUKASIK, 2000] Lukasik, E. **"Wavelet packets based features selection for voiceless plosives classification".** *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00*. vol.2, p. II689-II692, June 2000.

[MAIA, 1985] Maia, Eleonora Albano da Mota. **No Reino da fala: A línguagem e seus Sons.** São Paulo, *Ed. Ática*, 1985.

[MALBOS, 1994] Malbos, F.; Baudry, M.; e Montresor, S. **"Detection of stop consonants with the wavelet transform".** *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, p. 612-615, Oct. 1994.

[MARCHESI, 1996] Marchesi, B. Lippmann, L., Jr. e Nohama, P. **"Voice recognition method applied to Brazilian vowels".** *Proceedings of the 18th Annual International Conference of the IEEE. Engineering in Medicine and Biology Society*. Vol. 04, p. 1542-1543. Nov. 1996.

[MARINHO,2004] Marinho, Rafael.; Teixeira Jr, Talisman.; e Klautau, Aldebaro. **"Classificação Fonética Usando SVM e Seleção de Parâmetros".** *XXI Simpósio Brasileiro de Telecomunicações*, SBT'04, Belém, PA, Brasil. 2004.

[MARTIN, 1964] Martin, T. B.; Nelson, A. L.; e Zadell, H. J. **"Speech Recognition by Feature Abstraction Techniques".** *Tech. Report* Al-TDR-64-176, Air Force Avionics Lab, 1964.

[MCCULLOCH, 1943] McCulloch, W. S.; e Pitts, W. **"A logical calculus of the ideas immanent in nervous activity".** *Bulletin of Mathematical Biophysics.* vol. 5, pp. 115-133, 1943.

[MENG, 1991] Meng, Helen. M.; e Zue, Vitor W. **"Signal representation comparison for phonetic classification".** *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91*, vol.01, p. 285-288 April 1991.

[MERCER, 1909] Mercer, J. **"Functions of Positive and Negative type, and their connection with the theory of integral equations".** *Transactions of the London Philosophical Society,* vol. 209, pp. 415-446. 1909.

[MEYER, 1993] Meyer, Y. **Wavelets: Algorithms and Applications.** Philadelphia, SIAM. 1993.

[MINSKY, 1969] Minsky, M. L.; e Papert S. A. **Perceptrons.** Cambridge, MA: MIT Press. USA. 1969.

[MORETTIN, 1999] Morettin, Pedro A. **Ondas e Ondaletas: Da análise de Fourier à Análise de Ondaletas.** São Paulo. Ed. *Edusp. Universidade de São Paulo*. 1999.

[MPORAS, 2006] I. Mporas, T. Ganchev, P. Zervas, N. Fakotakis: **"Recognition of Greek Phonemes using Support Vector Machines".** *SETN 2006*, Crete, Greece.2006.

[NAGATA**,** 1963] Nagata, K.; Kato, Y.; e Chiba, S. **"Spoken Digit Recognizer for Japanese Language".** *NEC Res. Develop.,* No. 6, 1963.

[NILSSON, 1965] Nilson, N. J. **Learning Machines: Foundations of Trainable Pattern-Classifying Systems.** McGraw-Hill, New York, 1965.

[OLSON, 1956] Olson, H. F.; e Belar, H. **"Phonetic Typewriter".** *Journal on Acoustics Soc. Am*., 28(6): 1072-1081, 1956.

[OSHERSON, 1990] Osherson, D. N.; Weinstein, S. e Stoli, M. **Modular Learning.** Computacional Neuroscience, ed. E. L Schwartz, MIT Press, Cambridge, MA – USA, 1990.

[PAYTON, 1985] Payton, Karen L**. "Speech processing by a model of the auditory periphery***". Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85*. vol. 10, p.1109-

1112, Apr 1985.

[PICONE, 2002] J. Picone. e A. Ganapathiraju, **"Support Vector Machines For Automatic Data Cleanup"** *Proceedings of the International Conference of Spoken Language Processing*, vol. 4, pp. 210-213, Beijing, China, October 2000.

[PRIEBE, 1994] Priebe, R. D. e Baugh, K. W. **"Wavelet based detectors".** *IEEE- Acoustics, Speech, and Signal Processing, ICASSP-94.* vol. 04, p.:IV/105-IV/108. April 1994.

