Universidade Federal do Ceará
Departamento de Engenharia de Teleinformática
Curso de Graduação em Engenharia de Teleinformática
Projeto de Conclusão do Curso

Raphael Torres Santos Carvalho

Estudo Comparativo de Técnicas de Extração de Características para Reconhecimento de Fonemas

Fortaleza – Ceará
Abril 2009

Raphael Torres Santos Carvalho

Estudo Comparativo de Técnicas de Extração de Características para Reconhecimento de Fonemas

Projeto de Conclusão de Curso apresentado à Coordenação do Curso de Graduação em Engenharia de Teleinformática da Universidade Federal do Ceará como parte dos requisitos para obtenção do grau de ***Engenheiro de Teleinformática****.*

Área de Concentração: Sinais e Sistemas

Orientador : Prof. Dr. Paulo César Cortez

Co-orientador : M.Sc. Rodrigo Carvalho Souza Costa

Fortaleza – Ceará
Abril 2009

Resumo

As tecnologias de reconhecimento de voz, tanto na área de software quanto de hardware, vêm sendo desenvolvidas a passos largos, criando uma grande expectativa sobre a sua futura utilização e o impacto que devem causar em todos: usuários, empresas e instituições de ensino. Neste trabalho, é realizado um estudo comparativo de métodos de extração de características para detecção de fonemas, utilizando técnicas existentes na literatura e um método proposto baseado em *wavelets*. O método proposto consiste em uma modificação do método proposto por Farooq e Datta (2003) que consiste em aplicar a Transformada *Wavelet* para extração de características. Para os métodos estudados, são analisados o esforço computacional e o desempenho de classificação utilizando Redes Neurais Artificiais (RNA).

Palavras-chaves: Reconhecimento de Voz, Extração de Características, Fonemas, Redes Neurais Artificiais, Transformada *Wavelet*.

Abstract

The technologies of voice recognition, both software and hardware, have been developed at a fast pace, creating a large expectations about its future use and the impact that it should cause on everyone: users, enterprises and educational institutions. In this work, is conducted a comparative study of methods of feature extraction to detect phonemes using existing techniques in the literature and a proposed method based on wavelets. The method proposed is a modification of the method proposed by Farooq and Datta (2003) that consists of applying the Wavelet Transform to feature extraction. For the methods studied, the stress computational and performance of classification using Artificial Neural Networks are analyzed.

Keywords: Voice Recognition, Feature Extraction, Phonemes, Artificial Neural Networks, Wavelet Transform.

Dedico este trabalho a Deus e a minha família.

Agradecimentos

Agradeço primeiramente a Deus por todas as bençãos derramadas durante toda minha vida.

À minha família pelo carinho, apoio e incentivo que me permitiram chegar até aqui.

Aos amigos de faculdade que me ajudaram a crescer como pessoa e que estiveram presentes durante essa longa caminhanda em busca do diploma em Engenharia de Teleinformática.

Aos amigos e companheiros bolsistas e professores membros do projeto SMTRG, pela amizade, os quais pude conviver diariamente e me ajudaram de várias formas à desenvolver este trabalho.

Ao Professor Dr. Paulo César Cortez, meu Orientador Científico, pela amizade e a oportunidade de receber um pouco de sua sabedoria, pela disponibilidade apresentada e pelas condições que me proporcionou na realização deste trabalho.

Ao M. Sc. Rodrigo Carvalho Souza Costa, meu Co-orientador Científico, pela amizade, paciência e ajuda ao longo deste trabalho.

Por fim, aos demais professores do Departamento de Engenharia de Teleinformática pelos conhecimentos transmitidos e ajudas prestadas.

Não existe um caminho para a felicidade. A felicidade é o caminho.

Mahatma Gandhi

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

LP Predição Linear (*Linear Prediction*)

LPC Codificação Linear Preditiva (*Linear Predictive Coding*)

PDS Processamento Digital de Sinais

AR Auto-Regressivo (*Autoregressive*)

MMQ Método dos Mínimos Quadráticos

CWT Transformada *Wavelet* Contínua (*Continuous Wavelet Transform*)

DWT Transformada *Wavelet* Discreta (*Discrete Wavelet Transform*)

STFT Transformada de Fourier de Tempo Curto (*Short Time Fourier Transform*)

FFT Transformada Rápida de Fourier (*Fast Fourier Transform*)

RP Reconhecimento de Padrões

RNA Rede Neural Artificial

MLP Perceptron Multi-Camadas(*Multilayer Perceptron*)

MSE Erro Quadrático Médio(*Mean Squared Error*)

EEG Eletroencefalograma

Capítulo 1

Introdução

O reconhecimento da fala consiste em identificar fonemas, sílabas, palavras para formar a mensagem original, ou uma informação na qual existe uma seleção mais direta da resposta, sem interpretação dela. Assim, uma ação pode ser executada diretamente quando um padrão falado é reconhecido (PAULA, 2000).

O reconhecimento automático da voz consiste no processo de extrair automaticamente a informação lingüística do sinal da fala, a qual está codificada. Este processo normalmente acontece em três etapas conforme mostrado na Figura 1.1 (PAULA, 2000).

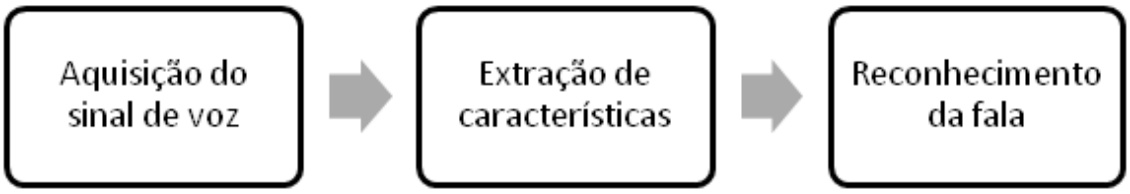


Figura 1.1: etapas do processo de reconhecimento de voz.

A primeira etapa consiste em obter digitalmente o sinal de áudio e convertê-lo para um padrão a ser utilizado pelas etapas seguintes.

A segunda etapa realiza a extração das características que descrevem adequadamente o sinal de voz ou o representam de forma mais compacta. Esta etapa é necessária devido ao fato do sinal de voz não ser usado diretamente para alimentar a etapa de reconhecimento, exceto quando sub-amostrado. Geralmente, o sinal de voz é ruidoso e pode possuir milhares de amostras, tornando difícil seu armazenamento e aumentando a complexidade do classificador.

A etapa de reconhecimento da fala consiste em classificar essas características e, em geral, é dividida em treinamento e classificação. No treinamento, as características dos fonemas são utilizadas para determinar um modelo que represente cada classe. A etapa de classificação usa o modelo gerado no treinamento para determinar qual fonema é pronunciado.

1.1 Motivação

As tecnologias de reconhecimento de voz, tanto na área de software quanto de hardware, vêm sendo desenvolvidas a passos largos, criando uma grande expectativa quanto à sua futura utilização e quanto ao impacto que causarão em todos: usuários, empresas e instituições de ensino (BATHAGLINI, 2009).

O desenvolvimento dessas tecnologias deve permitir a total interação entre usuários e computadores através de comandos de voz, eliminando, assim, a utilização dos atuais dispositivos de entrada de dados.

A interação homem-máquina sendo realizada através da voz, as pessoas com necessidades físicas especiais, como os deficientes visuais e os tetraplégicos, podem utilizar computadores com essa tecnologia para realizar praticamente qualquer trabalho de forma tão produtiva quanto as pessoas sem deficiência.

Os computadores não devem ser os únicos sistemas utilizando os métodos de reconhecimento de voz. Carros guiados pela fala, elevadores que se movem para um andar específico apenas com uma palavra e luzes que acendem ou apagam através da voz, são apenas alguns dos exemplos de tecnologias futuras que irão facilitar a vida do ser humano.

Atualmente, existem diversas técnicas de extração de características da voz para reconhecimento de fonemas presentes na literatura como banco de filtros de amplitude, Predição Linear (LP) seguida por ponderação percentual, Transformada Rápida de Fourier (FFT), coeficientes cepstrais e Transformada *Wavelet* (FAROOQ; DATTA, 2003). Em virtude disso, neste trabalho é estudado um conjunto destas técnicas que utilizam as Rede Neural Artificial (RNA) como classificadores e que podem ser utilizados nessas aplicações, analisando o desempenho computacional e os resultados obtidos.

1.2 Objetivos

O objetivo geral deste trabalho é realizar um estudo comparativo entre técnicas de extração de características para reconhecimento de fonemas com Rede Neural Artificial.

No desenvolvimento deste trabalho outros objetivos específicos são alcançados:

i. revisão bibliográfica;

ii. pesquisa de técnicas de extração de características baseadas em Predição Linear e Transformada *Wavelet*;

iii. pesquisa de redes neurais artificiais;

iv. estudo sobre o ambiente de simulação Matlab/Simulink;

v. simulação e implementação dos algoritmos estudados;

vi. avaliação do desempenho de classificação e esforço computacional dos métodos estudados;

vii. avaliação dos resultados.

1.3 Organização da Monografia

Esta monografia está dividida em 5 capítulos. O Capítulo 2 apresenta uma retrospectiva sobre o princípio de reconhecimento de fala e as principais técnicas de extração de características para reconhecimento de fonemas. O Capítulo 3 descreve as ferramentas utilizadas para a realização de trabalho e o método proposto. O Capítulo 4 descreve os resultados dos testes e por fim, o último capítulo apresenta as considerações finais e as perspectivas de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo são descritos as ferramentas matemáticas e os princípios de extração de características. Inicialmente são descritos os fundamentos da produção da fala e, em seguida, são descritas duas ferramentas matemáticas muito importantes na extração de características da fala, a Predição Linear (LP) e a Transformada *Wavelet*. Na seção 2.4, são descritas as técnicas de extração de características utilizando estas ferramentas matemáticas. Ao final do capítulo, é descrito o fundamento de reconhecimento de padrões utilizando Redes Neurais Artificiais (RNA).

2.1 Modelagem da Produção da Fala

A fala é uma das capacidades ou aptidões que os seres humanos possuem de comunicação, manifestando seus pensamentos, opiniões e sentimentos através dos vocábulos. Consiste no principal sinal entre os distintos sinais abordados pela linguagem natural, como por exemplo, ideogramas, gestos, gritos, trejeitos e outros tipos de linguagem corporal (PAULA, 2000).