[PROTAZIO, 2002] Protazio, João Marcelo Brazão. **Análise Wavelet Aplicada a Sinais Geofísicos.** *Dissertação de Mestrado.*Unicamp – Campinas-SP, 2002.

[RABINER, 1975] Rabiner, Lawrence R. e Schafer, R. W. ***"Digital representations of speech signals".*** Proc. IEEE, vol. 63, pp. 662-677, 1975.

[RABINER, 1989] Rabiner, Lawrence R. **"A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition".** *Proceedings of the IEEE*, 77 (2), p. 257–286, February 1989.

[RABINER, 1993] Rabiner, Lawrence R.; e Juang, Biing-Hwang. **Fundamentals of Speech Recognition.** New Jersey, USA. Ed. Prentice Hall, 1993.

[RAMAN, 1984] Raman, S.; e Yegnanarayana, B. **"Performance of isolated word recognition system for confusable vocabulary".** *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '84*. Vol. 09, Part 1, p. 17-20.Mar 1984.

[RANGOUSSI, 1995] Rangoussi, Maria; e Delopoulos, Anastasios. **"Recognition of unvoiced stops from their time-frequency representation".** *International Conference on Acoustics, Speech, and Signal Processing, ICASSP-95*, vol. 01, p. 792-795, May 1995.

[REDDY, 1966] Reddy, D. R. **"An Approach to Computer Speech Recognition by Direct Analysis of Speech Wave".** *Tech. Report* No. C549, Computer Science Dept., Stanford Univ., September 1966.

[ROCHA, 2005] Rocha, Oséas P. Thomé, Antonio C. G. e Barros, Sandro R. R**. "Reconhecimento de voz utilizando Wavelet e Classificador Neural".** *VII Congresso Brasileiro de Redes Neurais – CBRN*. Natal- RN. Out. 2005.

[RODRIGUES, 2002] Rodrigues, Gustavo F.; e Yehia, Hani C**. "Caracterização acústica das Vogais do Português Brasileiro visando a Normalização de Locutores".** *Seminário de Engenharia de Áudio* - Universidade Federal de Minas Gerais - Belo Horizonte - Junho de 2002.

[ROSENBERG, 1983] Rosenberg, Aaron E. Rabiner, Lawrence R. Wilpon, Jay G. e Kahn, Daniel. **"Demisyllable-Based Isolated Word Recognition System".** *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 31, nº 03, p. 713-726. 1983.

[ROSENBLATT, 1958] Rosenblatt, F. **"The Perceptron: A probabilistic model for information storage and organization in the brain".** *Psychological Review*, vol. 65, pp. 306-408. 1958.

[RUMELHART, 1986] Rumelhart, D. E.; Hinton, G. E; e Williams, R. J. **"Learning representations of back-propagation errors".** *Nature,* London – UK, vol. 323, pp. 533-536. 1986.

[RUSKE, 1982] Ruske, Günther. **"Automatic recognition of Syllabic Speech Segments using Spectral and Temporal Features".** *IEEE, Proceedings ICASSP*, vol. 01, p. 550-553. 1982.

[RUSSELL, 2003] Russell, Martin J. e Bilmes, Jeff A. **"Introduction to the special issue on new computational paradigms for acoustic modeling in speech recognition".** *Editorial, Computer Speech and Language*, nº 17, p. 107-112, March 2003.

[RUSSO, 1993] Russo, Ieda e Behlau, Mara. **Percepção da fala: análise acústica do português brasileiro.** São Paulo, Editora Lovise, 1993.

[SAKAI, 1962] Sakay, T.; e Doshita, S. **"*The Phonetic Typewrite – Information Processing*".** *Proc. IFIP Congress,* Munich, 1962.

[SAKOE, 1978] Sakoe, H.; e Chiba, S. **"*Dynamic Programming Algorithm Optimizations for Spoken Word Recognition*".** *IEEE Trans. Acoustics, Speech, Signal Proc.* ASSP-26 (1): 43-49, February 1978.

[SAKOE, 1979] Sakoe, H. **"*Two Level DP Matching – A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition*".** *IEEE, Trans. Acoustics, Speech, Signal Proc.* ASSP-27: 588-595, December 1979.

[SARIKAYA, 1998] Sarikaya, Ruhi. e Gowdy, John N. **"Subband based classification of speech under stress".** Proc. of the *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '98*. vol. 01, p.569-572. May 1998.

[SARIKAYA, 2000] Sarikaya, Ruhi.; e Hansen, John H. L. **"High resolution speech feature Parametrization for monophone-based stressed Speech recognition".** *Signal Processing Letters, IEEE*. vol. 07, Issue 07, p.182-185. July 2000.

[SCHOLKOPF, 1995] Scholkopf, B.; Burges, C. J.; e Vapnik, V. "**Extracting support data for a given task".** *Proc. 1st International Conference on Knowledge Discovery and Data Mining* (pp. 252 – 257). Menlo Park, CA - 1995.

[SILVA, 2003] Silva, Thais Cristofáro. **Fonética e Fonologia do Português.** 7º Edição, São Paulo, *Ed. Contexto*, 2003.

[STEVENS, 1937] Stevens, S. S. Volkman, J. e Newman, E. B. **"A Scale for Measurement of the Psychological Magnitude Pitch".** *Journal of the Acoustical Society of America*, vol. 08, pp. 185-190, Jan. 1937.

[SUZUKI, 1961] Suzuki J.; e Nakata, K. **"*Recognition of Japanese Vowels – Preliminary to the Recognition of Speech*".** Journal *Radio Res. Lab.* 37 (8): 193-212, 1961.

[TAN, 1996] Tan, Beng T.; Fu, Minyue; Spray, A.; e Dermody, Philipy. **"The use of wavelet transforms in phoneme recognition"**. *Fourth International Conference, on Spoken Language*, ICSLP 96. Proceedings. vol. 04, p. 2431-2434, Oct. 1996.

[THUONG, 1997] Thuong, Le-Tien. **"Some Issues of Wavelet Functions for Instantaneous Frequency Extraction in Speech Signal".** *IEEE TENCON- Speech and Image Technologies for Computing and Telecommunications*. p. 31-34. 1997.

[VAPNIK, 1971] Vapnik, Vladimir. N.; e Chervonenkis, Ya. **"On the uniform convergence of relative frequencies of events to theirs probabilities"**. *Theoretical Probability and Its Applications*, vol. 17, pp. 264-280. 1971.

[VAPNIK, 1982] Vapnik, Vladimir N., **Estimation of dependences based on empirical data.** Springer-Verlag New York. 1982.

[VAPNIK, 1992] Vapnik, Vladimir. N**. "Principles of risk minimization for learning theory"**. *Advances in Neural Information Processing Systems.* Vol. 04, pp.831-838, San Mateo, CA. 1992.

[VAPNIK, 1995] Vapnik, Vladimir N. **The Nature of Statistical Learning Theory**. Springer-Verlag, 1995.

[VELICHKO, 1970] Velichko, V. M.; e Zagoruyko, N. G. **"Automatic Recognition of 200 words".** Int. J. Man-Machine Studies, 2:223, June 1970.

[VINTSYUK, 1968] Vintsyuk, T. K**. "Speech Discrimination by Dynamic Programming".** *Kibernetika,* 4 (2): 81-88, 1968.

[WAIBEL, 1989] Waibel, Alexander. Hanazawa, Toshiyuki. Hinton, Geoffrey. e Shikano, Kiyohiro. **"Phoneme recognition using time-delay neural networks".** *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 37, nº 03, p.328-339. March 1989.

[WENDT, 1996] Wendt, Christopher. e Petropulu, P. Athina. **"Pitch determination and speech segmentation using the discrete wavelet transform".** *IEEE International Symposium on Circuits and Systems,* ISCAS '96. vol. 02, p. 45- 48, May 1996.

[YANG, 2001] Yang, Song. Er, Meng Joo e Gao Yang. **"A high performance neural-networks-based speech recognition system".** *International Joint Conference on Neural Networks*, IJCNN '01. vol. 2 p. 1527-1531. July 2001.

[YOUNG, 1996] Young, S. **"A Review of Large-Vocabulary Continuous-Speech Recognition".** *IEEE Signal Processing Magazine*, pp. 45-57. September 1996.

[ZHU, 2000] Zhu Q. e Alwan, Abeer. **"On the use of variable frame rate analysis in speech recognition".** *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '00. vol.03, p. 1783-1786. June 2000.