Existem duas principais fontes de características da fala específicas aos locutores, as físicas e as adquiridas (ou aprendidas). As características físicas relacionam-se principalmente ao trato vocal, estrutura formada pelas cavidades que vão das pregas vocais até os lábios e o nariz (SOUZA JÚNIOR, 2009). A Figura 2.1 ilustra o conjunto de órgãos que formam o trato vocal e compõem o sistema de produção da fala.

Na produção da fala, as cordas vocais situadas na laringe são excitadas pelo ar vindo dos pulmões. A vibração das pregas vogais geradas devido à passagem

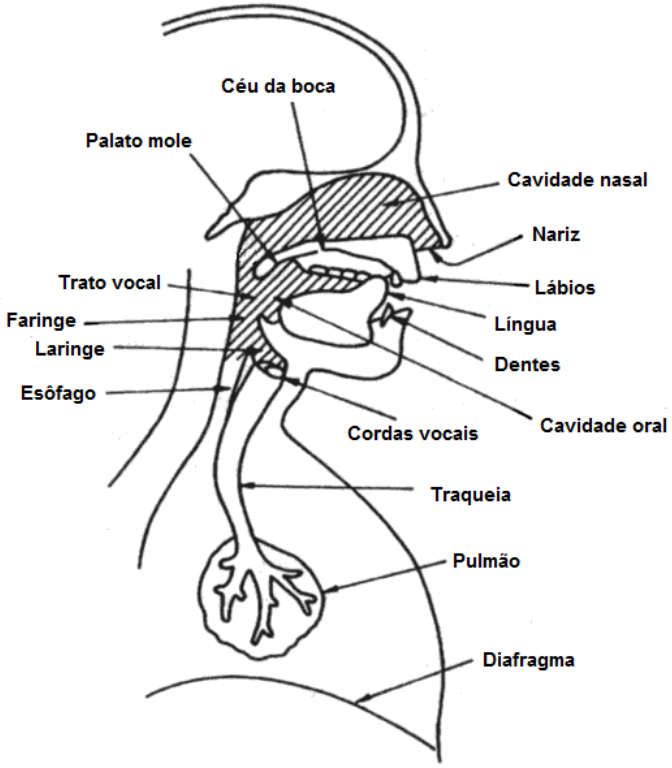


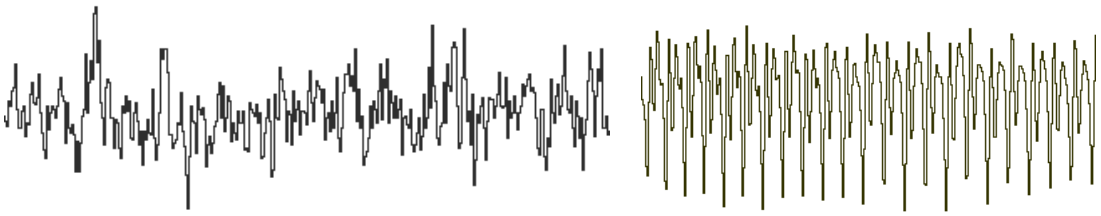
Figura 2.1: sistema de produção da fala (SOUZA JÚNIOR, 2009).

do fluxo de ar produz um som fraco e constituído de poucos harmônicos, que é amplificado quando passa pelas cavidades de ressonância (laringe, faringe, boca e nariz) e ganha "forma" final quando é articulado através de movimentos de língua, lábios, mandíbula, dentes e palato (SOUZA JÚNIOR, 2009).

Essa passagem pelas cavidades do trato vocal altera o espectro do som devido as ressonâncias, que formam picos de energia no espectro de frequência conhecidos como formantes. Através da análise espectral da fala produzida é possível estimar a forma do trato vocal.

Uma classificação comum dos eventos sonoros é feita quanto ao estado de vibração das cordas vocais. Adota-se uma convenção de três estados: silêncio, vozeados (sonoros) e não-vozeados (surdos). O silêncio representa a etapa em que nenhum som é produzido. Os sons ou fonemas sonoros são aqueles em que as cordas vocais são tensionadas e vibram de maneira aproximadamente periódica. Os sons surdos são produzidos quando não há vibração das cordas vocais, de modo que o som é formado basicamente nas cavidades do trato vocal, resultando em um sinal com natureza não-periódica ou aleatória (SOUZA JÚNIOR, 2009).

São ilustrados na Figura 2.2 exemplos de sons vozeados e não-vozeados, em que pode-se observar na Figura 2.2(a) a natureza aleatória dos sons não-vozeados e, na Figura 2.2(b), a forma quasi-periódica de um fonema vozeado.



(a) Elocução do fonema /s/: Som não-vozeado. (b) Fonema /z/: Som vozeado

Figura 2.2: exemplo de eventos sonoros vozeados e não-vozeados (SOUZA JÚNIOR, 2009).

Os sons vozeados não devem ser confundidos com os fonemas da língua portuguesa, uma vez que, o som é entendido como uma complexa realidade físico acústica de cada unidade sonora da fala, enquanto que os fonemas correspondem a percepção eclética e interpretativa realizada pelo falante e ouvinte, respectivamente (PAULA, 2000).

O fonema é a menor unidade sonora (fonética) de uma língua que estabelece contraste de significado para diferenciar palavras. Os fonemas, na língua portuguesa, são classificados em vogais, semi-vogais e consoantes.

As vogais são sons produzidos sem obstáculos para a passagem de ar, que passa livremente pela boca, oriundo do pulmão. Sua emissão é independente de outro fonema, por isso constitui a base da sílaba. Os sons das vogais produzem-se a partir de diferentes posicionamentos dos músculos da boca, constituídos pela língua, pelos lábios e pelo palato.

Na produção das vogais, a forma do trato vocal é constante com o tempo e uniforme, com as vibrações sustentadas das cordas vocais. Assim, para as vogais, o trato vocal pode ser, aproximadamente, modelado como uma concatenação de uma série de tubos cilíndricos de área transversal uniforme (KSHIRSAGAR; MAGNENAT-THALMANN, 2000).

Uma aproximação simples do modelo que consiste em m tubos acústicos é ilustrado na Figura 2.3. Os tubos têm as áreas transversais A_1 a A_m. Embora estes valores têm grande variação de pessoa para pessoa, a distribuição é semelhante em relação a uma dada vogal.

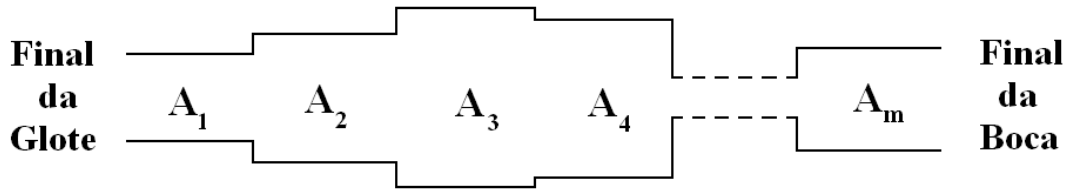


Figura 2.3: representação esquemática do sistema vocal e aproximação por tubos concatenados (KSHIRSAGAR; MAGNENAT-THALMANN, 2000).

As semi-vogais sempre acompanham uma vogal, formando sílaba com ela. Na língua escrita, as semi-vogais são representadas pelo ***i*** e ***u***, podendo em alguns casos serem representadas pelo ***e*** e ***o***.

As consoantes são fonemas produzidos através da obstrução do ar proveniente do pulmão, precisando de uma vogal para ser emitidos. Esses obstáculos podem ser totais ou parciais, a partir da posição da língua e dos lábios.

Com base no conhecimento sobre as características da formação da fala humana, é possível entender os fundamentos das técnicas de extração de características descritas na próxima seção.

2.2 Predição Linear

Predição Linear (LP) é uma operação matemática em que os valores futuros de um sinal de tempo-discreto são estimados como uma função linear de amostras passadas.

Em Processamento Digital de Sinais (PDS), LP é freqüentemente chamada Codificação Linear Preditiva (LPC) e pode então ser vista como um sub-conjunto de teoria de filtros.

Para um sinal discreto s_n, LP consiste em modelar o sinal como a saída de um sistema com entrada desconhecida u_n, representado matematicamente por (MAKHOUL, 1975)

$$s_n = -\sum_{k=1}^{p} a_k \cdot \hat{s}_{n-k} + G \cdot \sum_{l=0}^{q} b_l \cdot u_{n-1}, \qquad b_0 = 1 \tag{2.1}$$

em que a_k, $1 \leqslant k \leqslant p$, b_l, $1 \leqslant l \leqslant q$, e o ganho G são os parâmetros do sistema hipotético. Esta equação mostra que a saída s_n é uma função linear de saídas passadas e entradas presentes e passadas.

A equação 2.1 pode também ser especificada no domínio da freqüência aplicando-se a Transformada Z em ambos os lados dessa equação, obtendo com isso a função de transferência $H(z)$ do sistema (MAKHOUL, 1975)

$$H(z) = \frac{S(z)}{U(z)} = G\frac{1 + \sum_{l=1}^{q} b_l \cdot z^{-l}}{1 + \sum_{k=1}^{p} a_k \cdot z^{-k}}, \tag{2.2}$$

em que $U(z)$ é a Transformada Z da entrada desconhecida u_n e $S(z)$ é a Transformada Z de s_n dada por

$$S(z) = \sum_{n=-\infty}^{\infty} s_n \cdot z^{-n}. \tag{2.3}$$

Quando $b_l = 0$ para $1 \leqslant l \leqslant q$, tem-se um modelo chamado de todo-pólo, também conhecido como Auto-Regressivo (AR) (MAKHOUL, 1975).

Estimação dos parâmetros

O modelo todo-pólo é utilizado nos sinais da fala, em que o sinal de entrada u_n pode ser considerado como desconhecido. Desta forma, o sinal s_n pode ser predito somente através do somatório ponderado linearmente das amostras passadas. A aproximação de s_n por $\tilde{s}_n$ é dada por

$$\tilde{s}_n = -\sum_{k=1}^{p} a_k \cdot \hat{s}_{n-k}. \tag{2.4}$$

A estimação dos parâmetros desse modelo pode ser feita através do Método dos Mínimos Quadráticos (MMQ), em que assumi-se que o erro e_n, entre o valor atual s_n e o valor predito $\tilde{s}_n$, é dado por

$$e_n = s_n - \tilde{s}_n = s_n + \sum_{k=1}^{p} a_k \cdot \hat{s}_{n-k}. \tag{2.5}$$

No MMQ, os parâmetros a_k são obtidos com o resultado da minimização do Erro Quadrático Médio (MSE) ou do erro total em função de cada um dos parâmetros. A análise pode ser realizada através de duas abordagens. A primeira assume s_n como um sinal determinístico e a segunda como um processo aleatório.