Anexo I: Experimento: Reconhecimento de Dígitos

Dois experimentos com o reconhecimento de dígitos (números de zero a nove) foram realizados. O primeiro experimento foi o realizado no modo dependente do locutor e o segundo experimento foi realizado no modo independente do locutor. Os resultados obtidos nestes experimentos constam da segunda publicação gerada por este trabalho, a qual é citada no Anexo III. Alguns parâmetros utilizados neste teste de reconhecimento são diferentes dos descritos no Capítulo V.

Reconhecimento de Dígitos: Dependente do Locutor

Além do reconhecimento de dígitos, outro objetivo deste experimento foi encontrar qual a *Wavelet* mãe que apresenta os melhores resultados no reconhecimento de voz. Além disso, os resultados foram também obtidos com o descritor MFCC para efeito de comparação.

O sistema de reconhecimento desenvolvido é composto por 10 máquinas especialistas SVM treinadas no modo "um contra todos" com *kernel* polinomial e constante *C*=20. A Figura AI.1 apresenta o sistema desenvolvido. Como a rede neural SVM exige que os descritores tenham sempre um mesmo tamanho, foi realizado sobre cada palavra gravada um sistema de janelamento dinâmico [ZHU, 2000]. Isto se dá pelo fato de que o sinais das palavras (dígitos) de zero a nove possuem tamanhos diferentes.

O janelamento dinâmico é simples. Basicamente, divide-se o maior sinal de voz (para este caso a palavra "quatro") em janelas de 30ms com superposição de 50%, o objetivo aqui é respeitar o teorema da Amostragem [LATHI, 1987]. O número de janelas obtido do maior sinal é então fixado para os outros sinais, que para este exemplo é igual a 40 janelas. Deste modo, todos os sinais terão sempre 40 janelas, mas o tamanho de cada janela irá variar conforme o tamanho do sinal. Por exemplo, para o caso do maior sinal (palavra "quatro") as 40 janelas possuem 30 ms cada uma. Já para o caso do menor sinal (palavra "um"), as 40 janelas possuem um tamanho aproximado de 22 ms.

Deste modo, os descritores terão sempre a mesma dimensão pois são obtidos através da concatenação dos descritores obtidos nas 40 janelas. A dimensão final do descritor é de 1160, pois para cada janela o descritor possui uma dimensão de 29.

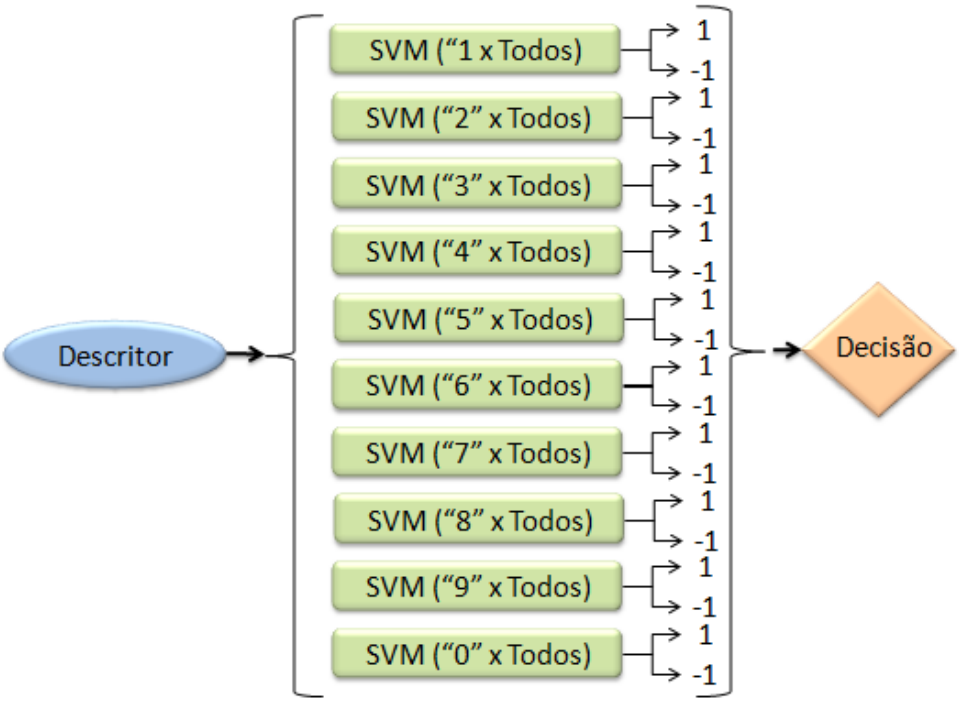

Figura AI.1: Sistema para o reconhecimento de dígitos (números de zero a nove).

Para o modo dependente do locutor, 700 arquivos foram gravados cada um contendo os dez sinais das palavras de zero a nove. Deste modo, 7000 sinais compõem o banco de dados que foi gravado por apenas um único locutor no período de três meses. Cada sinal possui somente um descritor obtido através do janelamento dinâmico descrito acima.

Com o objetivo de melhor avaliar o sistema, o método de validação cruzada foi aplicado para cada um dos tipos de descritores testados. Deste modo, o banco de dados foi dividido em quatro partes contendo cada um 25% dos descritores. Assim, em cada um dos quatro testes, 25% dos descritores são utilizados no treinamento das máquinas especialistas e os outros 75% dos descritores são utilizados na classificação.

A Tabela AI.1 apresenta os resultados obtidos em cada um destes testes de validação cruzada (G1, G2, G3 e G4) e o resultado final é apresentado na última coluna, o qual foi obtido pela média destes quatro testes de validação cruzada.

Dezenove diferentes experimentos foram realizados, sendo dezesseis com diferentes *Wavelets* mãe. O experimento número 18 foi realizado com o classificador MLP (*Multi Layer Perceptron*) treinado com o algoritmo *back propagation* com três camadas de [1160-20-10] neurônios. O experimento número 17 foi uma tentativa de utilizar a mesma rede neural MLP para resolver os problemas de empate encontrados pelas máquinas SVM do experimento 10. Nestes três experimentos (10, 17 e 18), a *Wavelet* mãe utilizada foi a que apresentou os

melhores resultados a Daubechies 5. Por fim, o experimento 19 foi realizado com os descritores obtidos dos coeficientes MFCC utilizando o classificador SVM.

Tabela AI.1: Reconhecimento de dígitos no modo dependente do locutor.

Experimentos	Descritores	Resultados: Validação Cruzada				Resultado
		% G1	% G2	% G3	% G4	% Média
1	Coiflet 1	93.11	92.39	93.67	92.83	93.00
2	Coiflet 2	95.56	93.72	94.06	94.11	94.36
3	Coiflet 3	95.28	94.28	94.11	94.00	94.42
4	Coiflet 4	95.94	94.56	95.11	94.06	94.92
5	Coiflet 5	97.06	96.50	97.11	96.56	96.81
6	Daubechies 1	92.06	92.11	91.50	91.17	91.71
7	Daubechies 2	93.89	93.72	93.44	93.83	93.72
8	Daubechies 3	94.17	92.89	93.89	93.06	93.50
9	Daubechies 4	94.94	93.56	93.67	93.56	93.93
10	Daubechies 5	97.44	98.22	97.67	98.28	97.90
11	Daubechies 6	95.72	94.17	94.22	93.61	94.43
12	Daubechies 7	94.72	93.78	94.33	93.83	94.17
13	Daubechies 8	95.33	94.83	94.83	93.56	94.64
14	Meyer	97.83	96.61	97.44	96.78	97.16
15	Biort 1.1	92.06	92.11	91.50	91.17	91.71
16	Biort 2.2	90.94	89.33	91.11	89.67	90.26
17	Daub. 5 (SVM + MLP)	98.94	99.19	98.67	99.23	99.00
18	MLP (Daub. 5)	82.89	84.28	83.56	80.78	82.87
19	MFCC com SVM	97.53	97.85	98.30	97.33	97.75

Dos resultados acima, pode-se ver que o melhor desempenho foi obtido com a *Wavelet* mãe *Daubechies* 5, inclusive melhor dos que os coeficientes MFCC. O pior desempenho foi obtido com a rede neural MLP. No entanto, quando utilizou-se esta mesma rede neural para resolver os problemas de empate apresentados pelas máquinas SVM, o resultado foi o melhor dentre todos. As *Wavelets Mayer* e *Coiflet* 5 também apresentaram bons resultados.