Na primeira abordagem, o erro quadrático E é dado por (MAKHOUL, 1975)

$$E = \sum_n {e_n}^2 = \sum_n \left(s_n + \sum_{k=1}^{p} a_k \cdot \hat{s}_{n-k} \right)^2, \tag{2.6}$$

podendo ser minimizado utilizando a expressão

$$\frac{\partial E}{\partial a_i} = 0, \qquad 1 \leqslant i \leqslant p \tag{2.7}$$

A partir das equações 2.6 e 2.7, obtém-se o seguinte conjunto de equações

$$\sum_{k=1}^{p} a_k \sum_n s_{n-k} \cdot s_{n-i} = -\sum_n s_n s_{n-i}. \tag{2.8}$$

O Erro Quadrático Médio (MSE) total, denotado por E_p, é obtido expandindo a equação 2.6 e substituindo na equação 2.8, resultando em

$$E_p = \sum_n {s_n}^2 + \sum_{k=1}^{p} a_k \sum_n s_n \cdot s_{n-k} \tag{2.9}$$

Existem dois métodos para estimar os parâmetros dependendo do intervalo de duração do erro: o método de autocorrelação e método da covariância.

No Método de Autocorrelação, o erro é minimizado para um intervalo infinito $-\infty < n < \infty$. Assim, as equações 2.8 e 2.9 são reduzidas a (MAKHOUL, 1975)

$$\sum_{k=1}^{p} a_k \cdot R(i-k) = -R(i), \qquad 1 \leqslant i \leqslant p \tag{2.10}$$

$$E_p = R(0) + \sum_{k=1}^{p} a_k R(k), \tag{2.11}$$

em que

$$R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+1}, \tag{2.12}$$

é a função de autocorrelação do sinal s_n. Note que $R(i)$ é uma função par, ou seja, $R(-i) = R(i)$. Os parâmetros $R(i-k)$ formam o que é conhecido como matriz

de autocorrelação. Na prática, o sinal s_n é conhecido ou utilizado apenas em um intervalo finito $0 \leqslant n \leqslant N-1$. Desta forma, a função de autocorrelação é dada por

$$R(i) = \sum_{n=0}^{N-1-i} s'_n s'_{n+1}, \qquad i \geqslant 0. \tag{2.13}$$

No Método da Covariância, o erro é minimizado para um intervalo finito $0 \leqslant n \leqslant N-1$. Assim, as equações 2.8 e 2.9 são reduzidas a

$$\sum_{k=1}^{p} a_k \cdot \varphi_{ki} = -\varphi_{0i}, \qquad 1 \leqslant i \leqslant p \tag{2.14}$$

$$E_p = \varphi_{00} + \sum_{k=1}^{p} a_k \varphi_{0k}, \tag{2.15}$$

em que a covariância do sinal s_n no dado intervalo é dada por

$$\varphi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k}. \tag{2.16}$$

Os coeficientes φ_{ki} na equação 2.14 formam a matriz de covariância. A partir da equação 2.16, pode-se mostrar que a matriz de covariância φ_{ik} é simétrica, ou seja, $\varphi_{ki} = \varphi_{ik}$. Com base na equação 2.16, observa-se que os termos ao longo da diagonal principal estão relacionados com o coeficiente anterior através da relação

$$\varphi_{i+1,k+1} = \varphi_{ik} + s_{-i-1} s_{-k-1} - s_{N-1-i} s_{N-1-k}. \tag{2.17}$$

A partir da equação 2.17, verifica-se que para determinar os valores do sinal s_n, para $-p \leqslant n \leqslant N-1$, deve ser conhecido um total de $p+N$ amostras. O método de covariância reduz-se ao método de autocorrelação quando N tende ao infinito.

A segunda abordagem considera o sinal aleatório e o erro e_n na equação 2.5 também é um processo aleatório. Pelo MMQ, minimiza-se o valor esperado $\mathcal{E}$ do quadrado da erro, ou seja

$$E = \mathcal{E}({e_n}^2) = \mathcal{E}\left(s_n + \sum_{k=1}^{p} a_k \cdot \hat{s}_{n-k} \right)^2. \tag{2.18}$$

Aplicar a equação 2.7 na equação 2.18 resulta em

$$\sum_{k=1}^{p} a_k \cdot \mathcal{E}(s_{n-k}s_{n-i}) = \mathcal{E}(s_n s_{n-i}), \qquad 1 \leqslant i \leqslant p. \tag{2.19}$$

O erro médio mínimo é então dado por

$$E_p = \mathcal{E}({s_n}^2) + \sum_{k=1}^{p} a_k \mathcal{E} s_n s_{n-k}. \tag{2.20}$$

A forma de resolução das equações 2.19 e 2.20 depende se o processo s_n é estacionário ou não estacionário. No caso em que s_n é um processo estacionário, tem-se que

$$\mathcal{E}(s_{n-k}s_{n-i}) = R(i-k), \tag{2.21}$$

em que $R(i)$ é a autocorrelação do processo. Com isso, as equações 2.19 e 2.9 reduzem-se a equações idênticas a 2.10 e 2.11, respectivamente. A única diferença é que, neste caso, a autocorrelação é de um processo estacionário em vez de um sinal determinístico. O caso estacionário fornece a mesma solução para os coeficientes a_k que o método de autocorrelação no caso determinístico (MAKHOUL, 1975). Além disso, se o processo é estacionário e ergódico, a autocorrelação pode ser calculada para um tempo médio.

Se s_n é um processo não-estacionário, tem-se que

$$\mathcal{E}(s_{n-k}s_{n-i}) = R(n-k, n-i), \tag{2.22}$$

em que $R(n-k, n-i)$ é a autocorrelação não-estacionária entre os tempos $n-k$ e $n-i$. Assumindo-se que a estimação de parâmetros de interesse ocorre no tempo n=0, as equações 2.19 e 2.9 podem ser reescritas como

$$\sum_{k=1}^{p} a_k R(-k, -i) = R(0, -i), \tag{2.23}$$

$$E'_p = R(0,0) + \sum_{k=1}^{p} a_k R(0,k). \tag{2.24}$$

respectivamente, em que E'_p é o erro médio mínimo do processo não-estacionário.

Na estimativa dos coeficientes de autocorrelação não-estacionário do sinal s_n, nota-se que os processos não-estacionários são não ergódicos e, portanto, não se pode substituir a média do conjunto por um tempo médio. No entanto, para uma determinada classe de processos não-estacionários, conhecidos como processos localmente estacionários, é razoável estimar a função de autocorrelação com relação a um ponto no tempo como um tempo médio de curto prazo. Exemplos de processos não estacionários, que podem ser considerados localmente estacionários, são de sinais da fala e de Eletroencefalograma (EEG) (MAKHOUL, 1975).

De maneira análoga ao caso estacionário, estima-se $R(-k,-i)$ através de φ_{ik} utilizando a equação 2.16. Usar esta aproximação para a autocorrelação de um processo não-estacionário conduz a uma solução para os parâmetros a_k na equação 2.23 que é idêntico ao que é dado pela equação 2.14 no método de covariância para o caso determinístico. Observa-se que, para um sinal estacionário, $R(t,t') = R(t-t')$ e, portanto, as equações 2.23 e 2.24 são reduzidas e resultam nas equações 2.10 e 2.11.

Cálculo dos Parâmetros Preditivos

Para cada uma das duas abordagens de predição linear apresentada anteriormente, os coeficientes preditivos a_k, $1 \leqslant k \leqslant p$, podem ser calculados resolvendo um conjunto de p equações em que p é desconhecido. Essas equações são mostradas na equação 2.10 para o método de autocorrelação (estacionário) e na equação 2.14 para o método da covariância (não-estacionário).

Existem diversos métodos para otimizar os cálculos necessários para solução dessas equaçõe, por exemplo, o método de redução ou eliminação de Gauss e o método de redução de Crout (MAKHOUL, 1975). Estes métodos gerais requerem $p^3/3 + O(p^2)$ operações (multiplicações ou divisões) e p^2 locais de armazenamento. Entretanto, evidencia-se através de 2.10 e 2.14 que a matriz de coeficientes em cada caso é uma matriz de covariância. Matrizes de covariância são simétricas e, em geral, semi-definidas positiva, embora na prática são, geralmente, definidas positiva. Portanto, as equações 2.10 e 2.14 podem ser solucionadas mais eficientemente através do método de decomposição de Cholesky. Este método requer $p^3/6 + O(p^2)$ cálculos e $p^2/2$ armazenamento, ou seja, representa metade do esforço computacional dos métodos gerais (MAKHOUL, 1975).

Maior redução no armazenamento e tempo de cálculo é possível resolvendo a equação 2.10 devido a sua forma especial. Esta equação pode ser expandida na forma matricial

$$\begin{bmatrix} R_0 & R_1 & R_2 & \dots & R_{p-1} \\ R_1 & R_0 & R_1 & \dots & R_{p-2} \\ R_2 & R_1 & R_0 & \dots & R_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & R_{p-3} & \dots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ \vdots \\ R_p \end{bmatrix}. \tag{2.25}$$

Através desta equação é possível observar que a matriz de autocorrelação $p \times p$ é simétrica e os elementos ao longo de qualquer diagonal são idênticos.

Levinson (1946) propôs um procedimento recursivo para solucionar esse tipo de equação. O procedimento foi mais tarde reformulado por Robinson (1967). Este método supõe que o vetor coluna no lado direito da equação 2.25 é um vetor coluna genértico. Um outro método, proposto por Durbin (1960), considera que este vetor coluna compreende os mesmos elementos encontrados na matriz de autocorrelação, possuindo um desempenho computacional duas vezes maior que o método de Levinson (1946). O método requer apenas $2p$ locais de armazenamento e $p^2 + O(p)$ operações: uma grande redução de complexidade em relação aos métodos gerais.

O procedimento recursivo de Durbin pode ser especificado através das seguintes equações (MAKHOUL, 1975):

$$E_0 = R(0), \tag{2.26}$$

$$k_i = -\frac{R(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)}{E_{i-1}}, \tag{2.27}$$

$$a_i^{(i)} = k_i, \tag{2.28}$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i \cdot a_{i-j}^{(i-1)}, \qquad 1 \leqslant j \leqslant i-1, \tag{2.29}$$

$$E_i = (1 - k_i^2) E_{i-1}. \tag{2.30}$$

Estas equações 2.27 a 2.30 são solucionadas recursivamente para $i = 1, 2, \dots, p$, cuja

solução final é dada por

$$a_j = a_j^{(p)}. \tag{2.31}$$

O cálculo dos coeficientes de autocorrelação ou de covariância requer $p \cdot N$ operações, que podem influenciar o tempo da computação se $N >> p$, como é freqüentemente o caso.

A solução da equação 2.25 não é afetada se todos os coeficientes de autocorrelação são multiplicados por uma constante. Em particular, se todos $R(i)$ são normalizados, isto é divididos por $R(0)$, são formados os coeficientes de autocorrelação normalizados $r(i)$

$$r(i) = \frac{R(i)}{R(0)}. \tag{2.32}$$

em que $|r(i)| \leqslant 1$.

Um sub-produto do algoritmo recursivo de Durbin (1960) é o cálculo do erro mínimo total E_i em cada etapa. Pode-se facilmente mostrar que o erro mínimo E_i diminui à medida que a ordem da predição aumenta. Neste caso E_i não é negativo, logo, é um erro quadrático. Portanto, tem-se que (MAKHOUL, 1975)

$$1 \leqslant E_i \leqslant E_{i-1}, \qquad E_0 = R(0). \tag{2.33}$$

Se os coeficientes de autocorrelação são normalizados utilizando a equação 2.32, então o erro mínimo E_i é também dividido por $R(0)$, sendo conhecido como erro normalizado V_i, dado por

$$V_i = \frac{E_i}{R(0)} = 1 + \sum_{k=1}^{i} a_k r(k). \tag{2.34}$$

Considerando-se a relação descrita na equação 2.33 é possível observar que

$$1 \leqslant V_i \leqslant p, \qquad i \geqslant 0. \tag{2.35}$$

Também, utilizando as equações 2.30 e 2.34, o erro normalizado final V_p é dado por

$$V_p = \prod_{i=1}^{p} 1 - {k_i}^2. \tag{2.36}$$

As quantidades intermediárias k_i, $1 \leqslant i \leqslant p$, são conhecidas como coeficientes de reflexão, também denominados de coeficientes de correlação parcial. O coeficiente k_i pode ser interpretado como a correlação parcial (negativa) entre s_n e s_{n+1}, mantendo s_{n+1}, $\cdots$, s_{n+i-1} fixos. O uso do termo "coeficiente de reflexão" vem da teoria de linha de transmissão, em que k_i pode ser considerado como coeficientes de reflexão no limite entre duas seções com impedâncias $Z_i + Z_{i+1}$, sendo dado por (MAKHOUL, 1975)

$$k_i = \frac{Z_{i+1} - Z_i}{Z_{i+1} + Z_i}. \tag{2.37}$$

Assim, a função de transferência $H(z)$ pode então ser considerada como uma seqüência de seções com taxas de impedância, usando a equação 2.37, sendo dada por

$$\frac{Z_{i+1}}{Z_i} = \frac{1 + k_i}{1 - k_i}, \qquad 1 \leqslant k \leqslant p. \tag{2.38}$$

A mesma explicação pode ser dada para qualquer tipo de situação em que existe transmissão de onda plana com incidência normal, em um meio consistindo de uma seqüência de seções com diferentes impedâncias. No caso de um tubo acústico com p seções de igual espessura, a taxa de impedância reduz ao inverso da taxa de áreas de seção-transversal consecutivas. Este fato pode ser usado na análise da fala.

Além da Predição Linear, outra ferramenta matemática importante na extração de características da fala é a Transformada *Wavelet*.

2.3 Transformada *Wavelet*

A análise por Transformada Fourier assume que o sinal é estacionário no tempo; contudo, no caso de sinais de fala isto não é verdade. Então, para sinais de fala, uma versão janelada do sinal é usada para processamento, com a suposição de estacionaridade durante esta janela. Este método é conhecido como Transformada de Fourier de Tempo Curto (STFT). Se uma janela de curta duração for escolhida, a freqüência de resolução é pequena e ao aumentar a duração da janela ocorre o

contrário. Fixando o tamanho da janela, a resolução tempo-freqüência conseguida pela STFT é também fixada (FAROOQ; DATTA, 2003).

Para superar o problema da resolução fixa, a Transformada *Wavelet* utiliza uma janela de tamanho adaptativo, que utiliza mais tempo para baixas freqüências e menos tempo para altas freqüências (RIOUL; VETTERLI, 1991). Essa análise pode ser usada para um sinal que tem componentes de alta-freqüência de curta duração e componentes de baixa-freqüência de longa duração, como no caso da fala (FAROOQ; DATTA, 2003).

A função base usada na Transformada *Wavelet* é localizada tanto no tempo como na freqüência. Todas as funções *wavelet* são versões escalonadas de uma função protótipo $\psi(t)$, também conhecida como *wavelet* 'mãe', de média zero e centrada na vizinhança de $t = 0$, dada por (FAROOQ; DATTA, 2003)

$$\psi_{\tau,a}(t) = a^{-\frac{1}{2}} \cdot \psi\left(\frac{t-\tau}{a}\right), \tag{2.39}$$

em que os parâmetros τ e a são chamados parâmetros de translação e escalamento, respectivamente. O termo $a^{-1/2}$ é usado para normalização da energia. A Transformada *Wavelet* Contínua (CWT) de um sinal $x(t)$, em que $x \in \mathbf{L^2}$, é dada por (RIOUL; VETTERLI, 1991)

$$CWT(\tau, a) = a^{-\frac{1}{2}} \int x(t) \cdot \psi^*\left(\frac{t-\tau}{a}\right) dt, \tag{2.40}$$

em que o parâmetro de escalamento a fornece a largura da *wavelet*, τ indica a posição e $\psi^*(t)$ é o complexo conjugado de $\psi(t)$. Tipicamente, a CWT é sobre-completa e uma amostragem apropriada pode ser usada para eliminar redundâncias.

A Transformada *Wavelet* Discreta (DWT) pode ser obtida a partir da equação (FAROOQ; DATTA, 2003)

$$D(j,k) = 2^{-\frac{j}{2}} \sum_i x(i) \cdot \psi^*\left(2^{-j}i - k\right), \tag{2.41}$$

em que i, j e k são inteiros. Escolhendo o fator de escalamento como diádico (2^j), a transformada resultante é conhecida como DWT diádica.

Se a decomposição *wavelet* for computada para uma escala 2^j, a representação do sinal resultante não é completa. As componentes de baixa freqüência

correspondentes as escalas maiores que 2^j são avaliadas usando (FAROOQ; DATTA, 2003)

$$A(j,k) = 2^{-\frac{j}{2}} \sum_i x(i) \cdot \phi^* \left(2^{-j} i - k\right), \tag{2.42}$$

em que $\phi^*(i)$ é o complexo conjugado da função de escalamento $\phi(i)$.

A DWT pode ser vista como um processo de filtragem do sinal, usando um filtro passa-baixas (escalamento) e um filtro passa-altas (*wavelet*). Então, o primeiro nível de decomposição DWT de um sinal divide em duas faixas, uma versão passa-baixas e uma versão passa- altas do sinal. A versão passa-baixas fornece a representação aproximada do sinal, enquanto a passa-altas indica os detalhes ou variações de altas freqüências. O segundo nível de decomposição é executado sobre a versão passa-baixas do primeiro nível de decomposição, como mostrado na Figura 2.4. Então, a decomposição *wavelet* resulta em uma árvore cuja estrutura é dita recursiva (FAROOQ; DATTA, 2003).

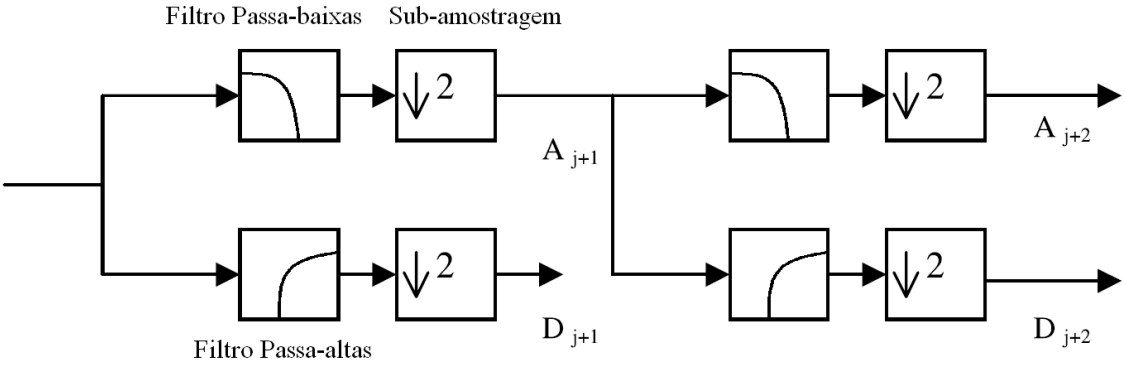


Figura 2.4: decomposição de sinal usando DWT diádica (FAROOQ; DATTA, 2003).

2.3.1 *Wavelet* Daubechies

As *wavelets Daubechies* são um família de *wavelets* ortogonais definidas para a DWT e caracterizadas por um número máximo de momentos nulos para um dado suporte. Para cada tipo de *wavelet* desta classe, existe uma função de escalamento (também chamada de *wavelet* 'pai') que gera uma análise multi-resolução ortogonal (DAUBECHIES, 1992).

Em geral, as *wavelets Daubechies* são escolhidas por ter um maior número N de momentos nulos para $\psi(t)$ e por dar largura de suporte $2N-1$, em que N é a ordem

da wavelet. Entre as $2N-1$ possíveis soluções, é escolhida a solução cujo filtro de escalamento tenha fase extrema (DAUBECHIES, 1992).

2.3.2 *Wavelet* Coiflet

As *wavelets* Coiflet são funções que, embora sejam similares em alguns momentos estatísticos, diferem-se por apresentarem variação de momentos não somente na função *wavelet* $\Psi(t)$, mas também na função de escalamento $\Phi(t)$, o que a destaca na exibição de características interpoladas, oferecendo boa aproximação para funções polinomiais (PEREIRA *et al.*, 2008).

As Coiflets são *wavelets* discretas projetadas por Ingrid Daubechies, a pedido de Ronald Coifman, para ter funções de escalamento com momentos nulos. A *wavelet* é aproximadamente simétrica e as suas funções wavelet tem $2N$ momentos nulos e as funções de escalamento $2N-1$, com largura de suporte $6N-1$ (DAUBECHIES, 1992).

Tanto a função de escalamento (filtro passa-baixas) e a função *wavelet* (filtro passa-altas) devem ser normalizadas por um fator $1/\sqrt{2}$. Os coeficientes *wavelet* B_k são derivados invertendo-se a ordem dos coeficientes da função escala e, em seguida, invertendo o sinal de cada termo, através da equação (DAUBECHIES, 1992)

$$B_k = (-1)^k \cdot C_{(N-1-k)}, \tag{2.43}$$

em que k é o índice do coeficiente, B é um coeficiente *wavelet*, C um coeficiente da função de escalamento e N é a ordem da *wavelet*.

2.4 Extração de características

Para identificar o fonema, alguma de suas características no tempo/frequência ou em algum outro domínio devem ser conhecidos. Assim, uma característica pode ser definida como uma unidade mínima, que distingue fonemas maximamente próximos (FAROOQ; DATTA, 2003).

Através da extração de características, o espaço de dados é transformado num espaço de características que possui a mesma dimensão do espaço de dados original, porém é representado por um número reduzido de características efetivas (COSTA, 2006). Nesta seção, são descritas as técnicas de extração de características estudadas neste trabalho.

2.4.1 Extração utilizando Predição Linear

Nesta seção é descrito o método proposto por Kshirsagar e Magnenat-Thalmann (2000). Este método consiste em extrair características relacionadas aos fonemas do tipo vogal.

Esse método realiza a extração da informação da forma do trato vocal, utilizando a análise por LP proposta por Wakita (1973). O método compara um modelo do filtro acústico, representado na Figura 2.3 pelos tubos conectados, com o modelo de produção da fala.

A comparação entre o modelo de tubo acústico e o modelo derivado da LP conduz a seguinte conclusão. Os coeficientes de reflexão k_i, calculados como um subproduto do algoritmo recursivo de LP, estão diretamente relacionados com a variação da área do trato vocal, conforme o modelo do tubo concatenadas, através da equação (KSHIRSAGAR; MAGNENAT-THALMANN, 2000)

$$k_i = \frac{A_{i-1} - A_i}{A_{i-1} + A_i}. \tag{2.44}$$

Estes coeficientes de reflexão são usados como características para classificação. As características extraídas fornecem a informação da forma do trato vocal para vogais sustentadas, através da equação 2.44.

2.4.2 Extração utilizando Transformada Wavelet

O método proposto por Farooq e Datta (2003) consiste em aplicar a DWT sobre um frame de 32 milisegundos (ms) de duração e extrai algumas características deste frame para classificação dos fonemas. A *wavelet*-base escolhida é a *Daubechies* com 6 momentos nulos por ter um tamanho mínimo de suporte de $2p - 1$ para um dado número de momentos nulos p.

Nesse método, o *frame* de 32 ms é dividido em 4 sub-*frames* de duração de 8 ms para poder acomodar rápidas oscilações nos fonemas e permitir acompanhar a evolução temporal da energia média por amostra em cada faixa. Como o sinal de fala é estacionário para duração de aproximadamente 10 ms (devido à limitação física do movimento das articulações da produção da fala), qualquer redução adicional na duração do frame não é útil (FAROOQ; DATTA, 2003).

A decomposição wavelet é aplicada em cada sub-frame e a energia dos coeficientes

wavelet em cada faixa de freqüência é calculada. Esta energia é normalizada pelo número de amostras na faixa correspondente, resultando desse modo uma energia média por amostra em cada faixa. A normalização é essencial porque cada faixa terá um número diferente de amostras. Estas energias médias por amostra, para diferentes faixas, são usadas como características para classificação. As características extraídas em um sub-frame base não fornecem somente a energia em cada faixa, mas fornecem também uma idéia da variação temporal da energia em cada faixa.

A quantidade de características extraídas depende do nível de decomposição de cada sub-frame. Uma decomposição nível N de um sub-frame extrai $N + 1$ características. Desta forma, sobre um fonema de duração 32 ms extrai-se um total de $4N+4$ características. Segundo Farooq e Datta (2003), a faixa de baixa freqüência de $0 - 62.5Hz$ é a menor faixa em que existe informação discriminatória. Assim, o maior nível de decomposição utilizado para extrair as características depende da taxa de amostragem do áudio.

Uma vez extraídas as características, o problema consiste em obter uma função discriminante para separar as diferentes classes presentes no espaço de características.

2.5 Reconhecimento de Padrões

O Reconhecimento de Padrões (RP) trata da classificação de uma estrutura de dados através de um conjunto de propriedades ou características. O RP através de máquinas envolve técnicas para a atribuição dos padrões a suas respectivas classes, de forma automática ou com a menor intervenção humana possível. Um padrão é uma descrição de um objeto e a classe de padrões é uma família de objetos que compartilham uma mesma propriedade (GONZALEZ; WOODS, 2008).

Exemplos de aplicações de RP são o reconhecimento de voz, impressão digital, identificação de caracteres, estrutura de íris, reconhecimento de palavras e escrita cursiva, reconhecimento de formas, supervisão de processos, detecção de falha em máquinas e diagnósticos médicos.

A classificação de padrões pode ser definida como um problema relacionado com a determinação de uma fronteira de decisão que consegue distinguir diferentes padrões em classes dentro de um espaço de características d-dimensional, em que d é o número

de características. A classificação, então, pode ser realizada, e pode ser entendida, de maneira geral, pela partição do espaço de atributos em um número finito de regiões de tal forma que objetos de uma mesma classe recaiam, pelo menos em tese, sempre dentro de uma mesma região.

A seguir são descritos os fundamentos do classificador para RP utilizado neste trabalho.

2.5.1 Redes Neurais Artificiais

Uma Rede Neural Artificial (RNA) pode ser vista como um modelo matemático composto de muitos elementos computacionais não lineares, chamados de neurônios, operando em paralelo e totalmente conectados por ligações caracterizadas por diferentes pesos (ARAÚJO, 2004).

Um simples neurônio V_c, calcula a soma das entradas (E_1, E_2, ..., E_n) ponderadas pelos pesos (W_{c1}, W_{c2}, ..., W_{cn}) que cada conexão possui, direciona este resultado para uma função de ativação não linear $\varphi(.)$ para produzir uma saída simples Y_c denominada nível de ativação daquele neurônio. Um modelo não linear de um neurônio pode ser observado na Figura 2.5.

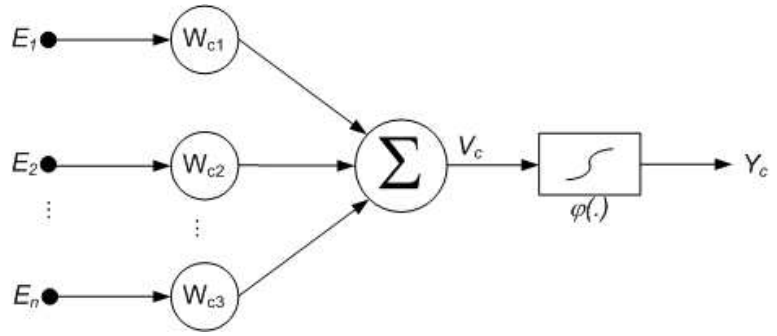


Figura 2.5: modelo não linear de um neurônio (HAYKIN, 2001).

Modelos de redes neurais são especificados pela topologia da rede, características dos neurônios e regras de aprendizagem ou treinamento. O termo topologia refere-se à estrutura da rede como um todo, especificando como as entradas, as saídas e as camadas escondidas são interconectadas (HAYKIN, 2001). Neste trabalho, é utilizada a topologia do Perceptron Multi-Camadas (MLP).

O MLP é uma das RNAs mais utilizadas para separar dados não-linearmente separáveis. Para iniciar o processo de aprendizagem da rede neural, faz-se necessária a seleção de um conjunto de amostras das classes padrões (conjunto de treinamento)

a serem reconhecidos pela rede e suas saídas desejadas correspondentes. Deve-se selecionar para treinamento amostras representativas de cada classe e um número suficiente destas amostras para que a rede possa aprender a identificar os padrões.

Basicamente, rede do tipo MLP apresenta três ou mais camadas de neurônios, a saber, um conjunto de unidades sensoriais (nós de fonte) que constituem a camada de entrada, uma ou mais camadas ocultas e uma camada de saída. A sua topologia é completamente interconectada na direção da camada de entrada para a saída sem retroalimentação. O sinal de entrada se propaga através da rede, camada por camada até a saída (HAYKIN, 2001).

A definição do número de neurônios das camadas de entrada e saída é realizada de acordo com o problema em questão. O número de neurônios da camada intermediária, ou mesmo o número de camadas intermediárias, é definido de forma intuitiva, não havendo portanto, uma regra que defina o seu número. Se a quantidade de neurônios escolhidos for muito pequena, isto, pode fazer com que apenas alguns neurônios especializem-se em características não úteis, tais como ruído. Se o número de neurônios for insuficiente, pode acontecer da rede não conseguir aprender os padrões desejados (HAYKIN, 2001).

O neurônio individual é o bloco construtivo de cada camada, que é caracterizado principalmente por sua função da ativação. A função de ativação mais comumente utilizada é a função logística, também utilizada neste trabalho, e é definida como (HAYKIN, 2001)

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2.45}$$

As redes MLPs são projetadas para aproximar uma relação entre entrada e saída não conhecida, através dos pesos de cada conexão, via regras de aprendizagem. Uma característica de grande importância deste modelo é o aprendizado supervisionado baseado em duas etapas: propagação e adaptação.

A propagação ocorre na fase de treinamento da rede e consiste em fornecer à rede um conjunto de estímulos (padrões de entradas) e a saída desejada correspondente ao padrão de entrada apresentado. Nesta fase, o primeiro padrão de entrada é propagado até a saída. Durante este passo os pesos sinápticos não mudam de valor.

Na fase de adaptação, o sinal do erro é computado (resultado da diferença entre a saída desejada e saída real da rede) e transmitido de volta para cada neurônio da camada intermediária que contribuiu para a saída obtida. Sendo assim,

cada neurônio da camada intermediária recebe somente uma parte do erro total, conforme a contribuição relativa que o neurônio obtém na saída gerada. Este processo repete-se camada por camada, até que cada neurônio da rede receba o seu peso correspondente. Tal processo é conhecido como retropropagação do erro, pois, o aprendizado baseia-se na propagação retroativa do erro, contra a direção das conexões sinápticas da rede (HAYKIN, 2001).

Os pesos existentes nas conexões entre os neurônios são atualizados de acordo com o erro recebido pelo neurônio associado. Esta atualização é um processo iterativo em que a rede ajusta seus pesos até que a informação do ambiente seja aprendida. O processo de aprendizagem termina quando a saída obtida pela rede neural, para cada um dos padrões de entrada, for próxima o bastante da saída desejada, de forma que a diferença entre ambas seja aceitável. Esta diferença é obtida pelo cálculo do MSE.

2.6 Resumo do Capítulo

Neste capítulo foram brevemente descritas as etapas de um sistema básico de reconhecimento de voz. Foram apresentadas a fundamentação matemática e as técnicas de extração de características propostas por Kshirsagar e Magnenat-Thalmann (2000) e por Farooq e Datta (2003) que são avaliadas neste trabalho, utilizando como classificador as Redes Neurais Artificiais (RNAs). No capítulo seguinte são apresentados a técnica proposta para extração de características, o conjunto de dados e metodologia utilizada nos testes de avaliação das técnicas estudadas.

Capítulo 3

Metodologia

Neste capítulo é descrita a metodologia utilizada na comparação das técnicas de extração de características para reconhecimento de fonemas apresentadas no Capítulo anterior. Na seção 3.1 é descrito o conjunto de dados utilizado para os testes dos algoritmos estudados. Em seguida, na seção 3.2 são descritos a metodologia e os parâmetros utilizados nos testes. Na seção 3.3 é brevemente apresentada a ferramenta de simulação do Matlab, o Simulink. Por fim, na última seção deste Capítulo é descrita a modificação proposta na técnica baseada em Transformada *Wavelet*.

3.1 Conjunto de Dados

Para comparar as técnicas de extração de características, utiliza-se um conjunto de dados formado pelas amostras de áudio dos fonemas do tipo vogal. Essas amostras são capturadas de 13 pessoas, sendo uma do sexo feminino, pronunciando cada vogal durante aproximadamente 3 segundos de forma constante e sem pausas, variando-se apenas a distância para o microfone. Para aquisição das amostras é utilizado um microfone para computadores pessoais da marca Leadership. As vogais pronunciadas consistem nas vogais da língua portuguesa a, ê, é, i, ô, ó, u. No total, o conjunto de dados possui 7 classes de vogais.

As amostras de áudio são extraídas a uma taxa de amostragem de 8kHz e 8 *bits* na quantização, ou seja, a configuração mínima de digitalização de um sinal de voz. Utiliza-se a configuração mínima para permitir uma melhor comparação entre as técnicas de extração de características, explorando ao máximo o potencial de cada

uma destas técnicas. Todas as amostras foram gravadas em uma sala fechada com ruído proveniente de condicionadores de ar e de forma espontânea.

3.2 Metodologia de Simulação

Para avaliar as técnicas de extração de características estudadas são realizadas simulações em um computador pessoal da marca Dell com processador Intel Pentium D de 2,80 GHz e 2 GB de memória RAM com sistema operacional Windows XP. Todas as simulações foram realizadas utilizando-se a plataforma de simulação Simulink do Maltab versão 2006b.

O intuito dessa avaliação é realizar um estudo estatístico de algoritmos de reconhecimento de voz, e a partir da análise desses resultados, determinar quais as técnicas possuem o menor custo computacional e os melhores resultados de classificação.

O sinal de voz é não-estacionário e ruidoso, de modo que a analogia com filtros digitais estabelecida na seção 2.2 somente é válida para um período de tempo aproximadamente estacionário da fala, que geralmente é em torno de 10 a 30 ms. Dessa forma, todos os métodos de extração de características utilizados neste trabalho, utilizam a análise de curta duração.

Para esse fim, implementa-se o janelamento do sinal utilizando uma janela retangular que é movida ao longo do sinal de voz sem sobreposição entre frames adjacentes. O tamanho dessa janela nos testes é de 32ms, ou seja, 256 amostras de áudio por frame para um taxa de amostragem de 8kHz.

Quanto à etapa de extração de características, varia-se a quantidade de características extraídas por frame de áudio. No método baseado em Predição Linear varia-se a quantidade de coeficientes de reflexão de 8 a 28. Para os métodos baseados em Transformada *Wavelet* utiliza-se o nível da decomposição *wavelet* para cada sub-frame de 1 a 6.

O classificador utilizado para avaliar as técnicas de extração é a rede neural Perceptron Multi-Camadas (MLP) com 2 camadas, 1 camada de entrada com quantidade de neurônios igual ao número de características extraídas e 1 camada de saída com 7 neurônios referentes às classes das vogais. Os parâmetros de treinamento da rede neural são escolhidos para se obter um estudo mais preciso e são: 100 épocas de treinamento, Erro Quadrático Médio (MSE) desejado de 10^{-5} e passo de

apredizagem de $0,01$.

Para avaliar essas técnicas no classificador, o conjunto de dados é dividido em dois conjunto: um de treinamento e outro de teste. As amostras são embaralhadas aleatoriamente e 80% delas são atribuías ao conjunto de treinamento, enquanto os 20% restantes são utilizados no teste.

A avaliação das técnicas é feita com base nas taxas de acerto média, máxima, mínima e desvio-padrão, e na matriz de confusão. Além disso, o tempo total médio de treinamento e de reconhecimento são determinados. Os resultados são extraídos de 10 simulações independentes.

3.3 Simulink

O Simulink é um sistema de simulação computacional baseado em ícones capaz de simular sistemas em geral, possuindo ou não características não lineares e dinâmicas. Através da combinação de vários diagramas de blocos o Simulink é capaz de simular sistemas de pequeno e grande porte e também de integrar-se com funções desenvolvida em Matlab (TAYLOR, 2004).

A estrutura modular do Simulink permite o agrupamento de modelos dentro de hierarquias que provê uma visão geral do sistema e uma fácil manutenção de componentes e sistemas complexos. O Simulink utiliza um ambiente gráfico baseado em diagramas de blocos que suportam diferentes operações, como, por exemplo, funções aritméticas, entrada e saída de dados, funções de transferência, modelos de estado de espaços, dentre outras (KALAGASIDIS *et al.*, 2007).

A partir de pequenos blocos, através deste ambiente, um conjunto de bibliotecas de várias áreas pode ser desenvolvido. Cada biblioteca é formada por vários sub-blocos ou sub-modelos que representam uma função específica de uma área específica. A partir do arranjo de vários destes blocos na tela e através das conexões dos blocos com algumas variáveis e constantes é possível simular vários sistemas de equação (TAYLOR, 2004).

Com o uso do Simulink, um desenvolvedor se preocupa mais com a implementação do modelo físico do projeto, esquecendo das questões de discretização e integração entre blocos, possibilitando assim um menor tempo para simular a eficiência de um algoritmo para em seguida ser implementado em um sistema específico (KALAGASIDIS *et al.*, 2007).

3.4 Técnica Proposta para extração de características

Este trabalho propõe uma modificação na técnica para extração de características proposta por Farooq e Datta (2003) para melhorar a decomposição do sinal em faixas de frequência.

A modificação consiste em utilizar como base a *wavelet* simétrica Coiflet de ordem 3, com 6 momentos nulos para a decomposição do sinal. Essa wavelet base é escolhida por permitir a detecção de características interpoladas no sinal de voz (PEREIRA *et al.*, 2008). Além disso, a Coiflet de ordem 3 é comparada a *Daubechies* de ordem 6 por possuirem a mesma quantidade de momentos nulos para a função *wavelet* $\Psi(t)$ (DAUBECHIES, 1992).

3.5 Resumo do Capítulo

Neste capítulo, foram brevemente descritos o conjunto de dados utilizado neste trabalho e a metodologia utilizada nos testes para avaliar as técnicas de extração de características. Também foi apresentado a técnica proposta para extração de características, que consiste em modificação no método proposto por Farooq e Datta (2003), utilizando a wavelet-base Coiflet com 6 momentos nulos para a decomposição do sinal. No capítulo seguinte são apresentados os resultados dos testes de simulação.

Capítulo 4

Resultados

Neste capítulo são mostrados os resultados experimentais obtidos utilizando a metodologia de simulação apresentada no Capítulo 3 para as diferentes técnicas de extração de características abordadas neste trabalho. Na seção 4.1 são apresentados os resultados de desempenho obtidos no reconhecimento dos fonemas. E por fim, são apresentados os resultados de custo computacional das técnicas estudadas.

4.1 Desempenho de Reconhecimento

Na Tabela 4.1 são mostrados as estatísticas da taxas de acerto e o desvio padrão da taxa de acerto das técnicas estudadas neste trabalho. Na tabela em questão, utilizam-se as notações LP-X, *Daubechies*-Y e *Coiflet*-Y, em que X denota o número de coeficientes de reflexão de Predição Linear (LP) e Y o número de níveis da decomposição *wavelet*. Para a técnica baseada em LP são X extraídas e para as técnicas baseadas na Transformada *Wavelet* são extraídas $4Y + 4$ características.

Através desta tabela, observa-se que a taxa média de acerto dos métodos estudados varia entre 52,58% e 89,29%. Verifica-se que a técnica com melhor desempenho classificatório é a proposta por Kshirsagar e Magnenat-Thalmann (2000) baseada em LP com resultados superiores ao das técnicas baseadas em decomposição *wavelet*, conforme representado na Figura 4.1. Contudo, a técnica baseada em LP possui uma grande variação dependendo do treinamento, isto é, desvio padrão superior à 5%, enquanto que os outros métodos estudados possuem no máximo este valor de desvio.

Tabela 4.1: desempenho de reconhecimento das técnicas estudadas.

Configuração	**Número de Características**	Taxas de Reconhecimento(%)			
		mínima	*média*	*máxima*	*desvio padrão*
LP-8	8	63,19	78,24	88,46	8,60
LP-12	12	72,53	83,57	92,86	6,40
LP-16	**16**	**68,13**	**89,29**	**95,60**	**8,25**
LP-20	20	78,57	89,23	95,05	5,92
LP-24	24	66,48	87,14	94,51	9,22
LP-28	28	60,99	85,88	98,35	12,29
Daubechies-1	8	48,9	52,58	59,34	3,46
Daubechies-2	12	60,44	65,55	71,98	3,00
Daubechies-3	16	61,54	67,8	72,53	3,46
Daubechies-4	20	60,99	65,22	70,33	3,46
Daubechies-5	24	58,79	65,38	71,43	4,69
Daubechies-6	**28**	**65,93**	**68,90**	**72,53**	**2,65**
Coflet-1	8	56,1	60,05	65,99	3,00
Coflet-2	12	58,85	66,04	74,78	4,90
Coflet-3	16	61,59	70,27	74,23	3,61
Coflet-4	20	64,89	70,33	79,73	5,10
Coflet-5	24	61,59	69,95	77,53	4,58
Coflet-6	**28**	**68,19**	**74,01**	**80,27**	**4,12**

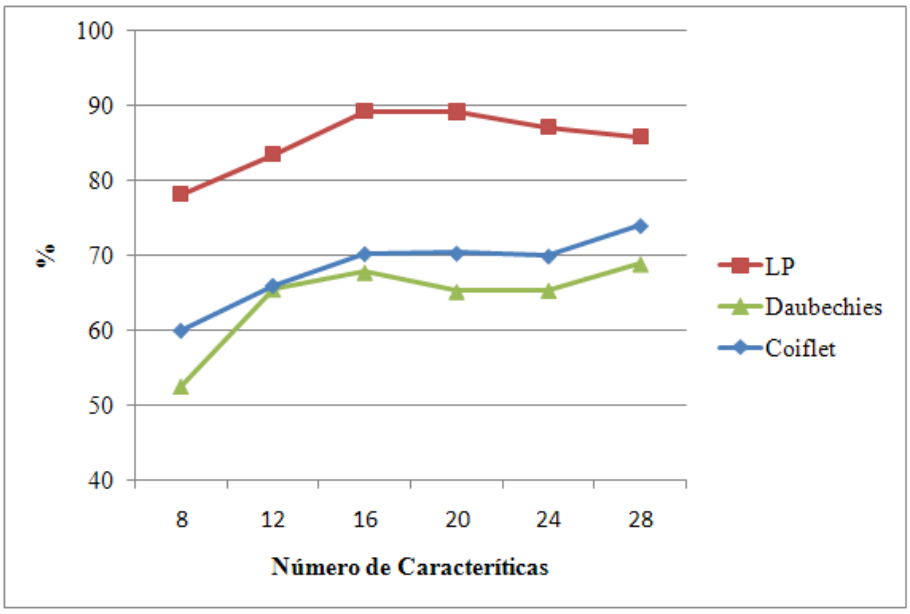


Figura 4.1: taxa média de acerto das técnicas estudadas.

O melhor desempenho classificatório médio da técnica de extração baseada em LP é atingido na configuração LP-16, que obteve uma taxa média de 89,29%. Através da Figura 4.1, observa-se que para valores de coeficientes de reflexão maiores que 16 ocorre uma diminuição na taxa média de reconhecimento.

Na Tabela 4.2 é mostrada a matriz de confusão para configuração com melhor taxa média, LP-16, em que observa-se que pelo menos 79% e no máximo 98% das vogais são classificadas corretamente, conforme mostrado na Figura 4.2. A vogal ***ó*** possui o pior reconhecimento com 79,8% de corretos reconhecimentos. A vogal ***u*** possui um desempenho intermediário com 84,7% das amostras classificadas

Tabela 4.2: matriz de confusão da configuração LP-16.

		Vogal Esperada						
		a	ê	é	i	ô	ó	u
Vogal Reconhecida	a	**252**	2	3	0	0	19	0
	ê	1	**243**	4	20	0	4	0
	é	4	3	**246**	0	1	1	0
	i	1	18	3	**210**	2	0	0
	ô	0	1	2	4	**227**	23	42
	ó	0	0	1	0	7	**209**	1
	u	0	0	0	1	21	6	**238**

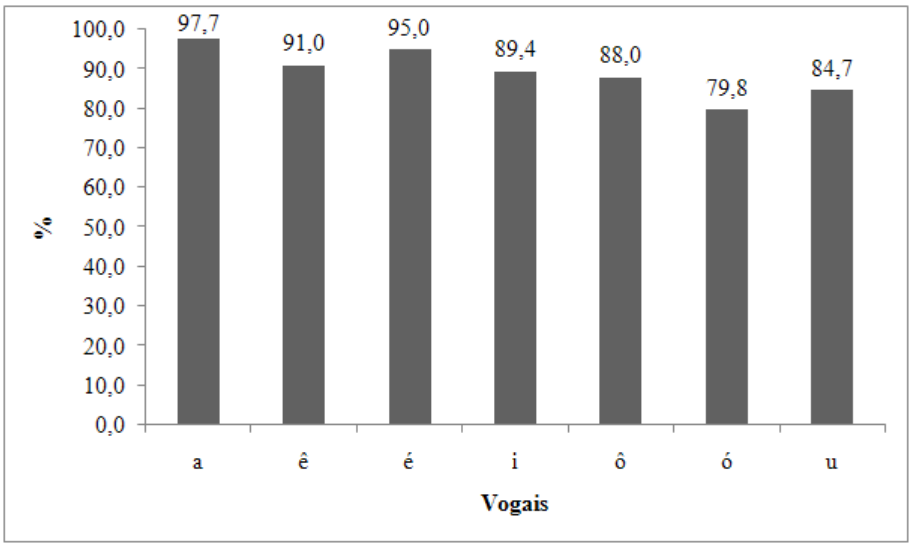


Figura 4.2: porcentagem de reconhecimento por vogal na configuração LP-16.

corretamente. As demais vogais conseguem reconhecer corretamente mais que 88% das amostras de teste.

Para a técnica de extração baseada em Transformada *Wavelet* proposta por Farooq e Datta (2003), o melhor desempenho classificatório médio é alcançado na configuração Daubechies-6, que obteve uma taxa média de 68,90%. Através da Figura 4.1, observa-se que a taxa de acerto médio aumenta com número de níveis de decomposição *wavelet* por sub-*frame*, ou seja, com o número de características extraídas por *frame*.

Na Tabela 4.3 é mostrado a matriz de confusão para configuração com melhor taxa média, *Daubechies*-6, em que observa-se que pelo menos 59% e no máximo 82%

Tabela 4.3: matriz de confusão da configuração Daubechies-6.

		Vogal Esperada						
		a	ê	é	i	ô	ó	u
Vogal Reconhecida	a	**231**	10	6	1	6	20	3
	ê	1	**158**	16	41	5	8	5
	é	5	22	**204**	9	7	26	7
	i	1	36	3	**201**	3	4	15
	ô	10	17	15	2	**136**	17	61
	ó	22	4	19	4	7	**176**	11
	u	12	11	5	11	53	25	**148**

das vogais são classificadas corretamente, conforme mostrado na Figura 4.3. A vogal ***u*** possui o pior reconhecimento com 59,2% de corretos reconhecimentos. A vogal ***é*** possui um desempenho intermediário com 61,2%, seguido das vogal ***ó*** com 62,7%, das amostras classificadas corretamente. As demais vogais conseguem reconhecer corretamente mais que 63% das amostras de teste.

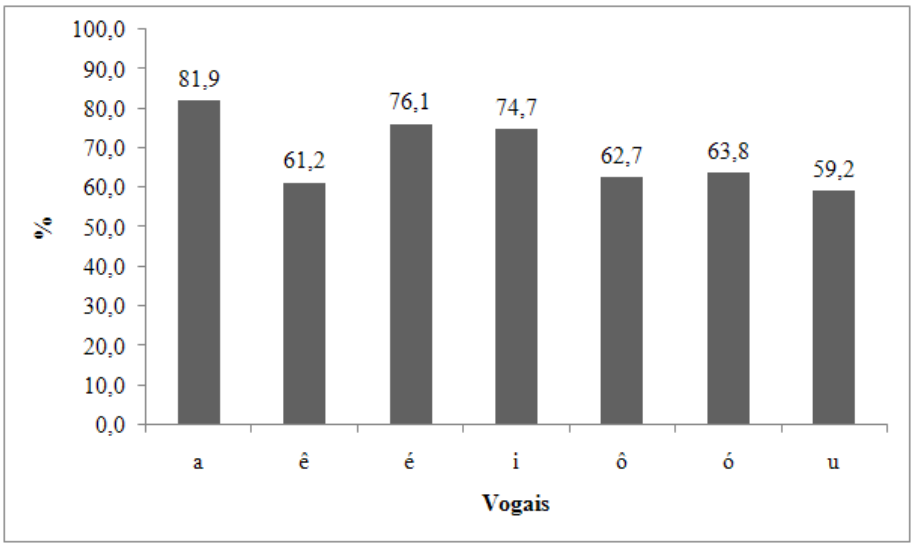


Figura 4.3: porcentagem de reconhecimento por vogal na configuração Daubechies-6.

Assim como no método proposto por Farooq e Datta (2003), a técnica proposta obteve o melhor desempenho classificatório médio para configuração com 6 níveis de decomposição *wavelet* por sub-frame, ou seja, na configuração Coiflet-6, com uma taxa média de 74,01%.

Na Tabela 4.4 é mostrado a matriz de confusão para configuração com melhor

Tabela 4.4: matriz de confusão da configuração Coiflet-6.

		Vogal Esperada						
		a	ê	é	i	ô	ó	u
Vogal Reconhecida	a	**219**	9	2	0	21	16	7
	ê	9	**184**	24	25	18	4	8
	é	5	23	**190**	15	15	16	11
	i	0	15	2	**192**	5	3	12
	ô	5	18	18	5	**186**	20	29
	ó	18	4	17	6	8	**182**	12
	u	10	7	6	3	12	10	**194**

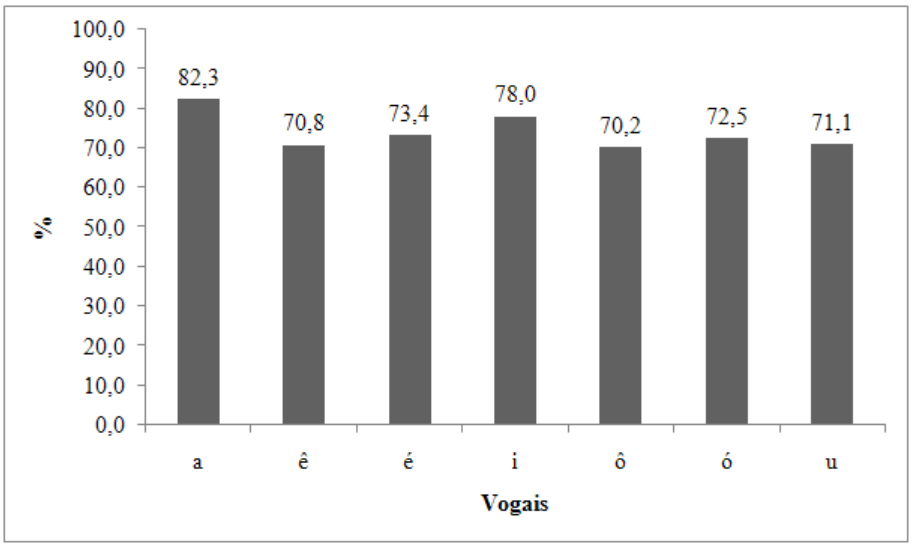


Figura 4.4: porcentagem de reconhecimento por vogal na configuração Coiflet-6.

taxa média, Coiflet-6, em que verifica-se que pelo menos 70% e no máximo 83% das vogais são classificadas corretamente, conforme mostrado na Figura 4.4. As vogais ***a*** e ***e*** conseguiram o melhor reconhecimento com 82,3% e 78% de corretos reconhecimentos respectivamente. As demais vogais tiveram reconhecimento correto inferior a 74% das amostras de teste, com o pior reconhecimento de 70,2% para a vogal ***ô***.

4.2 Custo computacional

Na Tabela 4.5 são mostrados o tempo total médio de treinamento e de reconhecimento das técnicas estudadas neste trabalho. A Figura 4.5 mostra a relação entre o numero de características extraídas e o tempo médio de treinamento do classificador para as técnicas estudadas.

Através dos resultados da Tabela 4.5 e da Figura 4.5, quanto maior o número de características extraídas em cada frame de áudio, maior o custo computacional de treinamento. Isto ocorre devido a configuração do classificador, em que a quantidade de neurônios da camada de entrada é igual ao número de características extraídas. Através da Figura 4.5, observa-se que o método proposto possui o maior tempo total médio de treinamento, enquanto a extração baseada em LP tem o menor.

Como se pode verificar comparando as Figuras 4.1 e 4.5, a extração baseada em LP tem o melhor custo-benefício, pois este método consegue altas taxas de acerto associado a um baixo custo computacional de treinamento do classificador.

Tabela 4.5: tempo total médio de treinamento e de teste das técnicas estudadas.

Configuração	**Número de Características**	Tempo Total Médio	
		Treinamento (s)	Teste (ms)
LP-8	8	20,35	11,7
LP-12	12	22,61	12,4
LP-16	16	27,80	21,8
LP-20	20	59,75	24,3
LP-24	24	72,51	22,1
LP-28	28	98,93	23,1
Daubechies-1	8	23,24	19,3
Daubechies-2	12	54,89	14,5
Daubechies-3	16	76,07	13,6
Daubechies-4	20	102,37	13,2
Daubechies-5	24	108,10	13,9
Daubechies-6	28	138,38	14,0
Coflet-1	8	53,45	21,6
Coflet-2	12	82,98	20,7
Coflet-3	16	132,28	21,2
Coflet-4	20	151,92	23,4
Coflet-5	24	176,76	20,0
Coflet-6	28	208,70	25,6

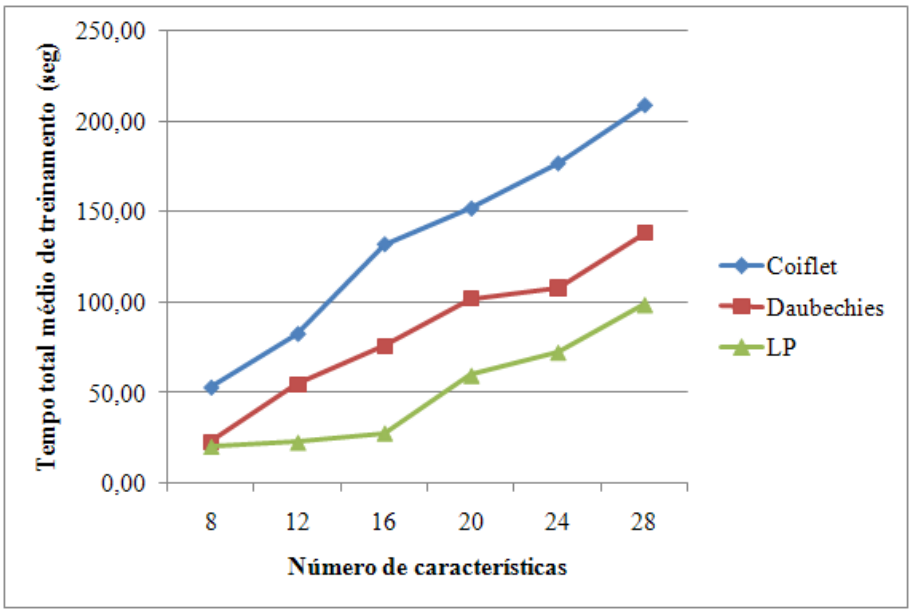


Figura 4.5: tempo total médio de treinamento.

4.3 Resumo do Capítulo

Neste capítulo, foram descritos os resultados obtidos para as técnicas de extração de características estudadas. Com base nesses resultados, observou-se que o método proposto por Kshirsagar e Magnenat-Thalmann (2000) obteve o melhor custo-benefício computacional, associando o melhor desempenho médio de classificação com menor custo computacional no treinamento e teste do classificador. Além disso, observou-se que o modificação proposta no técnica baseada Transformada Wavelet obteve um desempenho médio melhor que o método proposto por Farooq e Datta (2003). No capítulo seguinte são apresentandos as considerações finais e as perspectivas de trabalhos futuros.

Capítulo 5

Conclusões e Perspectivas

Neste capítulo são apresentadas as conclusões finais deste trabalho e as perspectivas futuras.

5.1 Conclusões Finais

Este trabalho realiza o estudo de algumas técnicas de extração de características para reconhecimento de fonemas e propõe uma modificação na técnicas baseada Transformada *Wavelet*.

Os resultados mostram que a técnica baseada em Predição Linear (LP) proposta por Kshirsagar e Magnenat-Thalmann (2000) obteve o melhor desempenho médio de classificação e menor custo computacional no treinamento do classificador.

Além disso, os resultados mostram que a modificação proposta na técnica baseada Transformada *Wavelet* foi capaz de melhorar em, aproximadamente, 6% o resultado do método proposto por Farooq e Datta (2003), possuindo um bom desempenho para as diversas classes de vogais.

O baixo desempenho das técnicas baseadas em Transformada *Wavelet* comparado com a técnica baseada em LP ocorre em função da taxa de amostragem estudada (8kHz). Contudo, segundo Farooq e Datta (2003), a taxa mínima de amostragem de 16kHz é necessária para o melhor funcionamento deste tipo de método. Desta forma, uma taxa de amostragem inferior ao recomendado resulta em perca de informação referente as componentes de alta frequência desse sinal e, por consequência, uma menor taxa de acerto.

Assim, para sinais de áudio com baixa taxa de amostragem, a técnica de análise no tempo, baseada em LP, consegue extrair melhor as informações das vogais do que as técnicas de análise na frequência, e, com isso, obteve melhores taxa de reconhecimento.

Com este trabalho, conclui-se que para as aplicações que utilizem a configuração mínima de digitalização de um sinal de voz, a técnica baseada em Predição Linear é a mais indicada, pois tem melhor custo-benefício: altas taxas de acerto associado a um baixo custo computacional de treinamento do classificador.

5.2 Perspectivas Futuras

Como proposta de trabalhos futuros sugere-se:

- extensão desse trabalho avaliando outras técnicas presentes na literatura de extração de características para reconhecimento de fonemas;
- utilizar as técnicas estudadas de extração de características da voz para detecção de doenças da laringe;
- estudo comparativo de técnicas de reconhecimento para fala contínua.

Referências Bibliográficas

ARAÚJO, R. T. S. *Detecção de Manchas de Óleo na Superfície do Mar em Imagens de Radar de Abertura Sintética*. Dissertação (Mestrado) — Universidade Federal do Ceará, 2004.

BATHAGLINI, M. G. *Reconhecimento de Voz*. 2009. Disponível em: <http://www-usr.inf.ufsm.br/~maicongb/trabalho.html>. Acesso em: 06 de agosto de 2009.

COSTA, R. C. S. Inspeção Automática de Laranjas Destinadas à Produção de Suco, Utilizando Técnicas de Processamento Digital de Imagens. *Monografia de Final de Curso, Centro Federal de Educação Tecnológica do Ceará*, 2006.

DAUBECHIES, I. *Ten lectures on wavelets*. Philadelphia, Pennsylvania: Society for Industrial Mathematics, 1992.

DURBIN, J. The fitting of time-series models. *Revue de l'Institut International de Statistique*, WP Van Stockum & Fils, p. 233–244, 1960.

FAROOQ, O.; DATTA, S. Phoneme recognition using wavelet based features. *Information Sciences*, ELSEVIER SCIENCE INC, v. 150, n. 1-2, p. 5–15, MAR 2003.

GONZALEZ, R. C.; WOODS, R. E. *Digital Image Processing*. 3rd. ed. New York, NY: Academic Press, 2008.

HAYKIN, S. S. *Redes Neurais: Princípios e Prática*. 2nd. ed. Porto Alegre: Bookman, 2001.

KALAGASIDIS, A. S. *et al.* The international building physics toolbox in simulink. *Energy and Buildings*, v. 39, p. 665–674, 2007.

KSHIRSAGAR, S.; MAGNENAT-THALMANN, N. Lip Synchronization Using Linear Predictive Analysis. *MIRALAB, CUI, University of Geneva*, 2000.

LEVINSON, N. The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction. *Journal of Mathematics and Physics*, Elsevier Science Inc, v. 25, n. 4, p. 261–278, 1946.

MAKHOUL, J. Linear Prediction: A Tutorial Review. *Proceedings of the IEEE*, IEEE-Institute Electrical Electronics Engineers Inc., v. 63, n. 4, p. 561–580, 1975.

PAULA, M. B. Reconhecimento de palavras faladas utilizando redes neurais artificiais. *Monografia de Final de Curso, UFPEL*, 2000.

PEREIRA, D. *et al.* Avaliação de Filtros Wavelets Aplicados no Pré-Processamento de Imagens Mamográficas. *Anais do XI Congresso Brasileiro de Informática em Saúde, Campos do Jordão*, 2008.

RIOUL, O.; VETTERLI, M. Wavelets and signal processing. *Signal Processing Magazine, IEEE*, 1991.

ROBINSON, E. Statistical Communication and Detection. *Griffin, London*, p. 249, 1967.

SOUZA JÚNIOR, A. H. D. Avaliação de Redes Neurais Auto-organizáveis para reconhecimento de voz em sistemas embarcados. *Dissertação de Mestrado, UFC*, 2009.

TAYLOR, C. J. Macroscopic traffic flow modelling and ramp metering control using matlab/simulink. *Environmental Modelling and Software*, v. 19, p. 975–988, 2004.

WAKITA, H. Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms. *Transactions on audio and eletroacoustics,IEEE*, v. 21, n. 5, October 1973.