Reconhecimento de Dígitos: Independente do Locutor

Para o modo independente do locutor quatro experimentos foram realizados com os quatro melhores descritores encontrados no experimento no modo dependente do locutor, sendo: as *Wavelets* mãe Daubechies 5, Mayer e Coiflet 5 além dos coeficientes MFCC. A estrutura do classificador utilizado foi a mesma do sistema do modo dependente do locutor apresentado na Figura AI.1 com as mesmas características.

Um banco de dados foi gravado durante três meses com 82 locutores do sexo masculino de idades entre 18 e 45 anos. Cada locutor gravou 10 arquivos com as dez palavras (zero a nove) gravadas em sequência. Deste modo, 8200 arquivos individuais foram gravados.

Novamente, a validação cruzada foi utilizada para melhor visualizar os resultados e o banco de dados foi dividido em quatro grupos de 25% (G1, G2, G3 e G4). A Tabela AI.2 apresenta os resultados obtidos nos quatro testes de validação cruzada (colunas), para cada um dos quatro descritores utilizados (linhas).

Tabela AI.2: Reconhecimento de dígitos no modo independente do locutor.

Experimento	Descritor	% G1	% G2	% G3	% G4	% Média
1	Coiflet 5	85.06	86.70	87.23	86.45	86.36
2	Daubechies 5	93.24	92.12	93.16	93.38	92.97
3	Meyer	88.38	89.78	91.14	91.17	90.12
4	MFCC	92.11	93.15	93.70	92.55	92.87

Novamente a *Wavelet* mãe *Daubechies* 5 obteve o melhor resultado seguida de perto pelos coeficientes MFCC. Vale salientar, que os resultados para o modo independente do locutor têm um desempenho bem inferior ao modo dependente do locutor devido a diversidade dos locutores.

Anexo II: Diagrama de Blocos do Sistema Desenvolvido

Anexo III: Publicações

As publicações citadas abaixo foram aceitas e publicadas nos respectivos congressos ou revistas nacionais e internacionais. O último trabalho citado teve seu aceite confirmado mas a publicação será feita apenas após a realização do respectivo congresso.

01- Bresolin, Adriano de Andrade; Dória Neto, Adrião D.; Alsina, Pablo Javier **Reconhecimento de Fonemas, utilizando análise multiespectral com SVM em uma estrutura de máquina de comité hierárquica.** *XVI- CBA Congresso Brasileiro de Automática,* SALVADOR – BA, Brasil, 2006.

02- Bresolin, Adriano de Andrade; Dória Neto, Adrião D.; Alsina, Pablo Javier**. A New Hierarchical Decision Structure using Wavelet Packet and SVM for Brazilian Phonemes Recognition.** *Lecture Notes in Computer Science*, v. 4233, p. 159/166, 2006.

03- Bresolin, Adriano de Andrade; Dória Neto, Adrião D.; Alsina, Pablo Javier. **Brazilian Vowels Recognition Using a New Hierarchical Decision Structure with Wavelet Packet and SVM**. *ICASSP/07. IEEE International Conference on Acoustic, Speech and Signal Processing Letters*, v. 2, p. 493/10.1109-496, Honolulu, Hawaii – USA, 2007.

04- Bresolin, Adriano de Andrade; Diamantino Rui da Silva Freitas; Dória Neto, Adrião D.; Alsina, Pablo Javier; Pera, Vitor. **European and American Audio-Visual Speech Recognition using SVM in Portuguese language.** *In: DCC/2008 - Data Compression Conference, 2008. Proceedings to be published by the IEEE Computer Society Press*, Snowbird, Utah, USA. 2008.

05- Bresolin, Adriano de Andrade; Dória Neto, Adrião D.; Alsina, Pablo Javier. **Digit Recognition using Wavelet and SVM in Brazilian Portuguese.** *ICASSP/08. IEEE International Conference on Acoustic, Speech and Signal Processing.* Las Vegas, Nevada-USA 2008.

06- Bresolin, Adriano de Andrade; Dória Neto, Adrião D.; Alsina, Pablo Javier. **Consonantal Recognition using SVM and New Hierarchical Decision Structure based in the Articulatory Phonetics.** *IEEE – ISM/2008, IEEE International Symposium on Multimedia.* Berkeley, California-USA. Dezembro, 2008.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo