Leonardo Alfredo Forero Mendoza

Redes Neurais e Máquinas de Vetores de Suporte no reconhecimento de locutor usando coeficientes MFC e características do sinal glotal

Dissertação submetida ao Programa de Mestrado em
Engenharia de Telecomunicações da
Universidade Federal Fluminense como parte
dos requisitos para obtenção do grau de Mestre.

Professores Orientadores:

Edson Luiz Cataldo Ferreira, D. Sc. (UFF)
Marley Vellasco, D. Sc. (PUC-Rio)

Niterói
2009

Declaração de Originalidade

Esta dissertação foi produzida por mim e relaciona trabalho original de minha própria execução. A menos que de outra forma mencionado, os gráficos e tabelas exibidos foram produzidos a partir de dados obtidos durante a pesquisa. Sempre que materiais, idéias, ou algoritmos computacionais de outros pesquisadores tiveram sido usados ou adaptados, a fonte de informação foi claramente especificada. Esta dissertação não foi submetida para graduação ou qualificação profissional em nenhum outro lugar.

Leonardo Alfredo Forero Mendoza

Agradecimentos

A Deus, por sempre fazer as coisas acontecerem para mim.

À minha família, que sempre me apoiou em todos os desafios que resolvi enfrentar, inclusive o início do mestrado.

Ao professor Edson Cataldo, pela orientação, por sempre mostrar boa vontade e por demonstrar confiança em meu trabalho.

À professora Marley, que me abriu as portas da PUC, pela orientação e pela constante disponibilidade para me atender.

Ao professor Andres Pablo, pelo incondicional apoio, pela amizade e por colocar à disposição o que eu precisasse para finalizar meu trabalho.

Ao curso de Mestrado em Engenharia de Telecomunicações da Universidade Federal Fluminense, que me concedeu esta grande oportunidade de aumentar meus conhecimentos.

À CAPES, por oferecer a bolsa de estudos, tornando possível a conclusão do mestrado.

Aos funcionários da UFF e a todos os companheiros do LACOP, por sempre me ajudarem a me sentir em casa.

Resumo

Este trabalho apresenta uma proposta de reconhecimento automático de locutor usando máquina de vetores de suporte e redes neurais. O vetor de entrada usado é híbrido composto de coeficientes MFC (Mel Frequency Cepstral Coeffiecients) e características extraídas do sinal glotal, obtida por filtragem inversa do sinal de voz. Os resultados são comparados com outros obtidos quando apenas os coeficientes MFC são usados na entrada

Palavras chave: Reconhecimento de locutor. Filtragem inversa. Maquina de vetores de suporte. Parameters of the glottal signal.

Abstract

This work presents a proposal for automatic speaker recognition using support vector machine and neural networks. The input vector is hybrid composed by MFC coefficients (Mel Frequency Cepstrum Coefficients) and features extracted from the glottal signal, obtained by inverse filtering of the voice signal. The results are compared with other obteined when only the MFC coefficients are used as input.

Keywords: Recognition speech. Inverse filtering. Support vector machine.

Dedicatória

Dedico este trabalho a:

Maria Tereza, Alexandra, Sofia e Nohemy,

Guillermo, Luis Alejandro,

Pablo, Luis Alfredo,
Vagales.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

1.1 Introdução

A partir da voz, torna-se possível identificar características próprias de cada pessoa como idade, sexo, língua e até mesmo sua identidade. Dessa forma, pode-se construir um sistema de reconhecimento, cujo objetivo é reconhecer um locutor a partir de sua voz, podendo ser utilizado, por exemplo, em aplicações de segurança e perícia criminal [1].

Os primeiros trabalhos descrevendo máquinas que podiam reconhecer, com certo sucesso, a pronúncia de determinadas palavras (reconhecimento de voz) datam dos anos 50 e tiveram seu auge nos anos 60 [2] graças às descobertas de propriedades da voz através do uso de espectrógrafos [3].

O reconhecimento automático do locutor (RAL) tem alcançado resultados bem satisfatórios, com o crescente aumento da capacidade computacional, tanto em velocidade de processamento digital, quanto em memória. De modo geral, ele é realizado em duas partes: Primeiro, a extração das características da voz do locutor, que busca obter impressões do locutor que sejam inerentes a ele, e em seguida o reconhecimento de padrões, que busca a separação entre os padrões verdadeiros e falsos.

O estudo do sinal glotal e de seus parâmetros, nos últimos anos, vem sendo utilizado em diferentes pesquisas e aplicações sobre a produção, codificação,

síntese, uso clínico da voz [4] e, também, em uma tentativa de quantificar a sua contribuição na transmissão de sentimentos [5]. Porém, sua utilização em reconhecimento automático de locutor é ainda restrita, devido à dificuldade de se obter o sinal glotal pois, em geral, as técnicas utilizadas para sua obtenção são invasivas como, por exemplo, a estroboscopia [6] [7], ou, quando não invasivas, têm a necessidade da utilização de aparelhos caros e difíceis de serem encontrados, como o eletroglotrógrafo [8] [9] [10].

O método de filtragem inversa semi-automático, desenvolvido em [11], conhecido como IAIF, apresenta uma técnica na qual uma estimativa do sinal glotal pode ser obtida a partir do sinal de voz do locutor, eliminando a influência do trato vocal. Dessa forma, tornou-se possível realizar o levantamento de características do sinal glotal a partir da gravação da voz do locutor.

Em reconhecimento automático de locutor, a técnica de extração de características denominada MFCC (Mel frequency Cepstral Coeficients), é bem conhecida e bem difundida e é uma técnica que vem proporcionado bons resultados [12], tanto na utilização com Redes neurais como em Modelos ocultos de Markov [13].

Neste trabalho, procurou-se unir os coeficientes MFC com as medidas obtidas diretamente do sinal glotal, de forma a verificar se ocorre uma melhora no desempenho do reconhecimento automático de locutor.

O reconhecimento de padrões, que é o caso de RAL, é realizado de modo que o computador reconheça padrões apresentados em sua entrada e gere em sua saída, resultados satisfatórios.

Nos últimos anos, há se tentado fazer algoritmos que simulem a capacidade de reconhecimento do cérebro humano em sistemas computacionais. Uma dessas áreas é conhecida como redes neurais artificiais [14], as quais são

utilizadas neste trabalho. Utilizamos, também, a técnica conhecida como "Máquina de Vetores de Suporte"(SVM) e comparamos os resultados obtidos.

1.2 Objetivos de Dissertação

-Obter uma estimação do sinal glotal por meio do método da filtragem inversa e extrair suas características para reconhecimento de locutor.

-Construir um vetor híbrido de características combinando os coeficientes MFC e algumas características do sinal glotal, para criar uma ferramenta para reconhecimento de locutor. Usar esse vetor híbrido como entrada em uma rede neural e, também, em uma máquina de vetores de suporte.

-A partir de uma "base de dígitos"segmentar cada palavra gravada extraindo o primeiro som vozeado pronunciado e construir uma base nova de vogais.

-Comparar o desempenho da rede neural e da técnica de máquina de vetores de suporte.

1.3 Estado da arte

Os reconhecimentos da voz e locutor, sem distorções, estão praticamente dominados utilizando as características MFC e utilizando HMM (Modelos Ocultos de Markov) para sua classificação [13], mas com ruído aditivo a situação muda por completo. Devido a isso, vem ocorrendo uma intensificação dos

estudos, visando aumentar a robustez dos reconhecimentos de voz e locutor em suas etapas de extração de características e em sua classificação. Novas técnicas incluindo a obtenção de outras características mais robustas têm sido objeto de estudos para serem combinados com os modelos estocásticos existentes [5].

O sinal glotal, que é a fonte geradora da voz, vem sendo estudada em pesquisas clínicas e para quantificar a contribuição do pulso glotal na transmissão de sentimentos [5], mas seu estudo para reconhecimento de locutor ainda não é muito explorado devido a dificuldade de sua obtenção, mas já foi comprovada que suas características funcionam para discriminação entre locutores [15]. Na classificação de padrões a intenção está concentrada no melhor aproveitamento das informações obtidas na fase de processamento da voz, utilizando sistemas inteligentes que analisam o significado do resultado obtido, em outras palavras, a coerência e o sentido. Atualmente vem-se utilizando técnicas hibridas combinando redes neurais e HMM com bons resultados [16], no entanto, pesquisadores vem trabalhando em diferentes técnicas de classificação que possam ser combinadas com as atuais para obter melhores resultados.

1.4 Contribuições desta Dissertação

Como contribuições desta dissertação, pode-se destacar

-Obter uma estimação do sinal glotal por meio do método da filtragem inversa e extrair suas características para reconhecimento de locutor.

- Um novo método para reconhecimento de locutor, usando redes neurais

e máquina de vetores de suporte, com um vetor de entrada híbrido, com coeficientes MFC e características extraídas do sinal glotal.

-A partir de uma “base de dígitos”segmentar cada palavra gravada extraindo o primeiro som vozeado pronunciado e construir uma base nova de vogais.

Capítulo 2

Fundamentos da produção da voz

2.1 A Voz Humana

A voz é uma característica humana, relacionada com a necessidade do homem de se comunicar e se agrupar, já que a voz não transmite só informação léxica, mas também emoções, como dor e alegria, através de sua entonação. A voz tem sons padrões que são associados com a fala e a comunicação verbal, e de acordo com as regras desta comunicação, mudam características da voz, como intensidade, altura, inflexão, ressonância, articulação e muitas outras características que são determinadas em sua produção [17]. Esta produção é um trabalho conjunto do sistema nervoso, respiratório e digestivo, além de músculos, ligamentos e ossos, que se unem apropriadamente.

Originalmente, o aparato fonador humano não foi feito para a produção da voz. Músculos, como as cordas vocais, foram desenvolvidos, em primeiro lugar, para as funções de respiração e alimentação. A evolução para a geração da voz foi detectada no *homo sapiens* e foi fundamental na formação das

sociedades.

2.2 O processo de produção da voz humana

A voz humana é produzida por meio do aparelho fonador, formado pelos pulmões, pela laringe, pela faringe, pelas cavidades orais (ou bucais) e nasais e por vários elementos articulatórios: os lábios, os dentes, o alvéolo, o palato, o véu palatino e a língua. A Fig 2.1 mostra um esquema do aparelho fonador. As cordas vocais, principais elementos para a geração da voz, são duas membranas situadas na laringe .

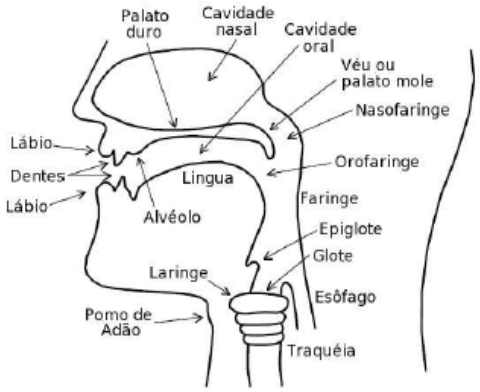

Figura 2.1: Aparelho fonador.

A abertura entre as cordas vocais se denomina glote. A produção da voz se inicia com uma expansão - contração dos pulmões, que gera uma diferença entre a pressão do ar nos pulmões e a pressão do ar próximo a boca, causando um escoamento de ar. O ar proveniente dos pulmões é forçado através do pequeno espaço existente entre as cordas vocais, causando o movimento das

cordas em uma frequência determinada pela tensão dos músculos associados. Este movimento modifica o fluxo de ar resultando em pulsos de ar (conhecidos como sinal glotal) que serão amplificados e modificados pelas cavidades oral e nasal, até serem irradiados pela boca. Os pulsos de ar são modulados pela língua, pelos dentes e lábios; de forma a produzir o que conhecemos por voz.

Fisiologicamente, são três subsistemas que atuam de modo sucessivo na produção da fala:

1- Respiratório: é o responsável pela produção do ar necessária para produzir a voz e é responsável pela passagem da corrente de ar nos pulmões pela traquéia e pela laringe. Em resumo, é a fonte de energia da voz e dele depende a intensidade e duração da voz.

O subsistema respiratório é composto pelos músculos intercostais (abdominais), diafragma, pulmões, brônquios e traquéia.

2- Laríngeo: é o mais importante do aparato fonador, pois nele se encontram as cordas vocais. Quando a corrente de ar passa pelas cordas vocais produz-se a onda sonora que é a energia acústica audível. As cordas vocais são as responsáveis pelo tom e dependem da abertura da glote e também de algumas das propriedades da voz, já que acelera ou desacelera a corrente de ar, como mostra a Fig 2.2.

O subsistema laríngeo é formado pela laringe, pelas cordas vocais, pela glote e a epiglote.

3-Supralaríngeo: é o responsável pela ressonância e pela articulação da voz. Na articulação, a onda sonora que se produz nas cordas vocais é filtrada. Este filtro atua modificando o espectro do som, aportando a característica

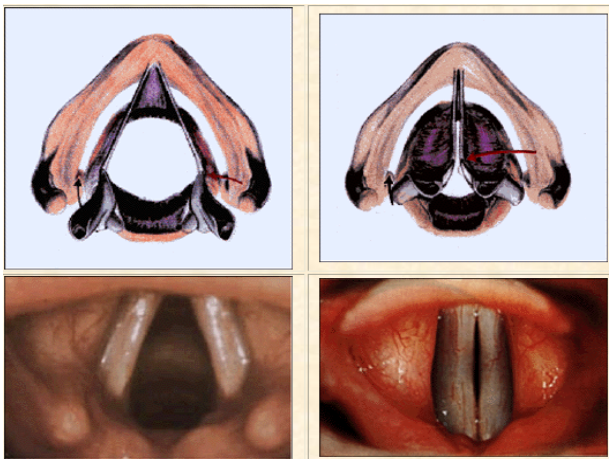

Figura 2.2: Cordas Vocais.

própria da voz para cada indivíduo. Isso ocorre nas quatro principais cavidades supralaríngeas.

A faringe, a cavidade nasal, a cavidade oral e a cavidade labial fazem a tarefa de ressonadores acústicos que dão ênfases a certas bandas de frequências do espectro gerado pelas cordas vocais. A forma do trato vocal determina suas frequências naturais e, consequentemente, as vogais a serem pronunciadas.

Os fonemas

A informação transmitida através da voz é intrinsecamente discreta, isto é, ela pode ser representada pela concatenação de elementos de um conjunto finito de símbolos, chamados fonemas. Um fonema é a menor unidade sonora de uma língua que estabelece contraste de significado para diferenciar palavras.

A maioria dos idiomas pode ser descrito em termos do conjunto de fonemas que possui. Este conjunto de símbolos básicos possui normalmente de

30 a 50 elementos que podem ser divididos basicamente em 4 classes: vogais, ditongos, semivogais e consoantes.

A fonética e a fonologia têm sido consideradas como distintas, estando a primeira voltada às propriedades físicas dos sons da fala e a segunda ao conjunto de representações dos sons distintivos na língua no sistema cognitivo. Neste trabalho, será detalhada a classe das vogais.

As vogais são produzidas pela excitação do trato vocal por pulsos de ar quase periódicos, causados pela vibração das cordas vocais. A Fig 2.3 mostra o sinal de voz correspondente à produção de uma vogal /a/.

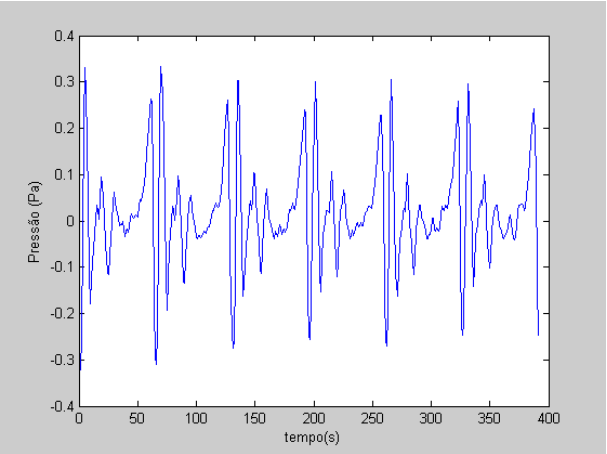

Figura 2.3: Sinal de voz correspondente a um trecho da vogal sustentada /a/ obtida com uma frequência de amostragem fs=11.025Hz.

Os segmentos vocálicos possuem fonte de sonoridade contínua e trato vocal supraglótico, sem obstrução à passagem do ar. A qualidade sonora de cada segmento vocálico passa a ser dependente da conformação das cavidades supraglóticas, as quais geram frequências de ressonância denominadas formantes [19].

Em Português, as vogais são classificadas de acordo com a Nomenclatura

Gramatical Brasileira (NGB), considerando: a zona de articulação (conforme o posicionamento da língua), o timbre, o papel das cavidades bucal e nasal e a intensidade (átonas ou tônicas).

2.2.1 Modelo de produção sonoro/surdo da voz

Para modelar detalhadamente o processo de produção da voz, os seguintes efeitos devem ser considerados: a variação da configuração do trato vocal com o tempo, perdas por condução de calor e fricção nas paredes do trato vocal, radiação de som pelos lábios, a maciez das paredes do trato vocal, junção nasal e a excitação do som no trato vocal [16]. Um modelo detalhado para geração de sinais de voz, que leva em conta os efeitos da propagação e da radiação conjuntamente pode, em princípio, ser obtido através de valores adequados para excitação e parâmetros do trato vocal. A teoria acústica sugere uma técnica simplificada para modelar sinais de voz, a qual é bastante utilizada. Essa técnica apresenta a excitação separada do trato vocal e da radiação. Os efeitos da radiação e do trato vocal são representados por um sistema linear variante com o tempo. O gerador de excitação gera um sinal similar a um trem de pulsos, ou sinal aleatório (ruído). Os parâmetros da fonte e sistema são escolhidos de forma a se obter na saída o sinal de voz desejado [16]. Na Fig 2.4, $u(n)$ é o sinal de excitação, $A_s(n)$ e $A_f(n)$ controlam a intensidade da excitação do sinal sonoro e do ruído, respectivamente, onde ocorre um chaveamento entre sonoro e surdo, alterando o modo de excitação.

2.2.2 A Teoria fonte-filtro

A teoria fonte-filtro considera a produção da fala dividida em duas partes independentes: a primeira é a fonte de sons, onde se produz o sinal de voz

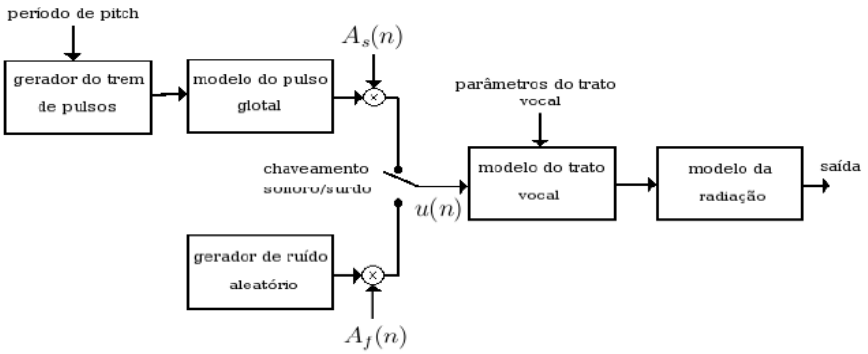

Figura 2.4: Modelo discreto da produção da voz.

(laringe) e a segunda é um sistema de filtros em série que modificam o sinal (trato vocal) [18] como mostra a Fig 2.5. Na prática, existe uma interação entre a fonte e o trato vocal. Porém, a validade da teoria pode ser considerada

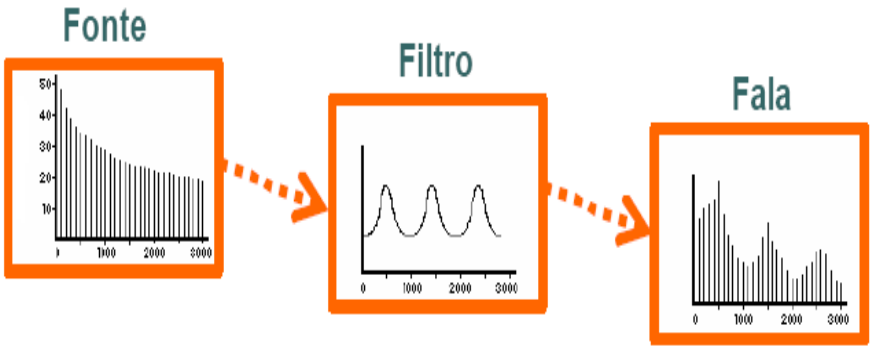

Figura 2.5: Representação da Teoria Fonte Filtro

suficiente para a maioria dos casos de interesse, sendo muito utilizada em processamento digital de sinais.

2.3 Pré-processamento da voz

A voz humana é um sinal de pressão acústica que varia com o tempo. Esse sinal, analógico, pode ser convertido em um sinal digital de modo a possibilitar seu processamento através de programas de computador. O processo de digitalização, como é mostrado na Fig 2.6, começa com a captura do sinal de áudio, por um microfone para converter o sinal de voz em sinal elétrico. Logo, passa por um filtro analógico chamado de *anti-aliasing* para eliminar as frequências altas e possibilitar o uso do Teorema da Amostragem.

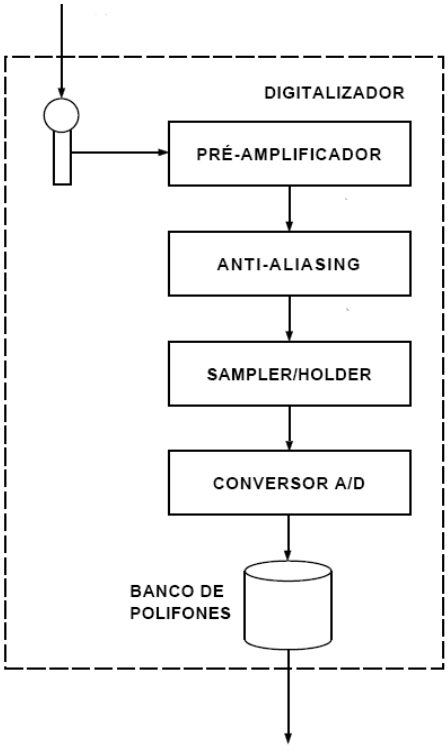

Figura 2.6: Diagrama de blocos do digitalizador.

Depois, o sinal obtido é amostrado com uma frequência de amostragem maior que o dobro da frequência máxima do sinal, segundo o Teorema da

Amostragem [22]. Em seguida, o sinal amostrado é quantizado com uma determinada resolução. Neste trabalho, o objetivo do pré-processamento é obter uma representação paramétrica dos sinais, que reduza redundâncias, mantendo informações estatísticas suficientes para o reconhecimento.

Pré-ênfase

A filtragem de pré-ênfase serve para atenuar as componentes de baixa frequência e incrementar as componentes de alta frequência do sinal de voz, prevenindo contra instabilidade numérica, também, minimizando o efeito dos lábios.

A pré-ênfase das frequências altas é necessária para que se obtenham amplitudes mais homogêneas das frequências formantes, porque informações importantes sobre a locução também estão presentes nas altas frequências[22].

A função de transferência mais usada para um filtro de pré-ênfase é dada por:

$$H(z) = 1 - az^{-1}, \quad 0 \leqslant a \leqslant 1. \tag{2.1}$$

Neste caso, a saída do sistema de pré-ênfase $\tilde{s}(n)$ está relacionada à entrada $s(n)$ pela equação de diferenças:

$$\tilde{s}(n) = s(n) - as(n-1) \tag{2.2}$$

onde o valor de a usualmente usado é 0,95.

Janelamento

Após a pré-ênfase, passa-se à etapa de "janelamento" na qual o sinal de voz é dividido em segmentos. Nesta etapa, são extraídos quadros de N amostras a partir do sinal $\tilde{s}(n)$, tendo uma superposição de M amostras (ver Fig. 2.7).

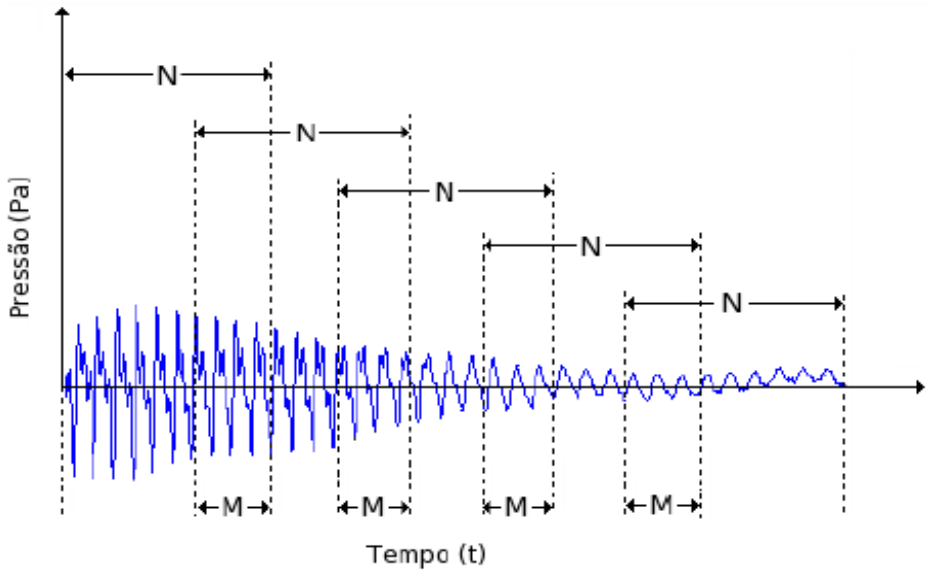

Figura 2.7: Divisão em quadros do sinal de voz.

O janelamento de pequenos segmentos, que variam de 10ms a 45 ms segundo aplicação, se precissa mais exatitude a janela é mais pequena.O janelamento possibilita minimizar as descontinuidades do sinal no começo e no final de cada janela (*frame*) e admitir que ele seja aproximadamente estacionário nesses intervalos, permitindo, assim, o uso de métodos tradicionais de análise espectral. Geralmente, para separar cada segmento do sinal de

voz, usa-se uma janela de Hamming [16], definida por:

$$h(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), & 0 \leq n \leq N-1 \\ 0, & c.c. \end{cases} \tag{2.3}$$

onde n é o índice da amostra e N é o número total de amostras da janela.

Capítulo 3

Coeficientes Cepstrais de Frequência Mel (MFCC)

3.1 Extração de características em sinais de voz

Um grande número de características da voz podem ser extraídas, para o uso no Reconhecimento Automático de Locutor (RAL), com técnicas tais como taxa de cruzamento de zeros, energia, frequência fundamental da voz e Coeficientes Cepstrais de Frequência Mel (MFCC). Neste capítulo, será apresentada a técnica de extração de características através dos Coeficientes Cepstrais de Frequência Mel (MFCC).

A técnica MFCC surgiu devido aos estudos na área de psicoacústica (a ciência que estuda a percepção auditiva humana). Esta ciência mostra que a percepção das frequências de tons puros ou de sinais de voz não seguem uma escala linear, impulsando assim a criação de uma escala que se aproxima desta percepção, sendo chamada escala *mel*.

3.1.1 Escala mel

No estudo da dinâmica do sistema auditivo humano definiu-se uma escala psicoacústica de sensibilidade do ouvido para diversas frequências do espectro audível, conhecida como escala *mel*. A escala "*Mel*" foi desenvolvida por Stevens e Volkman, em 1940 [21]. A escala mel baseia-se no sistema de audição humano, cuja sensibilidade aos sinais de voz se processa em uma escala não-linear de frequências. O mel é a unidade de medida de um tom, isto é, de uma frequência única percebida pelo ouvinte. Como referência, definiu-se a frequência de 1 KHz, 40 dB acima do limiar de audição do ouvido, como 1000 mels. Os outros valores subjetivos foram obtidos através de experimentos onde pedia-se a ouvintes que ajustassem a frequência física de um tom até que a frequência percebida fosse igual a duas vezes a frequência de referência, depois, 10 vezes a frequência de referência e assim por diante. Essas frequências teriam os valores de 2000 mels, 10000 mels e assim por

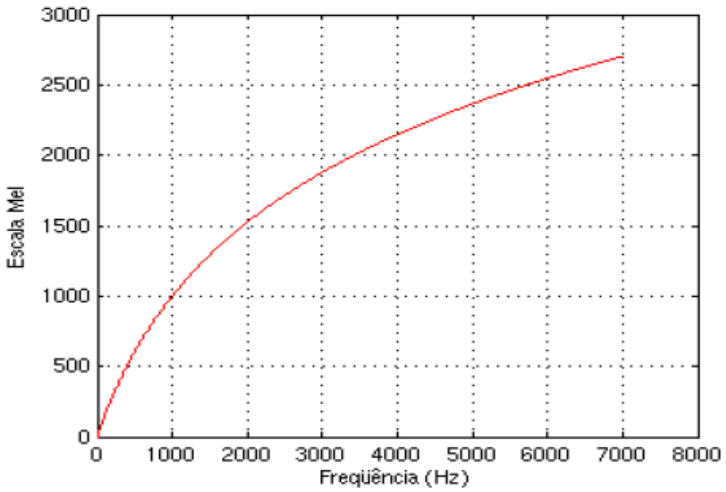

Figura 3.1: Escala Mel

diante como é mostrado na Fig 3.1.

A equação que descreve a escala mel é:

$$Mel(f) = 1127 \ln \left(1 + \frac{f}{700}\right). \tag{3.1}$$

3.1.2 Banda crítica

Alguns experimentos demostraram que a percepção humana de algumas frequências de sons complexos não podem ser individualmente identificadas, dentro de certas bandas [21]. Quando uma componente cai fora da banda, chamada de banda crítica, ela pode ser identificada. Uma explicação apresentada para esse fato é que a percepção de uma frequência particular pelo sistema auditivo, por exemplo f, é influenciada pela energia de certa banda de frequências em torno de f, o valor dessa banda varia nominalmente de 10% a 20% da frequência central do som, começando em torno de 100 Hz para frequências abaixo de 1 KHz e aumentando em escala logarítmica acima.

Cabe destaque à representação *cepstral* associada à escala mel apresentando maior eficácia computacional, sendo chamada de Mel-Cepstral.

3.1.3 Banco de filtros triangulares

A melhora do desempenho de sistemas de reconhecimento de voz e locutor, com o uso da escala mel aliada ao uso de bancos triangulares, tem se comprovado [16], originando a técnica MFCC [13]

Na Fig 3.2, apresenta-se a configuração de banco de filtros triangulares usado para o cálculo dos coeficientes MFC [20].

Para a faixa de frequências de interesse (300 Hz - 3.4 KHz), utilizam-se 20 filtros centrados nas frequências da escala mel. O espaçamento é de apro-

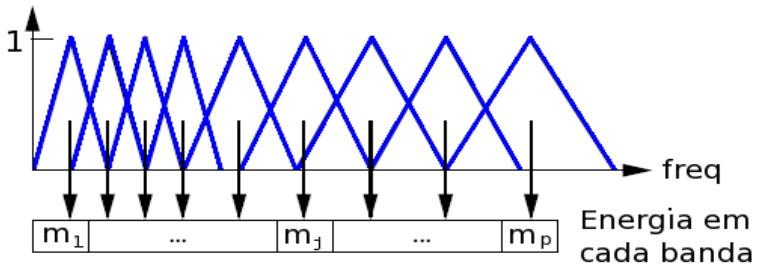

Figura 3.2: Banco de filtros usado na técnica MFCC.

ximadamente 150 mels e a largura de banda de cada filtro triangular é de 300 mels. Este banco de filtros simula a resposta em frequência da *membrana basilar*. Esses fenômenos (escala mel e banda crítica) sugeriram que seria mais interessante fazer algumas modificações na representação espectral do sinal. Tais modificações consistiram, primeiramente, em fazer uma ponderação da escala de frequência para a escala mel e, depois, incorporar a noção de banda crítica na definição de distorção espectral. Ou seja, ao invés de usar simplesmente o logaritmo da magnitude das frequências, passou-se a utilizar o logaritmo da energia total das bandas críticas em torno das frequências mel.

A aproximação mais utilizada para esse cálculo é a utilização de um banco de filtros triangulares, espaçados uniformemente em uma escala não linear (escala mel). A técnica de ponderação mel pode ser aplicada a vários tipos de representação espectral.

3.1.4 Cálculo dos MFCCs

Para o cálculo dos coeficientes MFC, primeiro, o sinal de voz $s(n)$ passa pela etapa de pré-ênfase, em seguida, o sinal resultante é dividido em pequenas janelas de Hamming. Para cada janela, m, estima-se o espectro $S(w, m)$,

utilizando a FFT. O espectro modificado $P(i),\ i = 1, 2, \ldots, N_f$, consistirá na energia de saída de cada filtro, expresso por:

$$P(i) = \sum_{k=0}^{N/2} |S(k,m)|^2 H_i \left(k \frac{2\pi}{N} \right) \tag{3.2}$$

onde N é o número de pontos da FFT, N_f é o número de filtros triangulares, $|S(k,m)|$ é o módulo da amplitude na frequência do k-ésimo ponto da m-ésima janela e $H_i(w)$ é a função de transferência do i-ésimo filtro triangular. Em seguida, define-se o conjunto de pontos $E(k)$ por:

$$E(k) = \begin{cases} \log[P(i)], & k = k_i \\ 0, & \text{qq outro } k \in [0, N-1] \end{cases} \tag{3.3}$$

onde k_i é o ponto máximo do i-ésimo filtro. Os coeficientes mel-cepstrais $c_{mel}(n)$ são então obtidos com o uso da Transformada Discreta de Coseno (DCT), dado por

$$c_{mel}(n) = \sum_{i=0}^{N_f} E(k_i) \cos \left(\frac{2\pi}{N} k_i n \right), \qquad n = 0, 1, 2, \ldots, N_c - 1 \tag{3.4}$$

onde N_c é o número de coeficientes mel-cepstrais desejado, N_f é o número de filtros e k_i é o ponto máximo do i-ésimo filtro. Por exemplo se $N_c = 15$ então se terá um vetor como é mostrado a seguir:

$$c_{mel} = c_0, c_1, c_2, \ldots, c_{13}, c_{14}.$$

Nesse vetor, considera-se o primeiro coeficiente, denotado por c_0 que pode carregar muita informação do meio de transmissão[34]. Este coeficiente por vezes é considerado e por vezes não; isto vai depender do tipo de reconhecimento que se deseja pode ser de voz ou locutor. A Fig.3.3 mostra o processo para obter os coeficientes MFC.

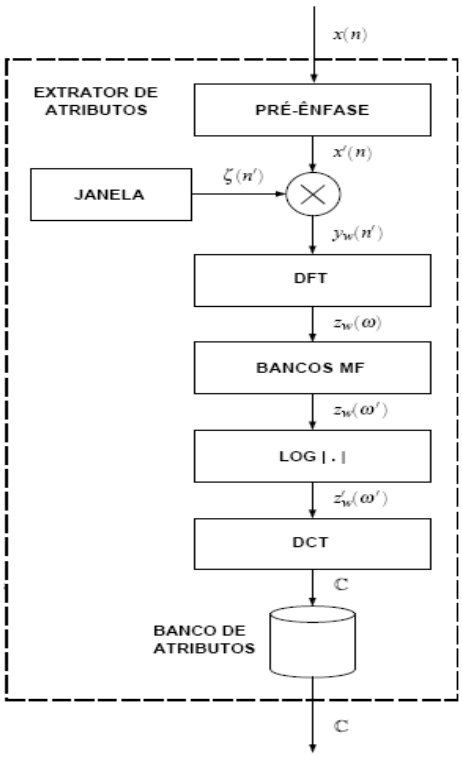

Figura 3.3: Diagrama de fluxo para o cálculo dos MFCCs.

3.1.5 Coeficientes Delta e Delta- Delta

Para melhorar o desempenho dos sistemas de reconhecimento adiciona-se mais informação do sinal, como por exemplo, a primeira e a segunda derivadas. Os coeficientes cepstrais, resultado do cálculo da DCT, são conhecidos também como coeficientes "estáticos"e os coeficientes obtidos a partir da primeira e segunda derivadas são chamados de coeficientes "dinâmicos", porque são utilizados para representar as mudanças dinâmicas no espectro da voz e, desse modo, detectar variações bruscas dentro do espectro. Uma equação

muito usada é a seguinte [24]:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}, \tag{3.5}$$

onde d_t é o coeficiente delta (Δ) no tempo t calculado em termos dos correspondentes coeficientes estáticos $c_{t-\Theta}$ até $c_{t+\Theta}$. O valor de Θ é o número de amostras necessárias para o cálculo dos coeficientes dinâmicos e este valor é normalmente achado de forma empírica; segundo a literatura , os valores mais típicos são de 2, 4 ou 8. Os parâmetros de segunda ordem são obtidos reaplicando a derivada sobre os resultados obtidos na primeira derivação.

Assim, por exemplo, se queremos calcular 12 coeficientes MFCC com seus respectivos coeficientes dinâmicos, teríamos no final 12 coeficientes estáticos, 12 coeficientes obtidos da primeira derivada (Δ) e mais 12 coeficientes obtidos da segunda derivada ($\Delta\Delta$); isto é, um vetor de 36 coeficientes[13]. Normalmente, o cálculo anterior é realizado sem considerar o primeiro coeficiente (c_0); por ter informações do meio do transmissão, portanto, se considerarmos o c_0, então teríamos um vetor de 39 coeficientes.

Capítulo 4

Extração de características do sinal glotal

4.1 Sinal glotal

Quando ocorre a expansão-contração dos pulmões, dá-se inicio à geração do sinal glotal, pois gera-se a diferença de pressão entre o ar nos pulmões e no ar próximo à boca. O fluxo do ar produzido por essa diferença de pressão passa através das cordas vocais que vibram em uma frequência relacionada à tensão dos músculos associados à produção da fala [25]. Esta vibração altera o fluxo de ar, transformando-o em um trem de pulsos ou sinal glotal. O processo da formação do sinal glotal é mostrado na Fig 4.1.

O sinal glotal possui propriedades importantes de difícil reprodução que estão intimamente ligadas às características anatômicas e fisiológicas da laringe. Atualmente, a teoria mais aceita para a descrição do sinal glotal (isto é, o aparecimento do trem de pulsos) é a teoria chamada de aerodinâmica mioelástica [26] [27]. Esta teoria postulou que os movimentos de abrir e fechar as cordas vocais são regidos pelas propriedades mecânicas dos tecidos

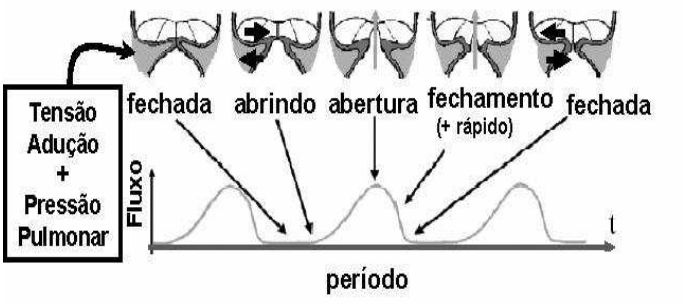

Figura 4.1: Formação do sinal glotal.

musculares que constituem, principalmente, as cordas vocais e pelas forças aerodinâmicas que se distribuem ao longo da laringe durante a fala.

A ação neural consiste apenas em aproximar as cordas vocais de tal forma que a superfície destas vibrem. O sinal glotal tem grande importância na determinação de sentimentos na voz e é utilizado em áreas de pesquisa clínica. A Fig. 4.2 é um exemplo de sinal glotal obtido através do sinal de voz, por filtragem inversa e será detalhado mais adiante .

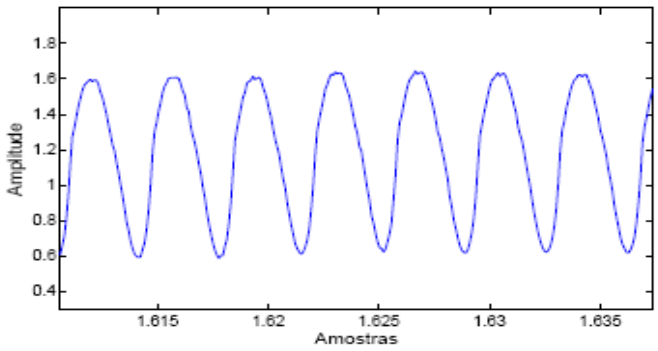

Figura 4.2: Sinal glotal da vogal sustentada representada na Fig.2.3, obtido por filtragem inversa.

4.2 Filtragem inversa

É uma técnica bastante empregada na estimação do sinal glotal. O fluxo de ar, proveniente dos pulmões, é alterado pela vibração das cordas vocais gerando o sinal glotal, que serve de excitação do trato vocal e gerando, finalmente, a voz. Portanto, o estudo do sinal glotal é de suma importância na compreensão da produção da voz.

Sua utilização em reconhecimento automático de locutor é ainda restrita, devido à dificuldade de se obter o sinal glotal pois, em geral, as técnicas utilizadas para sua obtenção são invasivas como, por exemplo, a estroboscopia [6], ou, quando não invasivas, têm a necessidade da utilização de aparelhos caros e difíceis de serem encontrados, como o eletroglotrógrafo [8]. A filtragem inversa tem a vantagem de dar uma estimativa do sinal glotal partindo do sinal de voz. As diversas versões desta técnica baseiam-se na mesma idéia; o pulso glotal é obtido cancelando os efeitos dos formantes na voz.

O trato vocal deve ser modelado e, então, os efeitos dos formantes são cancelados filtrando o sinal de voz através do inverso do trato vocal [11]. O PSIAIF (*Pitch Synchronous Iterative Adaptive Inverse Filtering*) é um método de filtragem inversa, semi-automático, desenvolvido por [11], que utiliza o sinal de voz como entrada e apresenta na saída uma estimação do fluxo glotal correspondente.

4.2.1 Algoritmo de filtragem inversa

A teoria fonte-filtro da produção da voz é a base teórica necessária para a criação da técnica de filtragem inversa. Se a função de transferência do filtro do trato vocal é conhecida, uma filtragem inversa poderá ser realizada. Em princípio, o sinal da excitação glotal pode ser reconstruído passando o sinal de

voz pelo inverso do filtro do trato vocal. Na prática, a função de transferência do filtro do trato vocal pode ser aproximada baseando-se no sinal de voz e no mecanismo de produção da voz. Aplicando a técnica de filtragem inversa ao sinal de voz, obteremos uma estimação da excitação glotal e a forma de onda do fluxo glotal, que também é conhecida como FGG (*flow glottogram*) [28] [29]. Atualmente, a maioria das técnicas de filtragem inversa são digitais devido à flexibilidade e facilidade de implementação quando comparadas aos filtros analógicos.

Os métodos de filtragem inversa digital podem ser divididos em técnicas manuais, semi-automáticas e automáticas. Os métodos manuais requerem o ajuste dos filtros para determinar os formantes do sinal de voz, diferentemente das técnicas automáticas que constroem um modelo do filtro do trato vocal e encontram os parâmetros dos filtros, normalmente por análise LPC (*linear prediction coefficients* [29]. Os métodos semi-automáticos encontram-se entre os dois extremos. O método proposto por [11] é um bom exemplo de método semi-automático, pois, basicamente, o filtro do trato vocal é encontrado automaticamente, mas o usuário pode controlar certos parâmetros que afetarão o resultado final do fluxo glotal. No trabalho, [30] comparou um método de filtragem inversa automático com um manual e concluiu que há extrema semelhança entre os resultados obtidos em cada método.

A filtragem inversa envolve, basicamente, a extração de dois sinais, o sinal glotal e o efeito do filtro do trato vocal, de uma única fonte de sinal. Entretanto, a técnica adota diversas aproximações a respeito do fluxo glotal e da função de transferência do trato vocal. Conseqüentemente, o resultado da filtragem inversa deve ser considerada como uma estimação do sinal glotal [31]. O fluxo glotal em si ainda não é conhecido exatamente. Em [15], há uma comparação entre a estimativa do sinal glotal e o sinal obtido pelo

electroglotógrafo encontrando muitas semelhanças e dando mais confiança na estimativa alcançada na filtragem inversa semi-automática.

A precisão da filtragem inversa se deteriora caso a frequência fundamental da voz seja alta, pois a estrutura espaçada dos harmônicos do espectro da excitação interfere nos formantes, que são ressonâncias locais no espectro[31].

4.2.2 Análise LPC

A predição linear (LPC) é uma técnica muito utilizada em processamento de sinais de áudio e processamento de voz e consiste em usar amostras anteriores do sinal para estimar a amostra atual. É usado em métodos de codificação de voz de alta qualidade. Estudos demonstraram [16] que um sinal de voz $s(t)$, pode ser visto como a saída de um filtro digital IIR (Resposta ao impulso infinita) cuja função de transferência é $1/A(z)$ (também chamado só-polo), excitado por uma sequência de impulsos que corresponde ao erro LPC do sinal $e(n)$, com transformada $E(z)$. Ou seja, a transformada z do sinal $s(t)$ é dada por:

$$S(z) = \frac{E(z)}{A(z)}. \tag{4.1}$$

O filtro $A(z)$ com coeficientes ate ordem M é dado por:

$$A(z) = 1 - \sum_{k=1}^{M} a_k z^{-k}, \tag{4.2}$$

e é chamado de filtro digital inverso [16]. De acordo com esse modelo, a n-ésima amostra do sinal de voz pode ser aproximada por uma combinação das M amostras anteriores. A diferença do valor real e do valor aproximado corresponde ao erro de predição linear do sinal. A energia do erro de predição do sinal é minimizada para determinar os pesos chamados coeficientes LP

(LPCs). Dessa forma, para o sinal de voz $s(nT)$, o valor predito pela n-ésima amostra é dado por:

$$s'(nT) = \sum_{k=1}^{M} a_k s(nT - kT), \tag{4.3}$$

onde a_k são os LPC´s. Estes coeficientes a_k do filtro são calculados de maneira a minimizar a soma dos quadrados dos erros:

$$e(n) = s(nT) - s'(nT). \tag{4.4}$$

A análise LP assume este modelo para representar o efeito combinado da resposta ao impulso do sistema do trato vocal e do formato do pulso glotal.

4.2.3 IAIF

O método de filtragem inversa semi-automático é conhecido como IAIF e foi desenvolvido por [11]. Utiliza o sinal de voz como entrada a fim de obter, na saída, uma estimação do fluxo glotal correspondente. O modelo de produção da voz, o qual o IAIF é baseado, está representado na Fig. 4.3 .

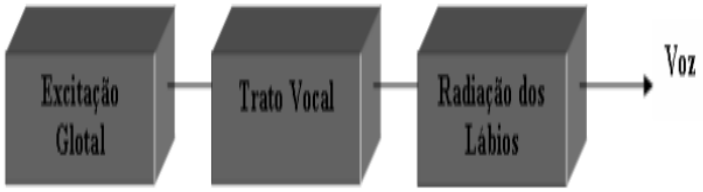

Figura 4.3: Formação do sinal glotal.

O IAIF é composto de três blocos fundamentais. São eles: análise LPC, filtragem inversa e integração. A análise LPC é responsável pela filtragem de

pré-ênfase, pela estimação do trato vocal e da contribuição glotal, definidas através da ordem de seus coeficientes e discutida, ainda, neste capítulo. A filtragem inversa é responsável pela eliminação do trato vocal e da contribuição glotal no sinal da voz. A integração é responsável pela eliminação da radiação dos lábios.

Como visto na Fig. 4.4, o sinal de entrada é passado por um filtro passa alta com intuito de eliminar as frequências baixas, que provocam flutuações na saída. O sinal filtrado é usado como entrada para os blocos subseqüentes (blocos 1, 2, 4, 7 e 9). A frequência de corte deve ser ajustada de modo que não seja maior que a frequência fundamental do sinal de voz, caso contrário perderá informações relevantes. O método IAIF é baseado no prévio conhecimento da função de transferência do trato vocal. Logo, se todo o efeito da fonte glotal é eliminado do espectro da voz, o trato vocal pode ser estimado, mais precisamente, por análise LPC ou outro método de predição linear. A estimação da contribuição glotal e a função de transferência do trato vocal é computada pelo algoritmo IAIF em uma estrutura que se repete duas vezes. Inicialmente, a primeira estimativa da contribuição glotal é obtida do sinal de voz por análise LPC de ordem um e, posteriormente, eliminada por filtragem inversa. A ordem da análise LPC, neste caso, se for maior que um pode modular os formantes, efeito indesejável por enquanto [11].

Um modelo preliminar do trato vocal é obtido aplicando análise LPC, de ordem elevada (no nosso caso, a ordem que apresento melhores resultados foi foi quarenta e cinco), ao sinal do qual o efeito da contribuição glotal inicial foi eliminado. A primeira estimativa da excitação glotal é obtida cancelando o efeito do trato vocal e da radiação dos lábios, por filtragem inversa e integração, respectivamente.

O resultado desta primeira estrutura é o sinal glotal (excitação glotal ou

contribuição glotal) que é usado como entrada da segunda estrutura a fim de estimá-lo de forma mais precisa. O espectro da excitação glotal é estimado no início da segunda estrutura usando análise LPC de ordem igual a dois ou quatro. Após cancelar a contribuição glotal, o modelo do trato vocal é encontrado, novamente usando análise LPC de ordem elevada. O resultado final é obtido pela fitragem inversa do efeito do trato vocal e da radiação dos lábios do sinal original da voz [11].

A primeira estrutura do algoritmo contém os blocos de 1 a 5 e a segunda os blocos de 6 a 10.

O processamento pelo IAIF é feito em janelas de 30ms com 75 por cento de superposição para aumentar a correlação entre janelas sucessivas, evitando variações bruscas entre características extraídas de janelas adjacentes. A Fig. 4.4 ilustra o diagrama do processamento do IAIF.

As fases de pré-ênfase e janelamento foram explicadas no capítulo de pré-processamento do sinal de voz e são aplicadas antes da entrada do sinal de voz no algoritmo IAIF.

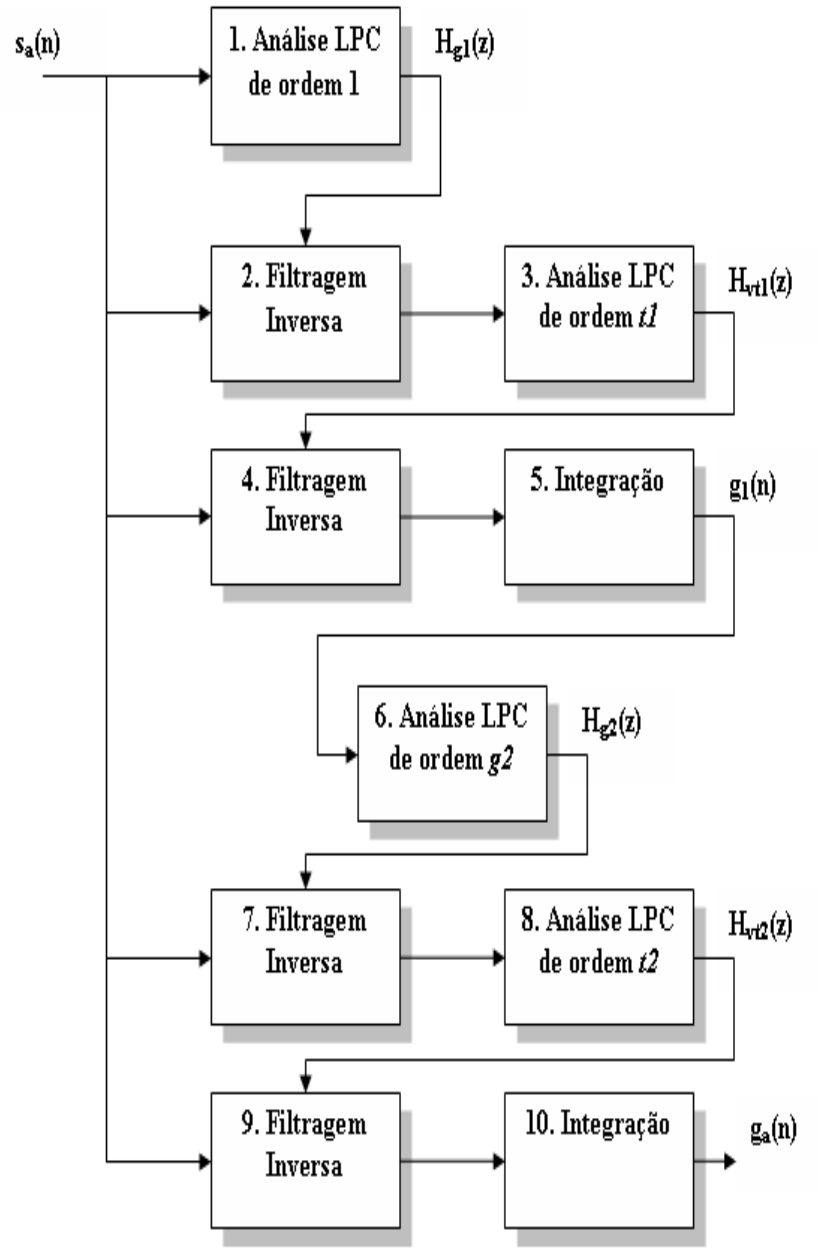

Figura 4.4: Algoritmo IAIF.

Bloco 1. Análise LPC de primeira ordem - O efeito da contribuição glotal no espectro da voz é preliminarmente estimado pela análise LPC de ordem 1. A saída deste bloco é representada pela Eq. 4.5.

$$H(z) = 1 - az^{-1}, \qquad 0,9 \leq a \leq 1,0. \tag{4.5}$$

onde o valor de a é 0,98.

Bloco 2. Filtragem Inversa - A contribuição glotal é eliminada passando $s_a(n)$ por $H_{g1}(z)$.

Bloco 3. Análise LPC de ordem t_1- a primeira estimativa do trato vocal é obtida, aplicando análise LPC à saída do bloco anterior. A saída deste bloco é dada pela Eq. (4.6) (no caso, $t_1 = 45$).

$$H_{vt1}(z) = 1 + \sum_{k=0}^{t1} b(k)z^{-k}. \tag{4.6}$$

Bloco 4. Filtragem Inversa - o efeito do trato vocal é eliminado passando $s_a(n)$ por $H_{vt1}(z)$.

Bloco 5. Integração - a primeira estimativa para a excitação glotal, $g_1(n)$, é obtida pelo cancelamento do efeito da radiação dos lábios através da integração. Este bloco marca o final da primeira estrutura usada no IAIF. Sua saída servirá de entrada para o bloco seguinte, diferentemente dos blocos 1, 2, 4, 7 e 9, que possuem o sinal de voz como entrada.

Bloco 6. Análise LPC de ordem g_2- a segunda estrutura se inicia pela nova estimação do efeito da fonte no espectro da voz, porém a análise LPC tem sua ordem alterada para dois ou quatro. O sinal no qual a contribuição glotal é estimada é $g_1(n)$. A saída deste bloco é representada pela Eq. (4.7) (no caso, $g_2 = 4$).

$$H_{g2}(z) = 1 + \sum_{k=0}^{g2} c(k) z^{-k} \,. \tag{4.7}$$

Bloco 7. Filtragem Inversa- o efeito da contribuição glotal é eliminado, passando $s_a(n)$ através de $H_{g2}(z)$. Eliminando a contribuição glotal, no espectro do sinal de voz, é possível estimar o trato vocal de forma mais precisa no próximo bloco.

Bloco 8. Análise LPC de ordem t_2 - o modelo final do trato vocal é obtido, aplicando análise LPC de ordem t_2 à saída do bloco 7. O bloco 8 tem saída representada pela Eq. (4.8). ($t_2 = 45$)

$$H_{vt2}(z) = 1 + \sum_{k=0}^{t2} d(k) z^{-k} \,. \tag{4.8}$$

Bloco 9. Filtragem Inversa - o efeito do trato vocal é eliminado da voz, passando $s_a(n)$ através de $H_{vt2}(z)$.

Bloco 10. Integração - o resultado final do algoritmo ou **sinal glotal**, $g_a(n)$, é obtido pelo cancelamento do efeito da radiação dos lábios, integrando a saída do bloco 9.

4.2.4 PSIAIF

No método IAIF, a contribuição glotal no espectro da voz é inicialmente estimada por uma estrutura iterativa. A função de transferência do trato vocal é modelada após eliminar a contribuição glotal média. A excitação glotal é obtida cancelando os efeitos do trato vocal e da radiação dos lábios, por filtragem inversa e integração, respectivamente. No método PSIAIF (*Pitch Synchronous Iterative Adaptive Inverse Filtering*), a forma do pulso glotal é obtida aplicando-se o algoritmo IAIF duas vezes, ao mesmo sinal, sendo o resultado da primeira aplicação servindo apenas para identificar o

período fundamental que será a base para o cálculo do novo janelamento, antes da segunda aplicação do IAIF. Isto é ilustrado na Fig. 4.5

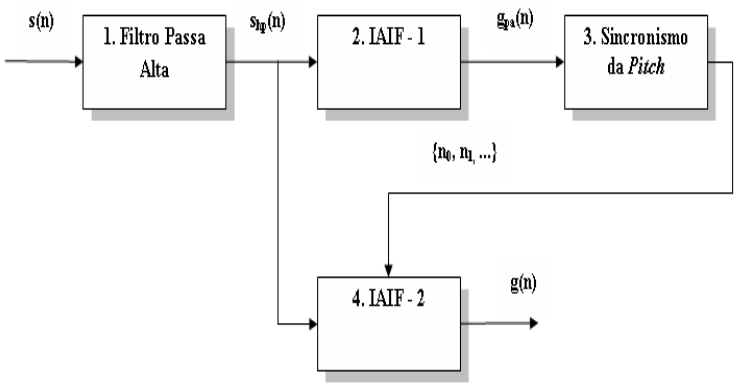

Figura 4.5: Método PSIAF.

A primeira análise realizada pelo IAIF fornece o resultado da excitação glotal que ocorre entre vários períodos da *pitch* ($g_{pa}(n)$), que tem como entrada o sinal de voz previamente filtrado ($s_{hp}(n)$ - bloco 1 da Fig. 4.5). Este pulso é usado para determinar posições e larguras de janelas para uma análise síncrona da *pitch* (frequência fundamental da voz). O resultado final será obtido analisando o sinal de voz original com o algoritmo IAIF em um período por vez, ou seja, a estimativa final da forma do pulso glotal será obtida aplicando o método IAIF ao sinal de voz, usando o intervalo de tempo entre dois máximos de abertura glotal consecutivos (n_0, n_1, ...) [11]. Outros tamanhos de janela podem ser utilizados, mas sempre tendo como referência o período fundamental. Neste trabalho foram usados três períodos fundamentais consecutivos. A principal vantagem na utilização do método PSIAIF está na obtenção do sinal glotal de forma mais precisa.

4.2.5 Modelo Discreto Só Polo(DAP)

DAP é uma alternativa à análise LPC. A idéia básica do modelo DAP é adaptar o modelo tudo-polo utilizado pela análise LPC usando unicamente o conjunto finito das localizações no espectro, relacionadas com a posição dos harmônicos da freqüência fundamental.

O modelo DAP é otimizado, tentando diminuir a distância entre o espectro original e a aproximação deste espectro, por meio de uma versão discreta da distância Itakura-Saito [43]. A distância é medida por:

$$D_{dap} = \sum_{m} \frac{S(w_m)}{\hat{S}(w_m)} - log(\frac{S(w_m)}{\hat{S}(w_m)}) - 1. \tag{4.9}$$

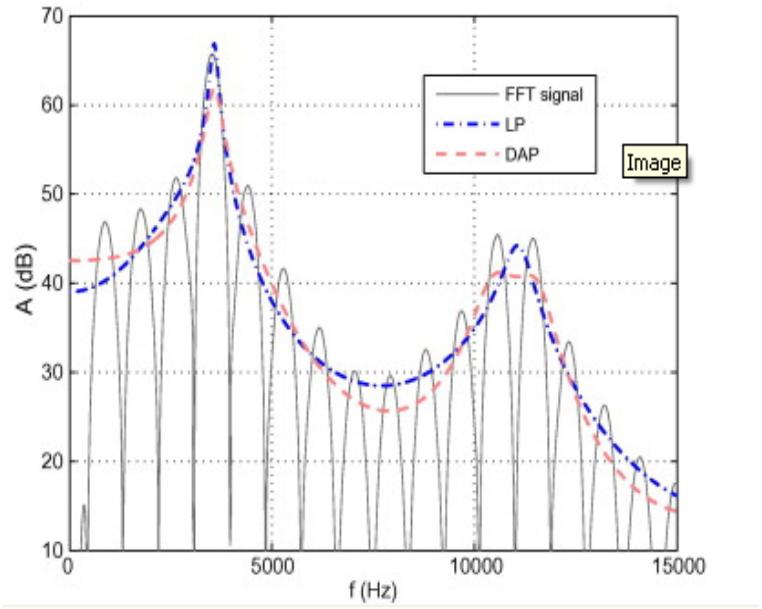

Figura 4.6: Estmações LP e DAP para uma sinal com periodo fundamental=50 e Fs=44100

Onde w_m são os harmônicos da frequência fundamental, Sw_m é o espectro

do sinal e $\hat{S}(wm)$ é a estimativa do espectro.

Esta distância é uma minimização adaptativa da estimação do modelo DAP. Atualmente o IAIF é implementado usando o modelo DAP por ser capaz de estimar o trato vocal com mais precisão que a análise LPC [31]. Na Fig 4.6 mostra-se uma comparaçaõ entre LPC e DAP.

4.3 Parâmetros do sinal glotal

Para discriminar o sinal glotal, extraem-se parâmetros desse sinal, aproveitando sua periodicidade, em vogais. Os parâmetros que descrevem o fluxo glotal podem ser usados em múltiplas aplicações, tais como: pesquisas sobre a produção da voz, codificação, síntese, reconhecimento automático de voz, uso clínico, verificação e identificação automática de locutor e para quantificar a contribuição do pulso glotal na transmissão de sentimentos.

4.3.1 Instantes de máxima abertura e máximo fechamento glotal

O instante de máximo fechamento é definido como o instante em que o fluxo glotal atinge seu valor mínimo. Fisiologicamente, corresponde ao instante que as cordas vocais começam a se separar. O instante de máxima abertura está associado ao máximo da excitação glotal, em outras palavras, corresponde ao instante que o fluxo glotal atinge seu valor máximo.

4.3.2 Duração fase de fechamento (Ko)

É a fase em que as cordas vocais são separadas e a área de abertura entre elas está diminuindo. A duração da fase de fechamento é indicada por Ko [11] [15].

4.3.3 Duração fase de abertura (Ka)

É a fase em que as cordas vocais estão, pelo menos parcialmente separadas e a área de abertura entre elas está aumentando. A duração da fase de abertura é denotada por Ka.

4.3.4 Periodo Fundamental (T)

É o tempo entre os ciclos de abertura e fechamento da glotis . Isto é indicado por T, e é o recíproco do frequência fundamental fo.

4.3.5 Amplitude de vozeamento (Av)

A amplitude de vozeamento (Av) é definida como a amplitude entre os valores mínimos e máximos do sinal glotal como é ilustrado na Fig.4.7. Este parâmetro na prática não teve êxito para discriminar o sinal glotal já que depende de vários fatores que entrariam como ruído na classificação, tais como a intensidade de voz em cada gravação e a distância entre o locutor e o microfone na hora da gravação.

4.3.6 Distância entre os instantes de máxima abertura glotal (pp)

Após a obtenção dos instantes de máxima abertura do sinal glotal, toma-se o valor da distância entre os máximos (pico-pico) do sinal glotal como se ilustra na Fig.4.7. Este parâmetro será definido como (pp) e é um aporte deste trabalho, somando-se aos parâmetros já descritos por [11][15].

4.3.7 Fase de abertura (Fa)

É a parte do ciclo glótico durante o qual as cordas vocais são separadas e passa o fluxo de ar através da glote. Onde $Fa = Ka + Ko$ [31].

4.3.8 Quociente de abertura (OQ)

É definido como a relação entre a fase de abertura e o comprimento total do ciclo glótico [31]. Onde $OQ = Fa/T$

4.3.9 Quociente de fechamento (CIQ)

É definido como a relação entre a fase de fechamento e um ciclo glotal completo. Onde $CIQ = Ko/T$ [31].

4.3.10 Quociente de velocidade (SQ)

É definido como a relação entre a fase de abertura e a fase de fechamento. Onde $SQ = Ka/Ko$ [31].

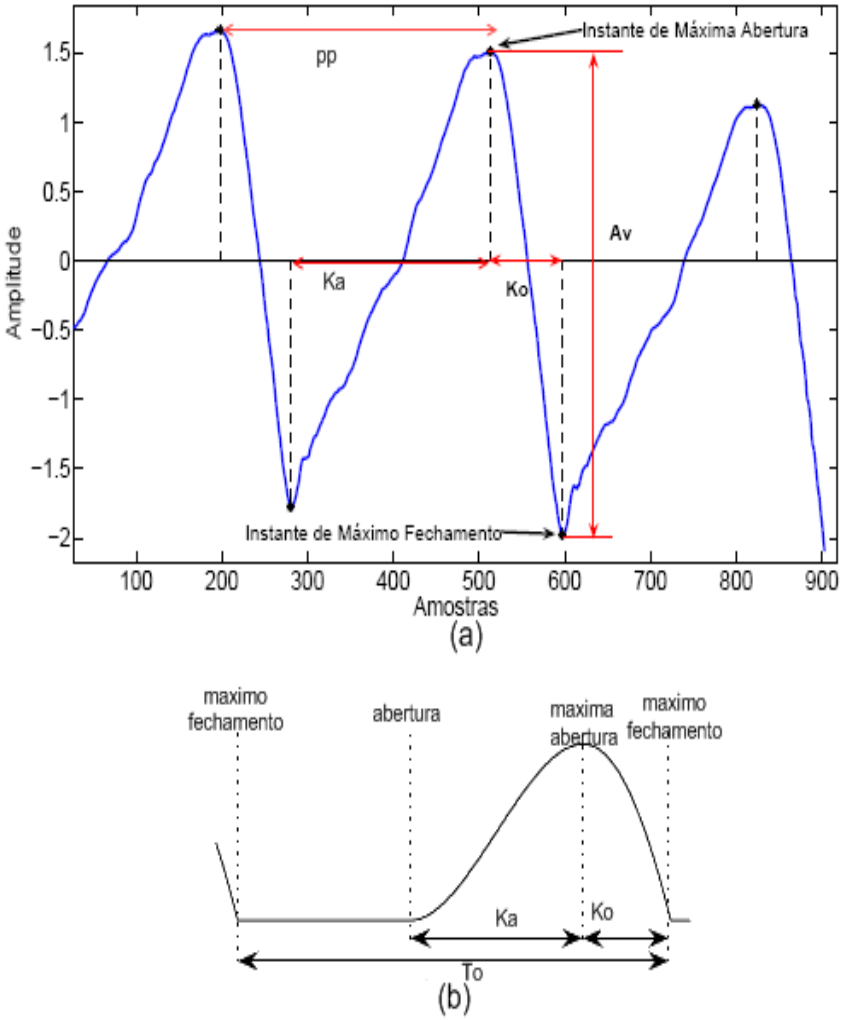

Figura 4.7: (a) Parâmetros do sinal glotal (b) Sinal glotal proposta [11].

Capítulo 5

Classificadores de padrões

5.1 Redes Neurais

O neurônio, a célula nervosa do sistema neural humano, é composto de três partes: o corpo, o axônio e os dendritos. Os dendritos têm por função receber as informações oriundas de outros nós, e conduzí-las até o corpo celular. Na parte do corpo, a informação é processada e novos impulsos são gerados. Estes impulsos são transmitidos a outros nós, passando através do axônio até os dendritos dos nós seguintes. O ponto de contato entre a terminação axônica de um neurônio e o dendrito de outro é chamado sinapse. É pela sinapse que os nós se unem funcionalmente, formando as redes neurais.

As redes neurais artificiais foram criadas para imitar, em um computador, a estrutura e a funcionalidade do cérebro. Dessa forma, os neurônios passam a ser representados como simples elementos de processamento, os dendritos como interconexões, as sinapses como pesos e o axônio pelos terminais de saída [14].

O processo de combinação dos sinais e geração de uma saída para o neurônio é modelado por uma função de transferência, as sinapses de cada

conexão são representadas por pesos que variam durante o treinamento.

As redes neurais têm certas propriedades às quais fazem com que seja uma das ferramentas mais estudadas na atualidade para classificação de padrões [32]. Algumas propriedades devem ser destacadas:

- Não-Linearidade : As redes neurais podem operar funções não lineares, habilitando-se assim em desenvolver funções complexas de transformação de dados.

- Adaptabilidade : A rede neural tem a capacidade de adaptar seus pesos de acordo com as variações do ambiente em que se encontra. Em particular, uma rede neural treinada para operar em um específico ambiente pode ser facilmente retreinada, com poucas modificações, para operar em condições ambientais diferentes.

- Robustez : As redes são tolerantes a falhas e dados ruidosos.

- Generalização : As redes não apenas memorizam os dados treinados, mas também podem generalizar para novos padrões. Isso é essencial no reconhecimento da voz, porque os padrões acústicos nunca são exatamente os mesmos.

- Paralelismo - As redes neurais são altamente paralelas por natureza, dessa forma são ideais para implementação em computadores de processamento paralelo, permitindo um rápido processamento.

5.1.1 Unidades de processamento

Um neurônio é uma unidade de processamento de informação que é fundamental para a operação de uma rede neural. A Fig. 5.1 mostra o modelo para um neurônio. Podem-se identificar três elementos básicos do modelo:

-Um conjunto de conexões ou sinapses cada uma das quais caracterizada

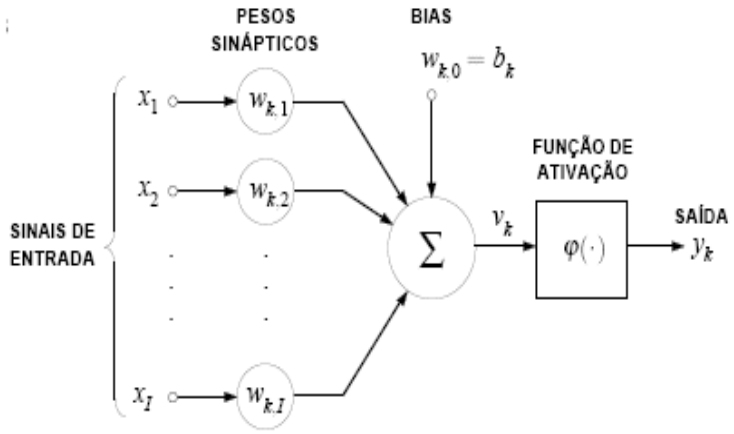

Figura 5.1: Modelo não linear de um neurônio.

por seu peso. Por exemplo, o sinal x_j, na entrada da sinapse j, conectado ao neurônio k é multiplicado por um peso $w_{k,j}$, onde k refere-se ao neurôinio em questão e o j á sinapse pela qual o peso refere-se.

-Uma função de ativação $\varphi(.)$ para limitar a amplitude do sinal de saída da unidade de processamento y_k e medir o estado de ativação para o neurônio.

-Uma função de propagação que se encarrega de propagar o estado de ativação do neurônio para os outros que estão conectados ao mesmo.

O neurônio se descreve nas Eq 5.1 5.2:

$$v_k = \sum_{j=1}^{p} w_{kj} x_j \tag{5.1}$$

$$y_k = \varphi(v_k - \Theta_k) \tag{5.2}$$

onde $x_1, x_2, ..., x_p$ são os sinais de entrada que representam os dendritos; $w_{k1}, w_{k2}, ..., w_{kp}$ são os pesos sinápticos do neurônio k que representam as sinapses; v_k é o estado de ativação do j-ésimo neurônio; Θ_k é o limiar; $\varphi(.)$ é a função de ativação; e y_k é a sinal de saida do neurônio k.

5.1.2 Funções de ativação

A função de ativação define a saída de um neurônio em termos do nível de ativação da sua entrada. As funções de ativação mais utilizadas são [14]:

-Degrau simetrico

$$y = \begin{cases} 1, & x \geq b \\ -1, & x \leq b \end{cases} \tag{5.3}$$

-Linear

$$y = x + b \tag{5.4}$$

-Logística Sigmoidal

$$y = 1/(1 + e^{-(n+b)}) \tag{5.5}$$

-Tangente Sigmoidal

$$y = (e^{(x+b)} - e^{-(x-b)})/(e^{(x+b)} + e^{-(x-b)}) \tag{5.6}$$

5.1.3 Arquitetura de redes neurais

Por arquitetura de uma rede neural entende-se a forma como estão conectadas suas unidades de processamento e como ocorre o fluxo do sinal dentro da rede. A arquitetura da rede neural está fortemente ligada ao algoritmo de aprendizado usado para treiná-la. As redes neurais classificam-se por seu numero de camadas e pelo tipo de conexão entre os nós, as mais comuns são explicadas a seguir:

Redes de camada única

Só existem os nós fontes da camada de entrada e qualquer saída da rede. Deve ser observado que a camada de entrada não deve ser considerada, pois nenhuma computação nela é realizada.

Redes de múltipla camada

Diferenciam-se das redes de uma camada pela presença de uma ou mais camadas escondidas (*hidden layers*).

Feedforward

Redes de uma ou mais camadas de processadores, cujo fluxo de dados é sempre em uma única direção, isto é, não existe realimentação.

Recorrentes

Redes com conexões entre processadores da mesma camada e/ou com processadores das camadas anteriores (realimentação).

5.1.4 Aprendizado nas redes neurais

Aprendizado é um processo pelo qual os parâmetros livres de uma rede neural são adaptados, através de estímulos do ambiente onde está a rede neural. O tipo de aprendizado é determinado pela maneira pela qual as mudanças nos parâmetros acontecem. Este processo é formalizado através de um algoritmo de aprendizado que define como a rede é estimulada, como os parâmetros se adaptam e como a rede deve responder a novos estímulos. Apresentamos a seguir as classes de aprendizado:

Aprendizado Supervisionado

A rede é treinada através do fornecimento dos valores de entrada e de seus respectivos valores desejados de saída; geralmente efetuado através do processo de minimização do erro calculado na saída.

O conhecimento está disponível para a rede sob a forma de exemplos de pares $t(t)$ do tipo vetor de entrada $x(t)$ e seu respectivo vetor de saida desejada d(t) reunidos em um conjunto de treinamento Γ como vemos na Eq. 5.7:

$$\Gamma = \{t(t)\} = \{(x(t), d(t))\}, 1 \leq t \leq T, \tag{5.7}$$

onde T é o número de elementos do conjunto de treinamento. Costuma-se atribuir um significado temporal ao índice do elemento de treinamento, pois estes elementos são apresentados seqüencialmente à rede neural para o seu treinamento. Neste contexto, uma apresentação do conjunto completo é chamada de época de treinamento, e T é a duração da época de treinamento. Quando um exemplo é apresentado á rede neural, é calculado um sinal de erro e(t), no qual à rede se baseia para tentar apresentar uma resposta mais correta na proxima vez que o exemplo for apresentado: $e(t) = d(t) - y(t)$

Desta forma, pode-se entender que a rede aprende a imitar o seu conjunto de treinamento (ambiente). Na Fig.5.2 mostra-se um diagrama para entender melhor o processo de aprendizagem supervisionado.

Aprendizado Não Supervisionado

No aprendizado não-supervisionado, utilizado em sistemas de classificação, não existe saída desejada. A rede é treinada através de excitações ou padrões de entrada e então, arbitrariamente, organiza os padrões em categorias. Para

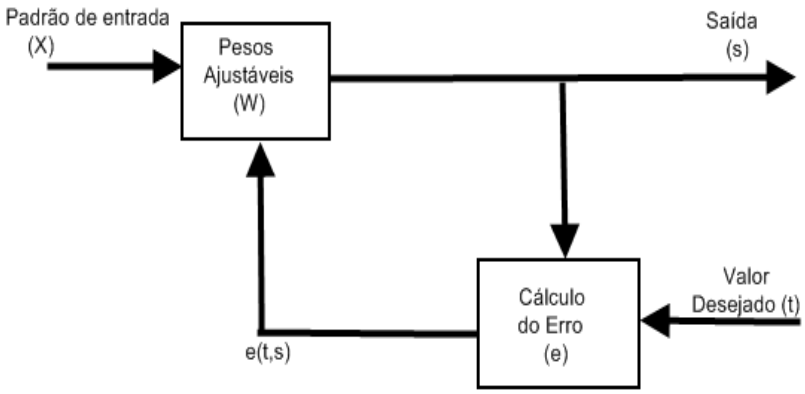

Figura 5.2: Aprendizado supervisionado

uma entrada aplicada à rede, será fornecida uma resposta indicando a classe a qual a entrada pertence. Se o padrão de entrada não corresponde às classes existentes, uma nova classe é gerada.

Neste caso, não há vetores de saídas desejadas $d(t)$ associados aos vetores de entradas $x(t)$.

$$\Gamma = \{t(t)\} = \{(x(t))\}, 1 \leq t \leq T \tag{5.8}$$

O sinal de erro gerado no aprendizado supervisionado é substituído por uma medida independente da tarefa que a rede deve aprender, e os parâmetros livres são adaptados para minimizar este medidor. Para isto pode ser usada uma regra de aprendizado competitivo. Uma vez que a rede tenha sido sintonizada às regularidades estatísticas dos dados de entrada, ela desenvolve a habilidade de formar representações internas para a codificação dos atributos da entrada e criar novas classes automaticamente.

O aprendizado não supervisionado (auto-organizado) baseia-se em modificar repetidamente os pesos sinápticos de uma rede neural em resposta aos

padrões de ativação, de acordo com regras predeterminadas, até que uma determinada configuração final seja atendida Na Fig.5.3 se mostra um diagrama para entender melhor o processo do aprendizagem supervisionado.

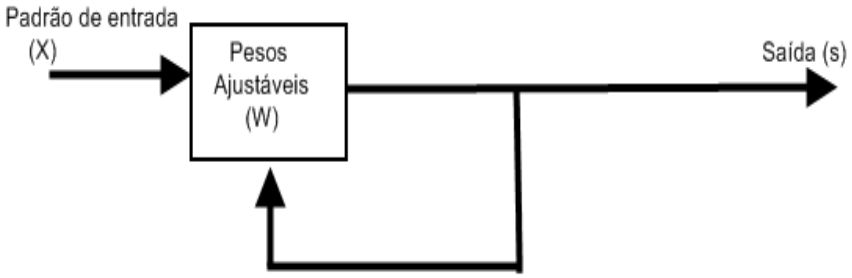

Figura 5.3: Aprendizado não supervisionado

5.1.5 Redes *Multilayer Perceptron*

As redes *multilayer perceptrons* (MLPs) têm sido aplicadas com sucesso em diversas áreas, desempenhando tarefas tais como classificação de padrões (reconhecimento), controle e processamento de sinais.

Uma rede neural artificial (RNA) do tipo MLP é constituída por um conjunto de nós fonte, os quais formam a camada de entrada da rede (*input layer*), uma ou mais camadas escondidas (*hidden layer*) e uma camada de saída (*output layer*). Com exceção da camada de entrada, todas as outras camadas são constituídas por neurônios e, portanto, apresentam capacidade computacional [14].

Em uma rede multi-camada, o processamento realizado por cada nó é definido pela combinação dos processamentos realizados pelos nós da camada anterior. Quando se segue da primeira camada intermediária em direção

à camada de saída, as funções implementadas tornam-se mais complexas. Estas funções definem como será realizada a divisão do espaço. Um exemplo é mostrado na Fig 5.4

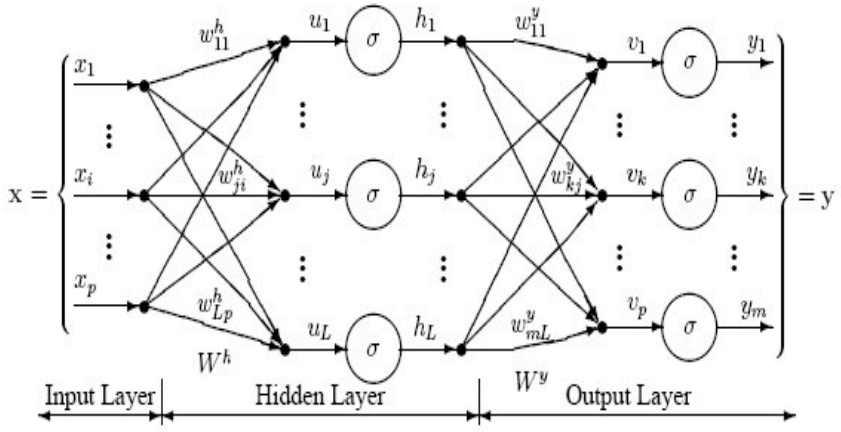

Figura 5.4: arquitetura *Multilayer Perceptron.*

Uma rede MLP apresenta três características distintas, de cuja combinação com a habilidade de aprender através da experiência (através do treinamento), deriva sua capacidade computacional:

- O modelo de cada neurônio do MLP inclui uma função de ativação não linear e diferenciável. Uma função comumente utilizada é a sigmoidal definida pela função logística mostrada na Eq. 5.9:

$$y_j = 1/(1 + exp(-v_j)) \tag{5.9}$$

Onde v_j é o potencial de ativação (isto é, a soma ponderada de todas as entradas sinapticas mais a polarização) do neuronio j, e y_j é a saída do neurônio.

- O MLP contém uma ou mais camadas de neurônios escondidos que

não são parte da camada de entrada ou da camada de saída da rede. Estes neurônios escondidos possibilitam que a rede aprenda tarefas complexas, extraindo progressivamente mais características significativas dos padrões de entrada (vetores de entrada).

-A rede MLP exibe um alto grau de conectividade, determinado pelas sinapses da rede. Uma mudança na conectividade da rede requer uma mudança na população das conexões sinápticas, ou pesos sinápticos.

5.1.6 Implementação de uma rede MLP

Em uma rede MLP o número de nós fonte na camada de entrada da rede é determinado pela dimensionalidade de espaço de observação, que é responsável pela geração dos sinais de entrada. O número de neurônios na camada de saída é determinado pela dimensionalidade requerida da resposta desejada. Assim o projeto de uma rede MLP requer a consideração de três aspectos:

-A determinação do número de camadas escondidas.

-A determinação do número de neurônios em cada uma das camadas escondidas.

-A especificação dos pesos sinápticos que interconectam os neurônios nas diferentes camadas de rede.

Os dois primeiros aspectos determinam a complexidade do modelo de RNA escolhido e infelizmente não há regras determinadas para tal especificação. A função das camadas escondidas em uma RNA é a de influir na relação entrada-saída da rede de uma forma ampla.

Uma RNA com uma ou mais camadas escondidas é apta para extrair as características de ordem superior de algum processo aleatório subjacente, responsável pelo comportamento dos dados de entrada, processo sobre o qual

a rede está tentando adquirir conhecimento.

A utilização de duas ou mais camadas escondidas pode facilitar o treinamento da rede, entretanto esta técnica não é recomendada, pois, a cada vez que o erro medido durante o treinamento é propagado para a camada anterior, ele se torna menos preciso.

O número de nós na camada intermediária depende de vários fatores como o número de exemplos de treinamento, a quantidade de ruído presente nos exemplos, a complexidade da função obtida e a distribuição estatística dos dados de treinamento.

Deve-se ter cuidado para não utilizar unidades intermediárias demais, o que pode levar a rede a memorizar os padrões de treinamento, ao invés de extrair as características gerais que permitirão a generalização ou reconhecimento de padrões não vistos durante o treinamento, nem um número pequeno, que pode forçar a rede a gastar tempo em excesso tentando encontrar uma representação ótima.

As RNAs MLPs têm sido aplicadas na solução de diversos e difíceis problemas através da utilização de tais algoritmos. O algoritmo de treino quase universalmente utilizado é o algoritmo de retro-propagação de erro, conhecido na literatura como Backpropagation Algorithm.

A operação da rede neural constitui de três etapas a primeira é de treinamento, que consiste no ajuste dos parâmetros do modelo. A de teste que é a validação dos parâmetros de esse modelo e, por último, a de produção que é a utilização do modelo.

5.1.7 Algoritmo de *Backpropagation*

O algoritmo de aprendizado mais conhecido para treinamento das redes MLP's é o algoritmo *Backpropagation*. Este algoritmo é supervisionado e seu treinamento ocorre em duas fases, onde cada fase percorre a rede em um sentido. Essas duas fases são chamadas de: *Forward* e *Backward*. A fase *Forward* é utilizada para definir a saída da rede para um dado padrão de entrada, nenhuma alteração nos pesos é feita. Na fase *Backward* são utilizadas a saída desejada e a fornecida pela rede para que os pesos sejam atualizados como mostra a Fig.5.5.

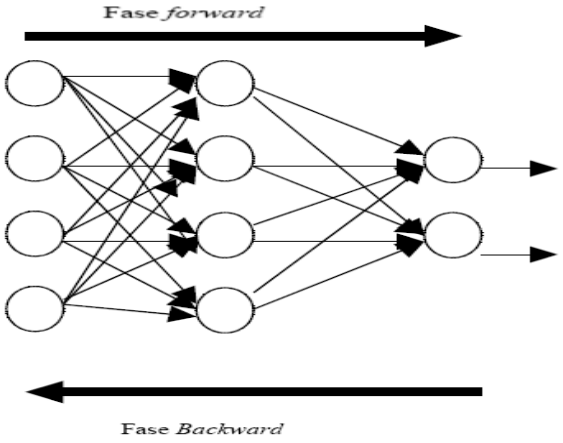

Figura 5.5: *Fluxo do processamento do algoritmo Back-propagation.*

Este procedimento de aprendizado é repetido diversas vezes, até que, para todos os processadores da camada de saída e para todos os padrões de treinamento, o erro seja menor do que o especificado. O aprendizado pode ser dividido em dois: *Batch* e incremental. O aprendizado por *Batch* somente ajusta os pesos após a apresentação de todos os padrões, cada padrão é avaliado com a mesma configuração de pesos. O aprendizado incremental atualiza os pesos a cada apresentação de um novo padrão e os pesos são

atualizados usando o gradiente do erro de um único padrão [14].

O desenvolvimento do *backpropagation* representa um marco fundamental em redes neurais, pois é um método computacionalmente eficiente para o treinamento de redes MLPs e por ter resolvido o problema de realizar a propagação reversa do erro em RNA com múltiplas camadas, problema este que atrasou por muitos anos o desenvolvimento da área das redes neurais.

5.1.8 Parâmetros utilizados no treinamento

Momento

O momento é introduzido no treinamento com o objetivo de acelerar o aprendizado sem causar oscilação. Possibilita a rede ignorar as variações de alta frequência na superfície de erro, diminuindo a probabilidade do processo de convergência parar em um mínimo local. A introdução do momento consiste em fazer com que as mudanças nos pesos das conexões sejam iguais à soma de uma fração da última alteração nestes pesos com a nova alteração determinada pela regra de aprendizagem.

Taxa de aprendizagem

A taxa de aprendizagem influencia a magnitude dos pesos, isto é, uma pequena taxa de aprendizagem implica em pequenas variações, tornando o treinamento lento e aumentando a probabilidade de paradas em mínimo local. Entretanto, ao utilizar altas taxas de treinamento a rede neural poderá saturar ou até mesmo oscilar. Uma alternativa é utilizar a taxa de treinamento adaptativa, isto é, quando o erro aumentar, o valor da taxa de aprendizagem diminuirá rapidamente; por outro lado, se o erro diminuir a taxa de

aprendizagem aumentará lentamente.

Correção do Erro no Treinamento

Seja $d_k(n)$ a resposta desejada para o neurônio k no tempo n e $y_k(n)$ a resposta atual desse neurônio produzida pelo estimulo x(n), aplicado na entrada da rede na qual o neuronio k está localizado. Pode-se definir o sinal de erro como a diferença entre a resposta desejada e a atual. Isto é mostrado na Eq 5.10:

$$e_k(n) = d_k(n) - y_k(n) \tag{5.10}$$

O propósito principal da correção do erro no treinamento é minimizar a função custo baseada no sinal de erro, $e_k(n)$, tal que a resposta atual de cada saida aproxime-se da resposta desejada para aquele neurônio. Um criterio comumente utilizado para a função custo é o critério do erro médio quadrático, definido como o valor medio quadratico da soma dos erros quadráticos, como se mostra na Eq 5.11:

$$E = (1/2) \sum_{p} \sum_{i=1}^{k} (d_i^p - y_i^p)^2 \tag{5.11}$$

onde E é a medida do erro total, p é o número total de padrões, k é o número de unidades de saída, d_i é i-ésima saída desejada e y_i é a i-ésima saída gerada pela rede. O fator 1/2 é utilizado para simplificar os cálculos em possíveis derivações que são resultantes de minimizações de E com respeito aos parâmetros livres da rede.

Derivação das fórmulas do algoritmo Backpropagation

O algoritma Back propagation estabelece o aprendizado de um MLP através da regra delta como sendo a correção efetuada em suas sinapses através de 5.12

$$\Delta w_{ji} = \eta \delta_j x_i, \tag{5.12}$$

onde Δw_{ji} é a correção aplicada á M-ésima sinapses do neuronio j, x é o sinal de entrada do neuronio i e $\delta(n)$ é o gradiente local do neuronio j.

Embora o erro total E seja definido pela soma dos nós de saída para todos os padrões, supõe-se, sem perda de generalidade, que a minimização do erro para cada padrão individualmente levará a minimização do erro total. Assim o erro passa a ser definido pela Eq 5.13.

$$E = \frac{1}{2} \sum_{j=1}^{k} (d_j - y_j)^2 \tag{5.13}$$

O simbolo w_{ji} denota o peso sinaptico que conecta a saida do neuronio i á entrada do neuronio j na iteração n. A correção aplicada a este peso em uma iteraçãoé denotada por Δw_{ji}. A regra delta sugere que a variação dos pesos seja definida de acordo com o gradiente descendente, isto é, de acordo com a Eq. 5.13.

$$\Delta w_{ji} = \alpha - \frac{\partial E}{\partial w_{ji}} \tag{5.14}$$

Utilizando a regra da cadeia, tem-se que:

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial net_j} \frac{\partial net_j}{\partial w_{ji}} \tag{5.15}$$

Como

$$net_j = \sum_{i=1}^{n} x_i w_{ji}. \tag{5.16}$$

Então a segunda derivada de Eq. 5.17

$$\frac{\partial net_j}{\partial w_{ji}} \tag{5.17}$$

É igual

$$\frac{\partial net_j}{\partial w_{ji}} = \frac{\partial \sum_{i=1}^{n} x_i w_{ji}}{\partial w_{jl}} = x_i \tag{5.18}$$

A primeira derivada localizada à direita da Eq. 5.15 mede o erro no nó j e o cálculo desta derivada também pode ser definida pela regra da cadeia:

$$\partial_f = \frac{\partial E}{\partial net_j} = \frac{\partial E}{\partial y_j}\frac{\partial y_j}{\partial net_j} \tag{5.19}$$

A segunda derivada da Eq. 5.19 é definida como:

$$\frac{\partial y_j}{\partial net_j} = \frac{\partial f(net_j)}{\partial net_j} = f'(net_j) \tag{5.20}$$

Já a primeira derivada vai depender da camada onde o nó j se encontra. Se o nó j estiver na última camada, o seu erro pode ser definido utilizando-se a Eq. 5.13.

$$\frac{\partial E}{\partial y_j} = \frac{\partial [\frac{1}{2}\sum_{j=1}^{k}(d_j - y_j)^2]}{\partial y_f} = -(d_j - y_j) \tag{5.21}$$

substituindo as Eq. 5.21 e 5.20 em 5.19 tem-se que:

$$\partial_j = -(\partial_j - y_j)f'(net_j) \tag{5.22}$$

Substituindo as Eq. 5.22 e 5.18 em 5.14 tem-se que

$$\Delta w_{ij} = x_i(\partial_j - y_j)f'(net_j) \tag{5.23}$$

Se o nó j não estiver na camada de saída tem-se que:

$$\frac{\partial E}{\partial y_j} = \sum_{i=1}^{M} \frac{\partial E}{\partial net_i} \frac{\partial net_i}{\partial y_j} = \sum_{i=1}^{M} \frac{\partial E}{\partial net_i} \frac{\partial \sum_{i=1}^{n} x_i w_{ji}}{\partial y_j} \sum_{i=1}^{M} \frac{\partial E}{\partial net_i} w_{jl} \tag{5.24}$$

$$\sum_{i=1}^{M} \frac{\partial E}{\partial net_j} w_{jl} = \sum_{i=1}^{M} \delta_i w_{ji} \tag{5.25}$$

Substituindo as Equações 5.24 e 5.20 em 5.19:

$$\partial_j = f'(net_j) \sum_i \delta_l w_{lj} \tag{5.26}$$

Pode-se então generalizar a Eq. 5.13 para

$$\Delta w_{ji} = \eta \delta_j x_i \tag{5.27}$$

ou

$$w_{ji}(t+1) = \alpha w_{ji}(t) + \eta \delta_j(t) x_i(t) \tag{5.28}$$

α é o momento e pode ir de $0 < \alpha < 1$ e η a taxa de aprendizagem. O processo de aprendizado pode ser entendido como uma combinação de pesos e limiares que irão corresponder a um ponto na superfície de erro. Considerando que a altura de um ponto é diretamente proporcional ao erro associado a este ponto, a solução está nos pontos mais baixos da superfície.

Critérios para parar o treinamento

Uma dúvida que surge naturalmente diz respeito a quando parar o treinamento da rede. Existem vários métodos para a determinação do momento onde o treinamento deve ser encerrado, entre eles pode-se citar:

-Encerrar o treinamento após M ciclos.

-Encerrar o treinamento após o erro quadrático médio ficar abaixo de uma constante.

-Encerrar o treinamento quando a porcentagem de classificações corretas estiver acima de uma constante.

-Encerrar o treinamento quando o erro médio quadrático não diminuir durante N ciclos.

-Combinação dos métodos acima.

5.2 Máquina de vetores de suporte(SVM)

A Máquina de Vetores de Suporte (SVMs, do Inglês *Support Vector Machine*s) constitui uma técnica de aprendizado que vem recebendo crescente atenção nos últimos anos [40]. Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as Redes Neurais Artificiais (RNAs). SVMs são baseadas na Teoria de Aprendizado Estatístico, desenvolvida por [40]. Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. Seu aprendizado é efetuado através do princípio de Minimização de erro estrutural, que demonstrou ser superior ao tradicional Minimização de Erro [39], princípio empregado pelas redes neurais convencionais.

5.2.1 A Teoria de Aprendizado Estatístico(TAE)

A Teoria de Aprendizado Estatístico estabelece condições matemáticas que auxiliam na escolha de um classificador particular $\hat{f}$ a partir de um conjunto de dados de treinamento. Essas condições levam em conta o desempenho do classificador no conjunto de treinamento e a sua complexidade, com o objetivo de obter um bom desempenho também para novos dados do mesmo domínio [40].

Seja f um classificador e F o conjunto de todos os classificadores que um determinado algoritmo de aprendizado de máquina (AM) pode gerar. Esse algoritmo, durante o processo de aprendizado, utiliza um conjunto de treinamento T, composto de n pares (x_i, y_i), para gerar um classificador particular $\hat{f}\epsilon F$.

Consideremos o seguinte exemplo baseado em [39]: encontrar um classificador que separe os dados das classes "círculo"e "triángulo"ilustradas na Fig 5.6. As funções ou hipóteses consideradas são ilustradas por meio das bordas, também denominadas fronteiras de decisão traçadas entre as classes.

Na Fig 5.6(a), mostramos uma hipótese que classifica corretamente todos os exemplos do conjunto de treinamento, incluindo dois possíveis ruídos. Por ser muito específica para o conjunto de treinamento, essa função apresenta elevada susceptibilidade a cometer erros quando confrontada com novos dados. Esse caso representa a ocorrência de um super ajustamento do modelo aos dados de treinamento.

Na Fig 5.6(c) temos outro classificador que comete muitos erros mesmo para casos que podem ser considerados simples. Tem-se assim a ocorrência de um sub-ajustamento, pois o classificador não é capaz de se ajustar mesmo aos exemplos de treinamento.

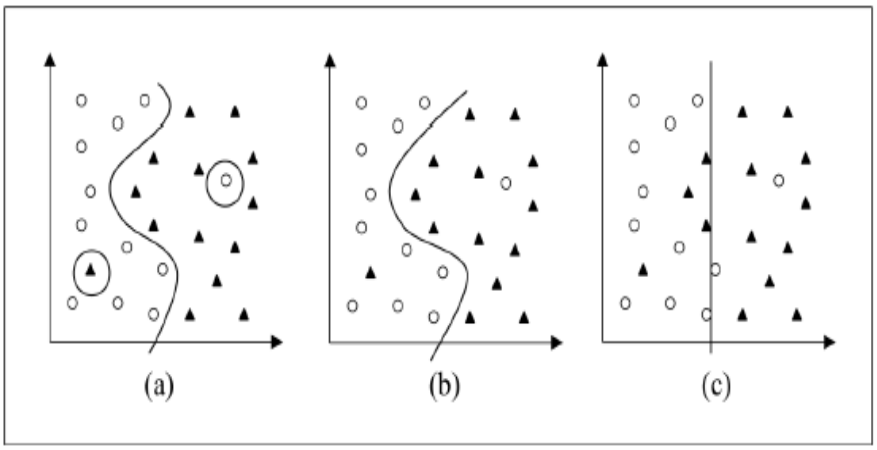

Figura 5.6: Diferentes hipóteses de configuração de treinamento.

A Fig 5.6(b) é um meio termo entre as duas funções descritas e é um classificador com complexidade intermediária e classifica corretamente grande parte dos dados, sem se fixar demasiadamente em qualquer ponto individual.

Na aplicação da TAE, assume-se inicialmente que os dados do domínio em que o aprendizado está ocorrendo são gerados de forma independente e identicamente distribuída de acordo com uma distribuição de probabilidade $f(x, y)$, que descreve a relação entre os dados (x) e os seus rótulos (y). O erro (risco) esperado de um classificador f denotado por $R(f)$, para dados de teste, pode então ser quantificado pela Eq. 5.29.

$$R(f) = \int_{\Omega} C(y, f(x)) f_{XY}(x, y) dxdy \tag{5.29}$$

O erro esperado mede então a capacidade de generalização de f. Na Eq. 5.29, $C(y, f(x))$ é uma função de custo, relacionando a previsão $f(x)$ quando a saída desejada é y.

Como é apresentado o erro esperado na Eq. 5.29 não é possivel minimizálo e $f_{XY}(x, y)$ é desconhecida. Normalmente, infere-se uma função $\hat{f}$

que minimize o erro sobre esses dados e espera-se que esse procedimento leve também a um menor erro sobre os dados de teste. Para tentar minimizar o erro é inserido o princípio de minimização de erro empírico [41] denotado por $R_{emp}(f)$ para a função f que é :

$$R_{emp}(f) = \frac{1}{n}\sum_{i=1}^{n} c(f(x_i), y_i). \tag{5.30}$$

A relação do erro esperado do classificador e o princípio do erro empírico é:

$$R(f) \leq R_{emp}(f) + \sqrt{\frac{h(ln(2n/h)+1) - ln(\Theta/4)}{n}}. \tag{5.31}$$

Esta relação é chamada de minimização de erro estrutural [40] que é a função de minimização de erro usada por SVM. Isto é, minimiza um limite superior sobre o erro esperado, contrário às redes neurais que minimizam o erro sobre a formação dos dados [41] . Um limite importante fornecido pela TAE relaciona o erro esperado de uma função ao seu erro empírico. Esse limite, apresentado na Eq. 5.31, é garantido com probabilidade $1 - \Theta$ em que $\Theta\epsilon[0, 1]$. O termo n representa a quantidade de exemplos no conjunto de treinamento.

5.2.2 SVMs Lineares

O objetivo da SVM Linear é separar duas classes por uma função que é induzida a partir dos exemplos disponíveis para aprendizagem. A meta é produzir um classificador que funciona bem em todos os exemplos, ou seja, que tenha boa generalização. Considere o exemplo na Fig 5.7

Aqui existem vários classificadores lineares possíveis que podem separar os dados, mas há apenas um que maximiza a margem de separação (maximiza a distância entre o classificador e os pontos mais próximos dos dados de cada classe). Este classificador linear é denominado de hiperplano ótimo de

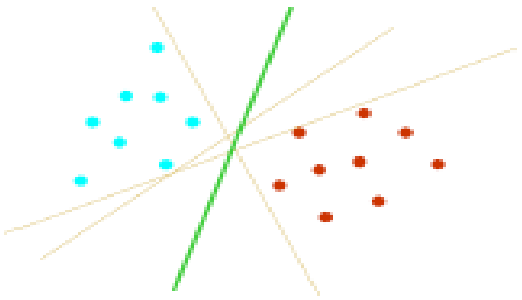

Figura 5.7: Hiperplano Ótimo de Separação.

separação. Os vetores de suporte são aqueles que servem de referência para encontrar a maior margem de separação para obter o hiperplano ótimo de separação.

Na Fig 5.8, mostra-se o hiperplano ótimo de separação e os vetores de suporte que são a referência para obter maior margem de separação entre duas classes.

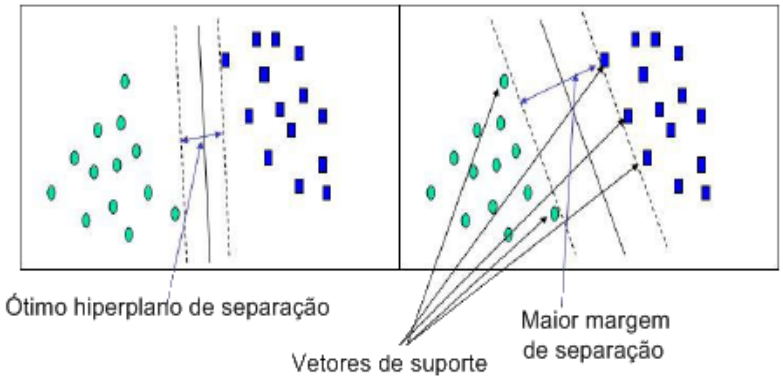

Figura 5.8: Vetores de suporte.

Para a formulação, vamos considerar o problema de separar o conjunto

de vetores de treinamento pertencentes as duas classes diferentes:

$$D = \{(x^1, y^1), ..., (x^l, y^l)\}, x\epsilon R^n, y\epsilon\{-1, 1\}, \tag{5.32}$$

com um hiperplano:

$$\langle w, x\rangle + b = 0. \tag{5.33}$$

O conjunto de vetores é considerado ótimo se são separados sem erro e a distância entre os mais próximos ao hiperplano de separação é máxima. Considera-se um hiperplano canônico para facilitar a formulação do problema [40], as variáveis w e b são limitadas por:

$$min_i|\langle w, x^i\rangle + b| = 1. \tag{5.34}$$

Para isso, [40] afirma que a norma do vetor deve ser igual ao inverso da distância, do ponto mais próximo no conjunto de dados até o hiperplano. A ideia é ilustrada na Fig 5.9 onde a distância do ponto mais próximo a cada hiperplano é mostrada.

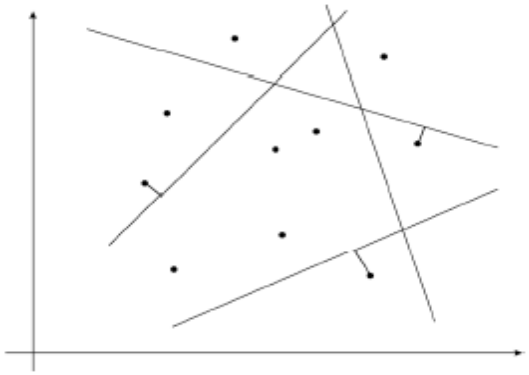

Figura 5.9: Hiperplanos canônicos

Um hiperplano separador em forma canônica deve obedecer às seguintes limitações:

$$y^i[\langle w, x^i\rangle + b] \geq 1, i = 1, ..., l. \tag{5.35}$$

A distância $d(w, b, x)$ de um ponto x do hiperplano (w, b) é:

$$d(w, b, x) = \frac{|\langle w, x^i\rangle + b|}{||w||}. \tag{5.36}$$

O hiperplano ótimo de separação é dado pela máxima margem de separação (ρ) , sujeito às limitações da Eq 5.35. A margem é dada por:

$$\rho(w, b) = \frac{2}{||w||}. \tag{5.37}$$

Assim, o hiperplano ótimo de separação ideal é aquele que minimiza:

$$\Phi = \frac{1}{2}||w||^2. \tag{5.38}$$

Esta função é independente de b como se mostra na Eq 5.35, ao mudar-se b irá movê-lo no sentido normal para si própria.

Para refletir sobre o modo como minimizar a Eq.5.38 com o princípio do erro mínimo estrutural, suponha o siguinte:

$$||w|| < A \tag{5.39}$$

Então, das Eqs 5.35 e 5.36:

$$d(w, b, x) \geq \frac{1}{A} \tag{5.40}$$

Um A qualquer dos pontos de dados é representado na Fig 5.10 e mostra como reduz os possíveis hiperplanos e é por isso sua capacidade.

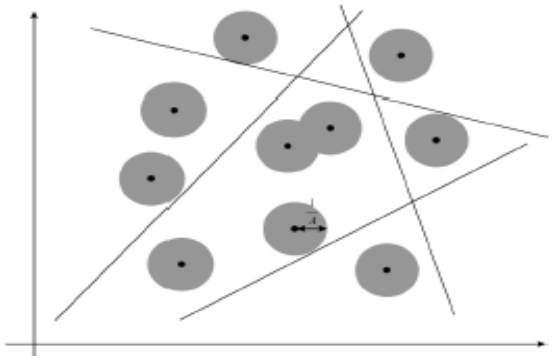

Figura 5.10: Hiperplanos canônicos

A solução para o problema de otimização da Eq 5.38, sob as condições da Eq 5.35, é dada por o ponto de sela da função Lagrange [40].

$$\Phi(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i(y^i[\langle w, x^i\rangle + b] - 1), \tag{5.41}$$

onde α são os multiplicadores de Lagrange. O função Lagrangeana tem de ser minimizada com respeito a w, b é maximizada com respeito $\alpha \geq 0$. A Eq 5.41 passa a ser transformada em um duplo problema, que é mais fácil de resolver. O problema dual é dado por:

$$max_\alpha W(\alpha) = max_\alpha - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle + \sum_{k=1}^{l} \alpha_k, \tag{5.42}$$

e, por conseguinte, a solução do problema é dado por,

$$\alpha^* = arg - min_\alpha \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_{k=1}^{l} \alpha_k, \tag{5.43}$$

com as seguintes restrições,

$$\alpha \geq 0, i = 1, ..., l. \tag{5.44}$$

$$\sum_{j=1}^{l} \alpha_j y_{j=0}. \tag{5.45}$$

Resolvendo as Eq 5.44 e Eq 5.45, determina-se os multiplicadores de Lagrange, e o hipérplano ótimo de separação é dado por:

$$w^* = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{5.46}$$

$$b^* = \langle -\frac{1}{2}(w^*, x_r + x_s)\rangle. \tag{5.47}$$

Onde x_r e x_s são qualquer vetor de suporte de cada classe que satisfaça:

$$\alpha_r, \alpha_s > 0, y_r = -1, y_(s) = 1. \tag{5.48}$$

Então o classificador é :

$$f(x) = sgn(\langle w^x, x\rangle + b). \tag{5.49}$$

Pelas condiçoes de Kuhn-Tucker [40] temos:

$$\alpha_i(y^i[\langle w, x^i + b] - 1) = 0, i = 1, ..., l, \tag{5.50}$$

e portanto apenas os pontos x^i que satisfazem:

$$y^i[\langle w, x^i\rangle + b] = 1, \tag{5.51}$$

terão multiplicadores de Lagrange não zero.

Estes pontos são designados Vetores de Suporte(SV). Se os dados foram linearmente separáveis todos os (SV) vão situar-se na margem e, consequentemente, o número de SVs pode ser muito pequeno. Consequentemente, o hiperplano é determinado por um pequeno subconjunto do grupo de treinamento.

Generalização do hiperplano òtimo de separação

Até agora, a discussão tem sido limitada aos casos em que a formação dos dados é linearmente separável. No entanto, em geral, este não será o caso, como é mostrado na Fig. 5.11.

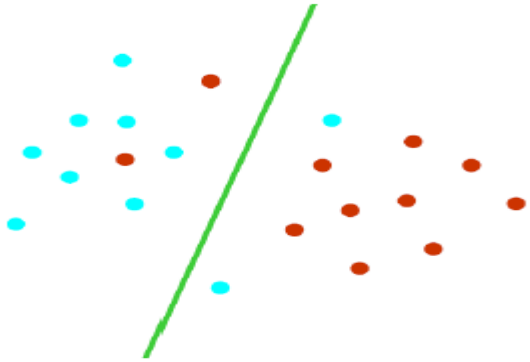

Figura 5.11: Dados não linearmente separáveis

No caso em que é esperado (ou possivelmente até mesmo conhecida) que um hiperplano pode separar corretamente os dados, o método mais indicado é unir uma função de custo a uma função erro adequada. Para permitir que o método do hiperplano ótimo de separação fora generalizado, Vapnik [40] introduziu uma variavel não negativa $\xi \geq 0$ e uma função de custo.

$$F_\sigma(\xi) = \sum_i \xi_i{}^\sigma \sigma > 0, \tag{5.52}$$

onde ξ_i é a medida dos erros desclassificados. O problema de optimização é agora colocado para minimizar o erro de classificação. As limitações da Eq. 5.35 são modificadas para o caso de não separável linearmente,

$$y^i[\langle w, x^i \rangle + b] \geq 1 - \xi_i, i = 1, ..., l, \tag{5.53}$$

onde $\xi_i \geq 0$. A generalização do hiperplano ótimo de separação é determinado

pelo vetor w e a função a otimizar fica

$$\Phi(w, \xi) = \frac{1}{2}||w||^2 + C\sum_i \xi_i, \tag{5.54}$$

A solução para minimizar este problema é idêntico ao caso de máquina de vetores de suporte linearmente separável. O coeficiente C é um parâmetro que introduz capacidade de controle dentro do classificador. C é o parâmetro regulador do classificador [40].

5.2.3 SVMs não Lineares

Há muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento por um hiperplano onde seria mais eficaz uma fronteira curva como é mostrado na Fig 5.12. As SVMs lidam com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como de entradas, para um novo espaço de maior dimensão, denominado espaço de características.

Seja $\Phi : X \rightarrow \Im$ um mapeamento, em que X é o espaço de entradas e $\Im$ denota o espaço de características. A escolha apropriada de Φ faz com que o conjunto de treinamento mapeado em $\Im$ possa ser separado por uma SVM linear. Em outras palavras dado um conjunto de dados não linear no espaço de entradas X, esse teorema afirma que X pode ser transformado em um espaço de características $\Im$ no qual com alta probabilidade, os dados são linearmente separáveis.

Para isso duas condições devem ser satisfeitas. A primeira é que a transformação seja não linear, enquanto a segunda é que a dimensão do espaço de características seja suficientemente alta.

Vamos a considerar o conjunto de dados mostrado na imagem da Fig 5.12. Transformando os dados de $\Re^2$ para $\Re^3$ com o mapeamento representado

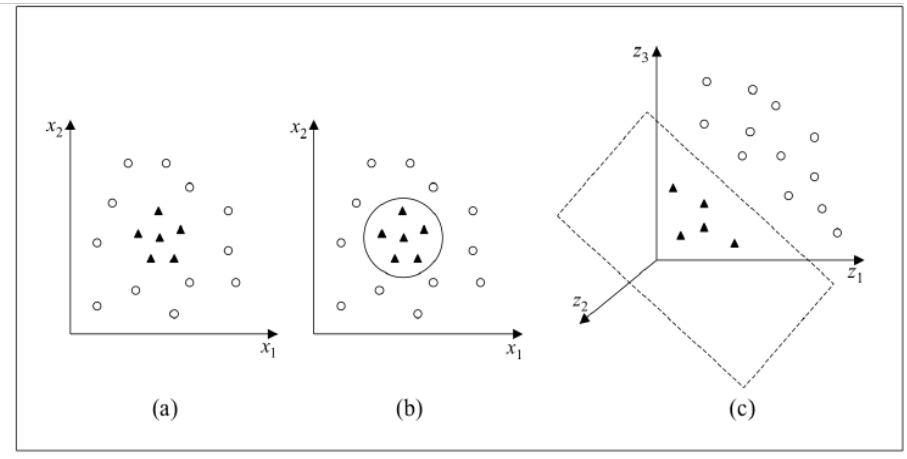

Figura 5.12: (a) Conjunto de dados não linear; (b) Fronteira não linear no espaço de entradas; (c)Fronteira linear no espaço de características

na Eq. 5.55, o conjunto de dados não linear em $\Re^2$ torna-se linearmente separável em $\Re^3$ como mostra a imagem (c) da Fig5.12 É possível, então, encontrar um hiperplano capaz de separar esses dados, descrito na Eq. 5.58. Pode-se verificar que a função apresentada, embora linear em $\Re^3$ corresponde a una fronteira não linear em $\Re^2$ como é mostrado na imagem (b) da Fig 5.12.

$$\Phi(x) = \Phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2), \tag{5.55}$$

$$f(x) = w \cdot \Phi(x) + b = w_1x_1^2 + w_2\sqrt{2}x_1x_2 + w_3x_2^2 + b, \tag{5.56}$$

aplicando o mapeamento o classificador se torna:

$$g(x) = sgn(f(x)) = sgn(\sum_{x_i \epsilon SV} \alpha_i^* y_i \Phi(x_i) \cdot \Phi(x) + b^*). \tag{5.57}$$

Como $\Im$ pode ter dimensão muito alta (até mesmo infinita), a computação de Φ pode ser extremamente custosa ou inviável. Por isso o mapeamento é

obtido por funções denominadas Kernels (K). Um Kernel K é uma função que recebe dois pontos x_i e x_j do espaço de entradas e computa o produto escalar desses dados no espaço de características

$$K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{5.58}$$

É comum empregar a função Kernel sem conhecer o mapeamento Φ, que é gerado implicitamente. A utilidade dos Kernels está, portanto, na simplicidade de seu cálculo e em sua capacidade de representar espaços abstratos.

Alguns dos Kernels mais utilizados na prática são os Polinomiais, os Gaussianos e os Sigmoidais, listados na Tabela 1.

Tabela 5.1: Funções Kernel mais comuns.

Tipo de Kernel	Função $K(x_i, x_j)$	Parâmetros
Polinomial	$(\delta(x_i \cdot x_j) + k)^d$	δ, k, d
Gaussiano	$exp(-\sigma\|x_i - x_j\|^2)$	σ
Radial Basis Function(RBF)	$(\frac{-\|x-x'\|^2}{2\sigma^2})$	σ^2

Capítulo 6

Resultados Experimentais

No trabalho [15] já se mostram algumas particularidades dos parâmetros obtidos do sinal glotal. Este trabalho teve como base a informação do sinal glotal de cada locutor, para tentar uma melhoria no desempenho da já conhecida técnica MFCC na tarefa de reconhecimento de locutor.

Neste trabalho utilizaram-se duas técnicas da área de inteligência computacional para comparar seu desempenho: a primeira, uma rede neural e a segunda, a técnica de máquina de vetores de suporte.

Foi usada uma base de dígitos para RAL obtida com o apoio do Instituto Militar de Engenharia (IME), produzidas em ambiente de laboratório e com a participação dos alunos de mestrado e de graduação do IME.

Dessa base foi criada outra, que contém vogais concatenadas de cada um dos dígitos da base original para poder extrair os parâmetros da estimação do sinal glotal. Esta base é um aporte do trabalho para próximas pesquisas e será detalhada a seguir.

Para o desenvolvimento das experiências utilizaram-se os software Weka [36] e Matlab.

6.1 Base de dígitos

A base de dígitos está estruturada para a realização de experimentos de RAL. Consta de 50 locutores femininos e 50 locutores masculinos, dos quais cada um deles repete três vezes as palavras: “zero”, “um”, “dois”, “três”, “quatro”, “cinco”, “seis”, “meia”, “sete”, “oito”, “nove”, em português falado no Brasil. Cada gravação tem uma taxa de amostragem de 11025 Hz e 16 bits de resolução com um só canal e gravados em ambiente de escritório. A formatação dos nomes dos arquivos é como segue:

D0R1LF01.wav

Onde D representa a palavra dígito; 0 é o dígito gravado; R representa a repetição; 1 é o número de repetições; L representa a palavra locutor; F(M) representa a palavra feminino (masculino); 01 é o número do locutor; logo um arquivo de nome D2R3LM02.wav contém a terceira repetição do dígito 2 pronunciada pelo locutor masculino 02.

6.1.1 Construção da base de sons vozeados, a partir de vogais concatenadas

Esta base foi construída pelo autor para a realização deste trabalho através da base de dígitos anterior, com o objetivo de poder trabalhar com vogais concatenadas para obter as características do sinal glotal. O trabalho para criação da base foi a seleção, corte e exportação manual das vogais de interesse.

Nesta etapa foi importante o uso de um software de voz que permitisse a visualização de ambos sinais e que póssuise um ambiente amigável que facilitasse a realização desta base. O software escolhido foi o Audacity [35].

As vogais extraídas foram: “o” para zero, “u” para um, “o” para dois, “e” para três, “a” para quatro, “i” para cinco, “e” para seis, “e” para sete, “o” para oito, “o” para nove e “e” para meia. A formatação utilizada para determinar o nome do arquivo é dada a seguir Vejamos, por exemplo, o arquivo de nome:

lm1n1r1vu.wav.

Nesse caso, o“l” representa que o locutor é do sexo masculino; o “n” representa o digito, nesse caso 1, o “r” representa o numero da repetição, nesse caso a primeira repetição, e o “v” representa a vogal, nesse caso a letra u.

6.2 Obtenção de características MFC

Para este trabalho foram usados 12 coeficientes MFC sem considerar seu primeiro coeficiente c_0 (que carrega muita informação do meio de transmissão [34]) e suas respectivas derivadas, conhecidas como coeficientes delta e delta-delta respectivamente, em um total de 36 coeficientes. Em [13], conclui-se que os coeficientes MFC mais suas duas derivadas têm um melhor desempenho na tarefa de reconhecimento de locutor, por terem mais informações e serem mais discriminantes.

Para a obtenção de características MFC, primeiro o sinal passa por um filtro de pré-ênfase com a=0,95, depois o sinal é janelado com janelas de Hamming com 30ms de largura. Para a extração dos coeficientes MFC foi utilizado um banco de 20 filtros triangulares. Para a tarefa de obtenção de características MFC foi utilizada a ferramenta [33] e o resultado é mostrado na Fig 6.1.

Depois de extraídas as características, Organizaram-se os dados em dois

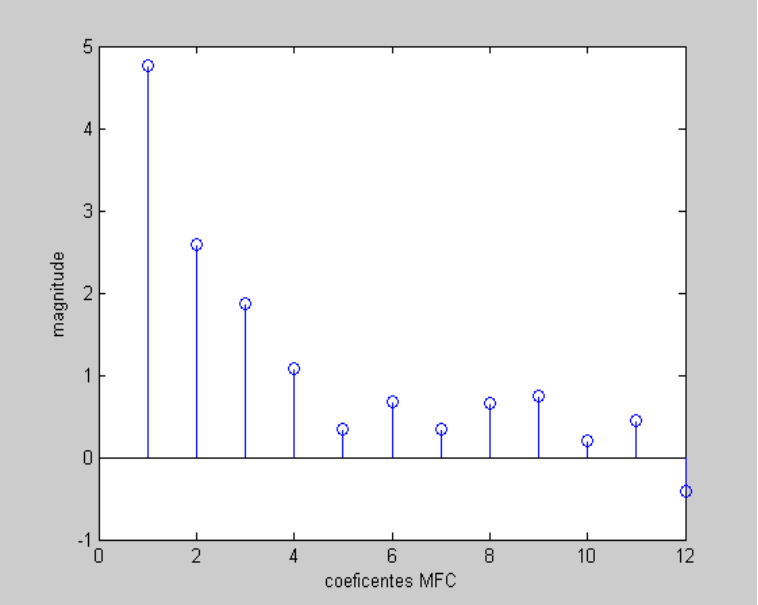

Figura 6.1: Gráfico dos coeficientes MFC da palavra nove.

grupos: o primeiro um grupo de vetores de 12 coeficientes MFC e um segundo grupo de Vetores de 36 coeficientes compostos por 12 coeficientes MFC e sua primeira e segunda derivadas. Estes serão utilizados depois para a tarefa de classificação.

6.3 Obtenção da estimativa do sinal glotal

Para a obtenção dos parâmetros do sinal glotal, primeiro deve-se obter a estimativa do sinal glotal pelo método PSIAIF (Pitch Synchronous Iterative Adaptive Inverse Filtering).

Para isso, foi utilizada a base de vogais concatenadas construída neste trabalho. Primeiro, estes sinais de voz passaram por um filtro de pré-ênfase para prevenir contra instabilidade numérica e, também, minimizar o efeito

dos lábios. Depois, segue a etapa de janelamento em pequenos trechos do sinal para poder capturar as características mais importantes de cada sinal. Neste trabalho, a janela utilizada foi de 30 ms ja , que foi a recomendadsa em [15].

Para obter a estimativa do sinal glotal pelo método PSIAIF, aplicam-se duas vezes o método IAIF como já foi comentado. A primeira com o objetivo de obter uma estimativa do período fundamental da voz, para poder analisar o sinal glotal de forma síncrona. Para isso, utilizaram-se janelas de 30 ms com 75% de superposição, ocasionando na saída o sinal glotal. Com esta primeira estimação do sinal glotal encontramos o período fundamental através dos picos máximos do sinal e este resultado é usado como base para um novo dimensionamento mais preciso da janela, para a segunda aplicação do método IAIF, para obter uma estimação do sinal glotal mais precisa. Para encontrar os picos do sinal usamos a rotina findpeaks do MATLAB [33]. Foram escolhidos, neste trabalho, três períodos fundamentais de cada sinal.

A estrutura do IAIF tem aplicação da técnica LPC que é a responsável pela filtragem de pré-ênfase, pela estimação do trato vocal e pela contribuição glotal. Nesta aplicação toma-se o valor de 45 coeficientes LPC, que é o valor sugerido em [15] e com o qual foram obtidos os melhores resultados comparados com outros. Na Fig 6.2, mostra-se um exemplo de estimativa glotal.

6.3.1 Extração de características do sinal glotal

O sinal glotal foi obtido por filtragem inversa, pelo método PSIAIF, tomando-se três períodos de cada sinal como em [15]. Neste trabalho, também como em

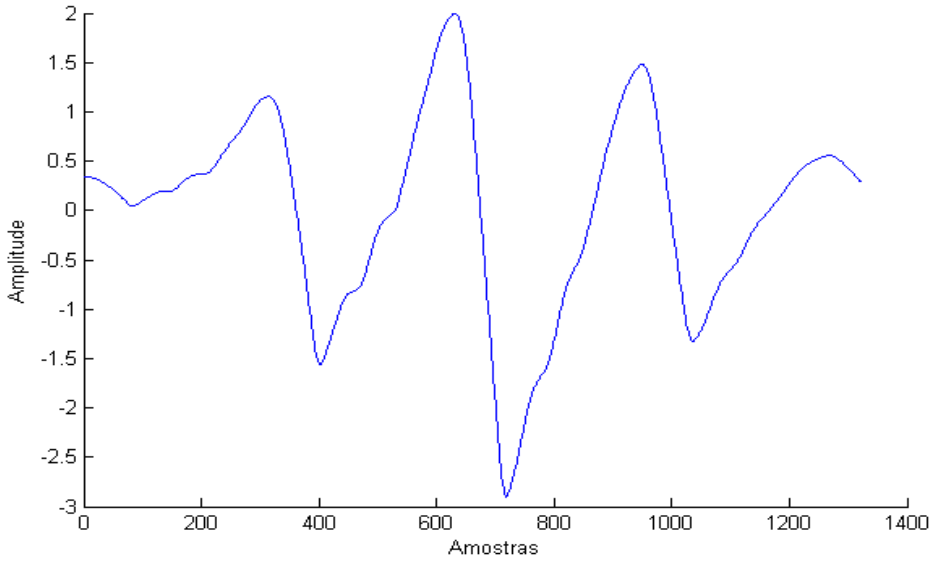

Figura 6.2: Gráfico de vogal /a/ concatenada com 45 coeficientes LPC.

[15], foram extraídos dois parâmetros de cada trecho de sinal glotal obtido. São eles: Ko, que é a diferença entre o instante de máximo fechamento e de máxima abertura, Av que é a amplitude de vozeamento, definida como a diferença de amplitudes entre os valores máximo e mínimo do sinal glotal. Esses parâmetros estão representados na Fig.6.3.

Porém, tomaremos ainda, como inovação, um outro parâmetro que chamamos de *pp*. Este parâmetro é a diferença entre os pontos de máxima abertura. Não vimos referências sobre esse parâmetro em outros trabalhos. Os instantes de máxima abertura e máximo fechamento foram encontrados com ajuda da rotina *findpeaks* com o qual se obtiveram os parâmetros mencionados, como se mostra na Fig 6.4.

Na Fig 6.4 pode-se observar que todos os máximos e mínimos encontrados pela rotina *findpeaks* não são os verdadeiros. Por exemplo, o primeiro máximo

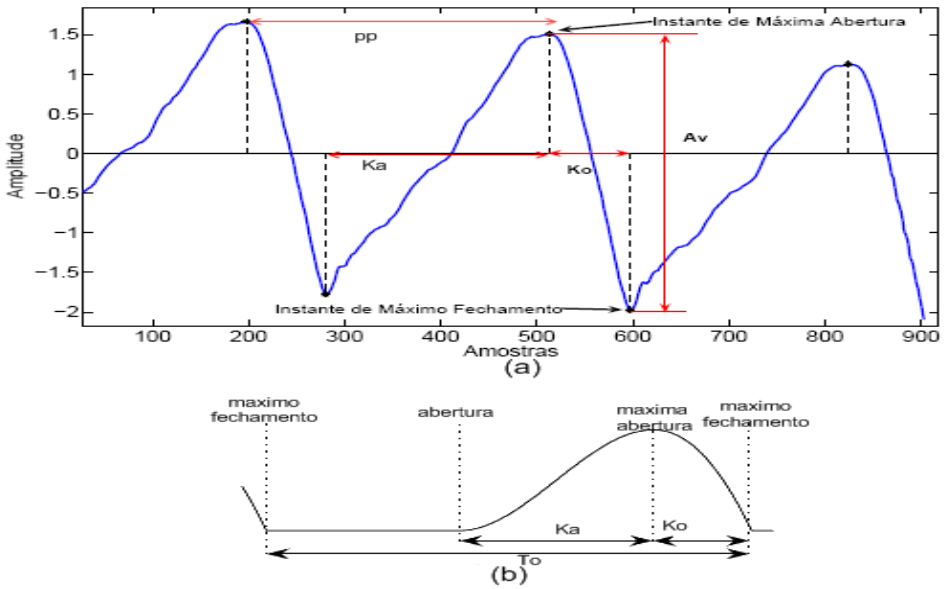

Figura 6.3: Sinal glotal e seus parâmetros.

não corresponde a uma máxima abertura isso deve-se ao fato de se escolher uma "media" de todas as variáveis para todos os locutores e, assim, algumas estimativas glotais aparecem com ruído, o que dificulta a escolha da máxima abertura e o máximo fechamento em algumas amostras. Dessa forma, tivemos que conferir visualmente, para obter os valores exatos. Isso torno a tarefa um pouco trabalhosa.

Como são três períodos, obtive-se por cada estimação do sinal glotal 3 valores (Av), 3 valores (Ko) e 2 valores (pp), totalizando 8 parâmetros para cada estimativa do sinal glotal.

Os passos para extrair as características do sinal glotal são mostrados na Fig 6.5.

Ao discriminar visualmente cada um dos parametros glotais , observamos que para a tarefa do reconhecimento de locutor o parâmetro (Av) não era

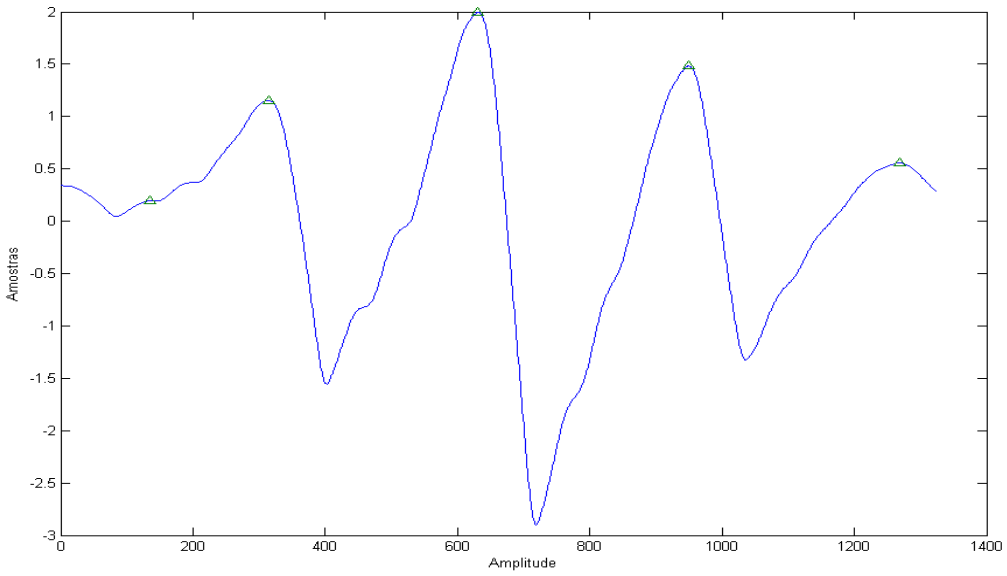

Figura 6.4: Gráfico mostrando os instantes de maxima abertura achados pela rotina *findpeaks*.

um bom discriminante. Chegando-se à conclusão de que a amplitude do sinal (Av) varia por causas diversas, como a distância do locutor ao microfone e a intensidade da voz.

Por outro lado, o parâmetro *pp* mostrou-se ser um bom discriminante. A Fig 6.6 mostra a distribuição do parâmetro *pp* para cinco locutores. Cada cor é um locutor (classe), as divisões da parte de baixo mostram os diferentes valores de *pp* neste caso os valores estão na faixa entre 65 e 131 para primeiro parâmetro pp e de 63 até 128 para o segundo parâmetro *pp*. Cada locutor é representado por 33 vetores (amostras). Nos gráficos percebe-se, por exemplo, que para o locutor representado pela cor vermelha, os valores dos parâmetros *pp* estão entre a faixa 65 e 87 e para o locutor representado pela cor azul claro, os valores de *pp* estão entre 98 e 131, o que evidencia usar

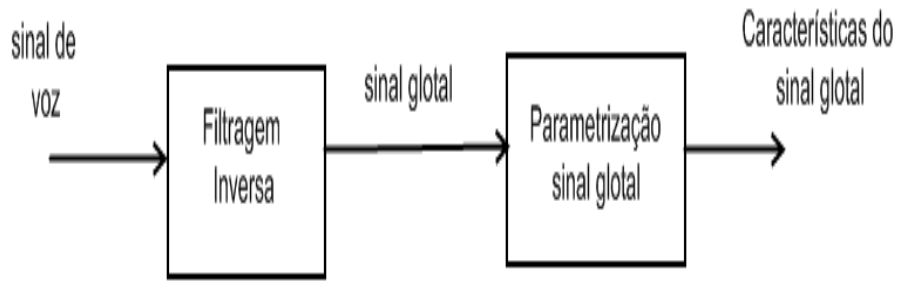

Figura 6.5: Extração de características do sinal glotal

pp como discriminante entre vários locutores.

As características obtidas logo serão agregadas às características MFC para formar o vetor híbrido.

6.3.2 Vetor Híbrido de características: Coeficientes MFC e características do sinal glotal

Como um dos principais objetivos deste trabalho, inclui-se um vetor com os parâmetros da estimativa do sinal glotal e os coeficientes MFC.

No trabalho de [15], foram mostradas algumas características do sinal glotal com muito poder de discriminação, porém foi feita com poucos locutores. A grande dificuldade de utilizar o sinal glotal para a tarefa de reconhecimento de locutor é a complexidade de obter o sinal, já que como foi citado, antes precisava-se de aparelhos para obtê-la, mas isso foi solucionado com a técnica de filtragem inversa. O vetor híbrido aqui construído é composto de coeficientes MFC, suas derivadas (Δ e $\Delta\Delta$) e os parâmetros da estimativa do sinal glotal Ko e *pp* de cada sinal. Por exemplo, consideremos o digito do arquivo:

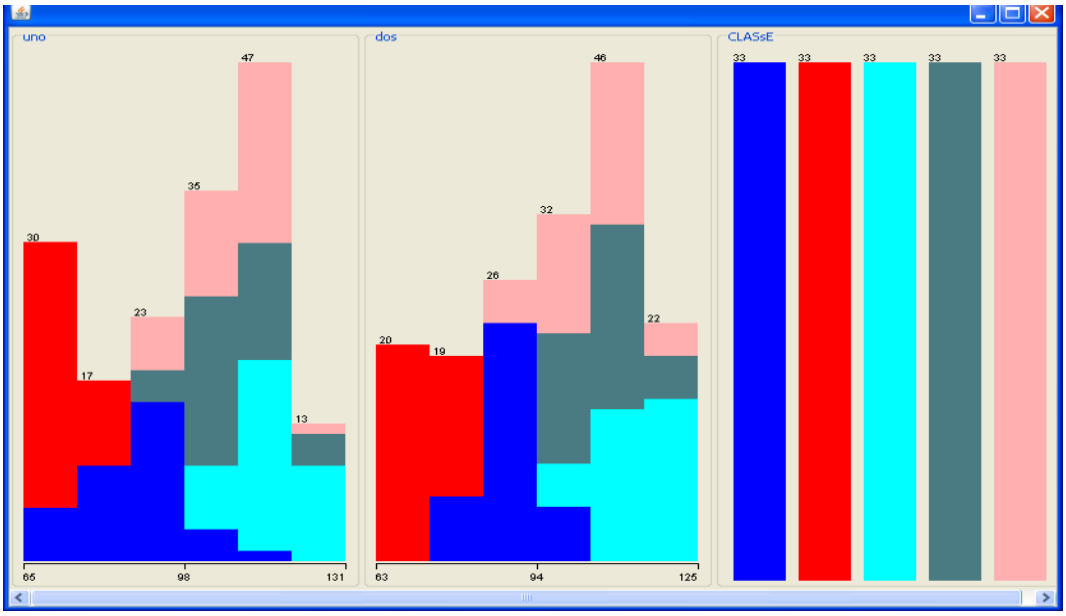

Figura 6.6: distribução do parâmetro *pp* do sinal glotal

"D1R1LM01"

que é o digito um . Desse sinal, primeiro, foram extraídos os coeficientes MFC. Além disso, foi extraída a vogal "u" e guardada na base de vogais como:

"lm1n1r1vu"

São extraídos, então, os parâmetros da estimativa do sinal glotal da vogal u e todos unidos em um só vetor. Este é o Vetor Híbrido de características: Coeficientes MFC características do sinal glotal. As diferentes configurações deste vetor utilizadas neste trabalho são as seguintes:

-12 coeficientes MFC + Ko.

-12 coeficientes MFC + Ko + pp.

-12 coeficientes MFC + coeficientes Δ e $\Delta\Delta$ + Ko.

-12 coeficientes MFC e suas derivadas Δ e $\Delta\Delta$ + Ko + pp.

Todas essas configurações foram testadas em uma rede neural construída.

6.4 Rede Neural Artificial(RNA)

Para utilizar uma rede neural artificial (RNA) na tarefa de reconhecimento de locutor deve-se levar em conta os seguintes aspectos:
-Parâmetros de uma Rede *Multilayer Perceptron*.
-Normalização dos pesos.
-Critério de parada do treinamento de RNA
-Taxa de aprendizagem e momento.
-Variações no treinamento da rede MLP

6.4.1 Parâmetros de uma rede *Multilayer Perceptrons*(MLP)

A rede neural que se utilizou neste trabalho tem um número de entradas correspondentes ao numero de características do vetor híbrido, quer disser que depende do comprimento do vetor de características glotais e de coeficientes MFC.

A escolha feita neste trabalho foi dependente de cada aplicação, além disso, foram consideradas diferentes tentativas até chegar à melhor configuração. Para todas as aplicações, a quantidade de neurônios da camada escondida foi escolhida pela média aritmética do número de atributos do vetor de entrada e do número de classes de saída da rede neural. Outras heurísticas também foram testadas com diferentes números de neuronios na camda escondida, mas o melhor desempenho foi obtido com a média aritmética, seguindo [23].

Um exemplo de rede neural com as características de arquitetura *multi-*

layer percptrons é mostrado na Fig 6.7, onde se tem um vetor de entrada de 17 características (composta por 12 coeficientes MFC $+Ko+pp$), uma camada escondida sendo a media aritmética de entradas e saídas e por último, e cinco saidas que representam os 5 locutores (as cinco classes). O software utilizado nas experiências com redes neurais foi o Weka [36].

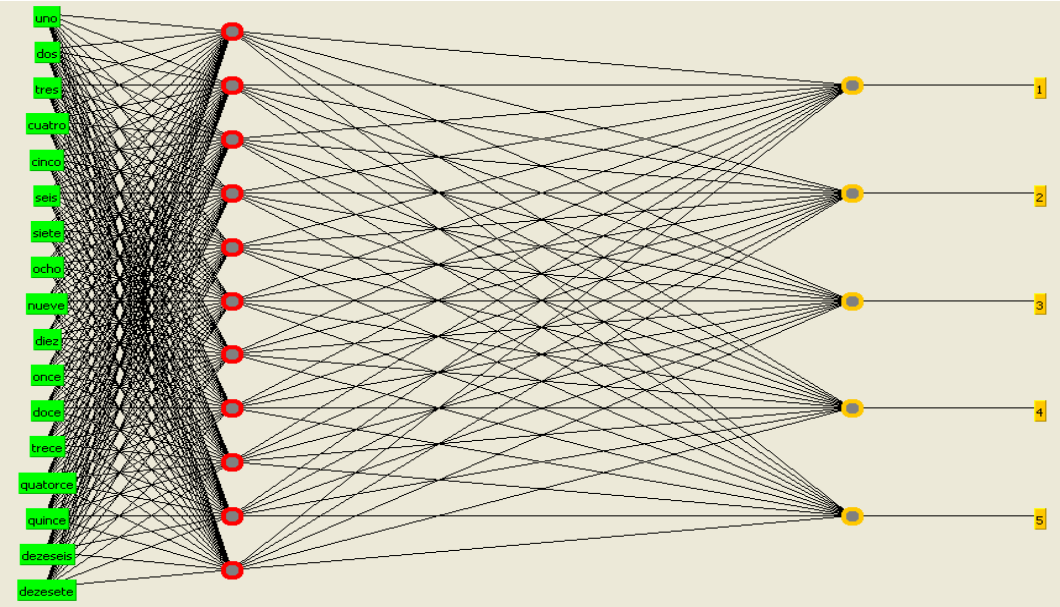

Figura 6.7: Exemplo de rede neural com arquitetura *multilayer perceptrons*

6.4.2 Normalização dos Pesos

A fase de normalização é de suma importância para o sistema, pois é nela que ocorrerá o tratamento dos valores coletados. A partir dos dados adquiridos, no módulo de captura, cabe ao módulo de normalização o preparo para que sejam entregues à análise da rede neural. Esta fase ajusta a escala de valores obtidos. As vantagens da normalização é a economia no processamento além de encaixar todos os dados em um faixa, ou seja, menor dispersão.

A faixa utilizada para as aplicações neste trabalho é de [-1, 1].

6.4.3 Critério de parada do treinamento do RNA

O treinamento pode ser interrompido em três circunstâncias, quando é atingido o número máximo de épocas do treinamento, ou quando é atingido o erro mínimo desejado ou quando uma determinada porcentagem dos dados de treinamento é corretamente classificada. O principal problema nas épocas de treinamento é o sobre-treinamento (*overfitting*) como já se comentou anteriormente. Para evitar este problema, neste trabalho, utilizou-se uma estratégia que une os três critérios de parada de treinamento e utiliza-se um conjunto de validação. A validação cruzada consiste em dividir os padrões de treinamento em três conjuntos

- Um conjunto de treinamento que é grupo dos padrões usados para modificar os pesos.
- Um conjunto de validação que são os padrões usados para verificar o problema de *overfitting*.
-Um conjunto de teste que são os padrões para testar o desempenho do modelo final.

Coloca-se a quantidade de épocas de treinamento e as épocas só devem ser interrompidas quando o erro dos padrões do conjunto de validação começam a subir de forma consistente.

6.4.4 Momento e taxa de aprendizagem

A taxa de aprendizagem pode ser constante ou adaptativa. O momento foi colocado para as aplicações deste trabalho em 0,9 e a taxa de aprendizado foi em media 0,3.

6.5 Experiências

A primeira aplicação divide-se em duas experiências: Primeiro, foram utilizados 30 locutores masculinos falando todos os dígitos com três repetições cada. Para cada palavra, foram extraídos os coeficientes MFC e sua primeira e segunda derivadas utilizando vinte filtros passa-faixa triangulares [13]. De cada palavra foram extraídas vogais para depois proceder a aplicação do algoritmo de filtragem inversa discutido. Extraíram-se sons vozeados de cada palavra. Escolheram-se partes periódicas de cada palavra, descritos a seguir: "o" para zero, "u" para um, "o" para dois, "e" para três, "a" para quatro, "i" para cinco, "e" para seis, "e" para sete, "o" para oito, "o" para nove e "e" para meia. Todas foram extraídas de forma semi-automática.

Obtenção da estimativa do sinal glotal

Os melhores resultados para o sinal glotal foram obtidos com 45 coeficientes LPC [8] e, a partir desta estimativa, foram extraídos dois parâmetros *Ko* e *pp*. Como são três períodos fundamentais, foram obtidos para cada palavra três valores para *Ko* e dois valores para *pp*.

Rede neural construída para o reconhecimento/classificação

Foram testadas duas configurações de treinamento da rede neural: Na primeira, dividiu-se a base de dados em três grupos. Um grupo com 75% (742 vetores) dos dados para treinamento; outro grupo, com 20% (198 vetores) dos dados para validação e os restantes 5% (50 vetores) para teste. Na segunda configuração foram usadas 95% (940 vetores) das gravações da base para treinamento e 5% (50 vetores) para teste, tomando como referência as épocas de treinamento, antes do sobretreinamento (*overfitting*), da primeira

configuração. Para escolher estes valores levou-se em conta que a base é ruidosa por isso a base de treinamento e validação deixarem-se com os valores maiores possiveis. Na segunda configuração não foram usados dados para validação.

As duas primeiras experiências foram considerados apenas locutores masculinos, sendo a primeira experiência apenas levando-se em consideração os 12 coeficientes MFC e as características glotais e a segunda experiência foi realizada levando-se em consideração os 36 coeficientes MFC e as características glotais. A terceira e a quarta experiência foram realizadas nos mesmos moldes da primeira e da segunda, considerando-se, porém, locutores femininos. Para cada uma das quatro experiências consideram-se as duas configurações de rede neural.

6.5.1 Primeira experiência

A primeira experiência pode ser dividida em quatro partes, de acordo com o vetor usado na entrada da rede neural: (I) apenas os 12 coeficientes MFC, (II) os 12 coeficientes MFC mais três valores correspondentes às características glotais Ko, totalizando 15 componentes, (III) os 12 coeficientes MFC mais dois valores correspondentes às características glotais pp, totalizando 14 componentes, (IV) 12 coeficientes MFC, mais três valores correspondentes às características glotais Ko, mais dois valores correspondentes às características glotais pp, totalizando 17 componentes.

Nas Tabs. 6.1 e 6.2 são mostrados os resultados da primeira experiência com a primeira configuração e com a segunda configuração, respectivamente.

Figura 6.8: Configuração rede neural

Tabela 6.1: Primeira experiência com a primeira configuração da rede neural.

Atributos do vetor	classificação
12 MFCC	52%
12 MFCC+pp	54%
12 MFCC+Ko	54%
12 MFCC+Ko+pp	58%

6.5.2 Segunda experiência

A diferença dessa experiência em relação à primeira refere-se apenas ao número de coeficientes MFC. Na segunda experiência foram considerados 36 coeficientes, pois incluíram-se a primeira e a segunda derivadas dos coeficientes MFC.

Nas Tabs. 6.3 e 6.4 são mostrados os resultados da segunda experiência com a primeira configuração e com a segunda configuração, respectivamente.

Tabela 6.2: Primeira experiência com a segunda configuração de rede neural.

Atributos do vetor	classificação
12 MFCC	56%
12 MFCC+pp	56%
12 MFCC+Ko	56%
12 MFCC+Ko+pp	60%

Figura 6.9: Configuração rede neural

Estes resultados foram os melhores depois de provar diferentes configurações da rede neural. Através dos resultados mostrados nas Tabs. 6.1, 6.2, 6.3 e 6.4, conclui-se que o vetor incluindo todas as características glotais é o mais efetivo para reconhecimento de locutor, pois apresenta as maiores porcentagens de acerto. Os valores são ainda mais significativos quando são incluídos, além dos coeficientes MFC, também as suas derivadas. A característica Ko foi a que mais influenciou o resultado, quando comparado aos resultados com a utilização do vetor de entrada apenas com os coeficientes MFC.

Tabela 6.3: Segunda experiência com a primeira configuração da rede neural.

Atributos do vetor	classificação
36 MFCC	64%
36 MFCC+pp	70%
36 MFCC+Ko	84%
36 MFCC+Ko+pp	88%

Tabela 6.4: Segunda experiência com a segunda configuração da rede neural.

Atributos do vetor	classificação
36 MFCC	62%
36 MFCC+pp	72%
36 MFCC+Ko	88%
36 MFCC+Ko+pp	92%

Observou-se também que a segunda configuração da rede apresentou melhores resultados, o que já era esperado devido ao maior número de padrões de treinamento.

6.5.3 Terceira experiência

A diferença desta aplicação com a primeira é só a inclusão de locutores femininos na base de dados. Foram utilizados 30 locutores divididos em 20 locutores

masculinos e 10 femininos. Como na primeira experiência, dividimos em quatro partes, de acordo com o vetor usado na entrada da rede neural: (i) apenas os 12 coeficientes MFC, (ii) os 12 coeficientes MFC mais três valores correspondentes às características glotais Ko, totalizando 15 componentes, (iii) os 12 coeficientes MFC mais dois valores correspondentes às características glotais *pp*, totalizando 14 componentes, (iv) 12 coeficientes MFC, mais três valores correspondentes às características glotais Ko, mais dois valores correspondentes às características glotais *pp*, totalizando 17 componentes. Nas Tabs. 6.16 e 6.6 são mostrados os resultados da terceira experiência com a primeira configuração e com a segunda configuração, respectivamente.

Figura 6.10: Configuração rede neural

Tabela 6.5: Terceira experiência com a primeira configuração da rede neural.

Atributos do vetor	classificação
12 MFCC	40%
12 MFCC+pp	50%
12 MFCC+Ko	50%
12 MFCC+Ko+pp	54%

Tabela 6.6: Terceira experiência com a segunda configuração de rede neural.

Atributos do vetor	classificação
12 MFCC	42%
12 MFCC+pp	52%
12 MFCC+Ko	54%
12 MFCC+Ko+pp	56%

6.5.4 Quarta experiência

A diferença dessa experiência em relação à terceira refere-se apenas ao número de coeficientes MFC. Na terceira experiência foram considerados 12 coeficientes, pois incluíram-se a primeira e a segunda derivadas dos coeficientes MFC.

Nas Tabs. 6.7 e 6.8 são mostrados os resultados da segunda experiência com a primeira configuração e com a segunda configuração, respectivamente.

Figura 6.11: Configuração rede neural

Tabela 6.7: Quarta experiência com a primeira configuração da rede neural

Atributos do vetor	classificação
36 MFCC	56%
36 MFCC+pp	64%
36 MFCC+Ko	68%
36 MFCC+Ko+pp	80%

Estes resultados foram os melhores depois de provar diferentes configurações da rede neural. Através dos resultados mostrados nas Tabs. 6.6, 6.7 e 6.8, tal como nas experiências 1 e 2, conclui-se que o vetor incluindo todas as características glotais é o mais efetivo para reconhecimento de locutor, pois apresenta as maiores porcentagens de acerto. Quando os locutores femininos foram incluídos, o desempenho do classificador foi inferior comparado com os resultados das experiências 1 e 2. Isso deve-se ao fato da estimativa glotal dos locutores femininos ser pior do que com os locutores masculinos, talvez causado pelas frequências mais altas do sinal glotal em locutores femininos.

Tabela 6.8: Quarta experiência com a segunda configuração da rede neural

Atributos do vetor	classificação
36 MFCC	64%
36 MFCC+pp	68%
36 MFCC+Ko	74%
36 MFCC+Ko+pp	82%

Isso quer dizer que a base é ruidosa com a precisa de locutores femininos.

Um exemplo de estimativa do sinal glotal de um locutor feminino é mostrado na Fig 6.12, que é uma estimativa bastante ruidosa comparada com as estimativas de vozes masculinas.

Nas experiências 3 e 4 pode-se ver o bom desempenho dos parâmetros da estimativa do sinal glotal no reconhecimento de locutor, já que os resultados somente com os coeficientes MFC foram muito menores que nas experiências 1 e 2, e com os parâmetros da estimativa do sinal glotal subiram consideravelmente seu desempenho.

Para o desenvolvimento das experiências foram utilizados dois softwares MATLAB e WEKA [36].

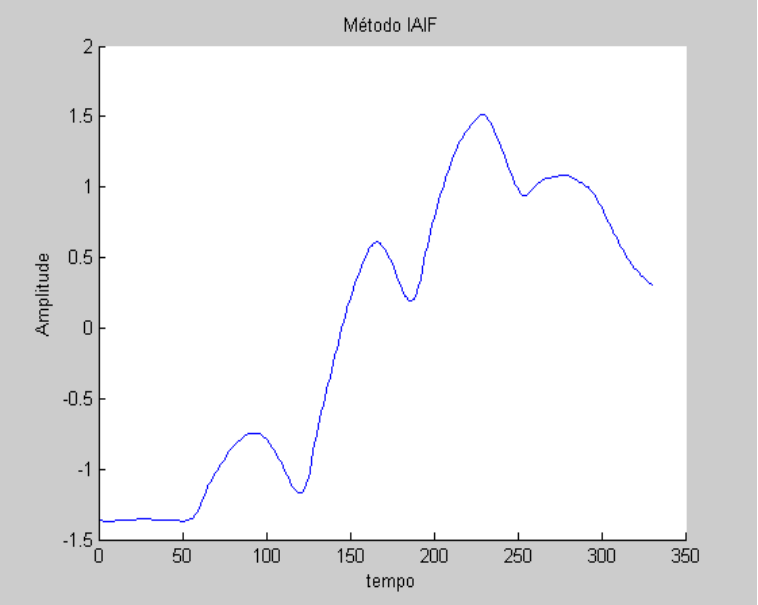

Figura 6.12: estimativa do sinal glotal do locutor feminino

6.5.5 Experiências com máquina de vetores de suporte

A estratégia resumida do SVM consiste em, dado um conjunto de vetores de treinamento pertencente a duas classes separáveis, o SVM encontra o hiperplano com a máxima distância Euclidiana do conjunto de treinamento. Para a tarefa de classificação com SVM é crucial a escolha da função Kernel(escolhida pelo usuário segundo sua natureza).

Neste trabalho foram escolhidas duas funções Kernel, linear e RBF (Radial Basis Function).

SVM linear

A teoria mais relevante do SVM linear já foi mostrada neste trabalho. Mostra-se aqui aspectos importantes para levar em conta a escolha do classificador SVM linear que tenha um bom desempenho. A SVM linear também é conhe-

cida como Função Kernel polinomial de potencia 1. Para projetar o melhor classificador SVM linear, os dados de entrada têm que ser normalizados para um melhor desempenho do algoritmo. Um valor que influencia o desempenho do algoritmo é o termo de regularização, denotado por C, que impõe um peso à minimização dos erros no conjunto de treinamento em relação à minimização da complexidade do modelo. Para este trabalho, provaram-se vários valores da constante C para cada uma das configurações dos vetores de entrada apresentados.

Apresenta-se a seguir duas experiências realizadas com SVM, o que chamaremos de quinta e sexta experiencias.

6.5.6 Quinta experiência

Para esta experiência utilizaram-se duas bases de dados já trabalhadas nas experiências anteriores, uma de 30 locutores masculinos e outra dividida em 20 locutores masculinos e 10 locutores femininos. Nesta experiência, como vetores de entrada foram utilizados: (i) apenas os 12 coeficientes MFC, (ii) os 12 coeficientes MFC mais três valores correspondentes às características glotais Ko, totalizando 15 componentes, (iii) os 12 coeficientes MFC mais dois valores correspondentes às características glotais *pp*, totalizando 14 componentes, (iv) 12 coeficientes MFC, mais três valores correspondentes às características glotais Ko, mais dois valores correspondentes às características glotais *pp*, totalizando 17 componentes, (v) 36 coeficientes MFC, (vi) os 36 coeficientes MFC mais três valores correspondentes às características glotais Ko, totalizando 39 componentes, (vii) os 36 coeficientes MFC mais dois valores correspondentes às características glotais *pp*, totalizando 38 componentes, (viii) 36 coeficientes MFC, mais três valores correspondentes às características glotais

Ko, mais dois valores correspondentes às características glotais pp, totalizando 41 componentes. Utilizou-se como teste o 5% da base total 50 vetores. A experiência consiste em classificar os vetores de entrada com a técnica SVM, usando função Kernel linear. Consideramos diversos valores de C para verificar o desempenho do algoritmo.

Tabela 6.9: Quinta experiência com a base de dados de 30 locutores masculinos.

Atributos do vetor	C=1	C=10	C=100
12 MFCC	44%	48%	48%
12 MFCC+Ko	52%	58%	60%
12 MFCC+pp	52%	64%	60%
12 MFCC+Ko+pp	54%	68%	64%
36 MFCC	64%	62%	48%
36 MFCC+pp	74%	70%	70%
36 MFCC+Ko	74%	68%	68%
36 MFCC+Ko+pp	78%	74%	74%

Tabela 6.10: Quinta experiência com a base de dados de 20 locutores masculinos e 10 locutores femininos.

Atributos do vetor	C=1	C=10	C=100
12 MFCC	38%	44%	46%
12 MFCC+Ko	60%	66%	72%
12 MFCC+pp	44%	66%	72%
12 MFCC+Ko+pp	58%	66%	74%
36 MFCC	48%	58%	52%
36 MFCC+pp	70%	70%	68%
36 MFCC+Ko	60%	62%	58%
36 MFCC+Ko+pp	74%	74%	78%

Observa-se que dependendo das características usadas no vetor de entrada os melhores resultados variaram, de acordo com os valores de C considerados. Os resultados são muito menores aos obtidos por RNA.

Maquina de Vetores de suporte com Kernel RBF

A função Kernel do tipo RBF tem como equação característica $\Phi = (\frac{-\|x-x'\|^2}{2\sigma^2})$. Aqui entram dois parâmetros que tem-se que determinar, o parâmetro C que já foi explicado, e o parâmetro σ^2. A saída da função Kernel depende da distância euclidiana de x' ate x (um destes pontos será o vetor de suporte e o outro será um ponto de teste). O vetor de suporte será o centro da RBF e σ^2 vai determinar a área de influencia dela sobre os dados. Um maior valor de σ^2 (maior variância) significa que a área de influência do vetor de suporte será maior. Quando um vetor suporte influência uma área maior, todos os outros vetores de suporte na área vão aumentar em valor para compensar essa influência, até que todos os valores cheguem ao equilíbrio.

Um maior valor também reduz o número de vetores. O problema do valor de σ^2 ser muito alto é o problema da generalização. Na Fig 6.13 é mostrado o gráfico de uma função RBF com $\sigma^2 = 1.5$.

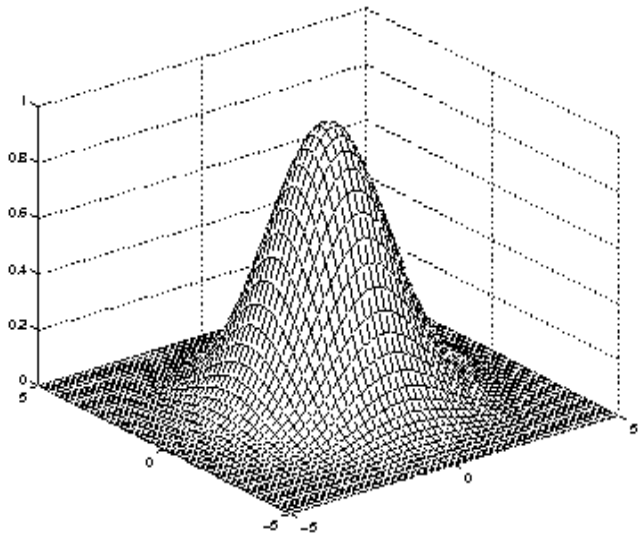

Figura 6.13: Função Kernel RBF

6.5.7 Sexta experiência

A diferença desta experiência com relação à quinta refere-se à troca da função Kernel linear por uma função Kernel RBF onde tomaram-se três valores diferentes de σ^2. Para cada valor de σ^2 tomaram-se vários valores de C, para entender o desempenho do algoritmo.

Tabela 6.11: Sexta experiência com a base de dados de 30 locutores masculinos com $\sigma^2 = 0.01$ e variando C.

Atributos do vetor	C=100 $\sigma^2 = 0.01$	C=150 $\sigma^2 = 0.01$	C=200 $\sigma^2 = 0.01$
12 MFCC	46%	52%	52%
12 MFCC+Ko	54%	56%	58%
12 MFCC+pp	58%	58%	58%
12 MFCC+Ko+pp	60%	64%	64%
36 MFCC	66%	60%	60%
36 MFCC+pp	78%	70%	70%
36 MFCC+Ko	78%	72%	70%
36 MFCC+Ko+pp	78%	82%	74%

Tabela 6.12: Sexta experiência com a base de dados de 30 locutores masculinos com $\sigma^2 = 0.1$ e variando C.

Atributos do vetor	C=1 $\sigma^2 = 0.1$	C=10 $\sigma^2 = 0.1$	C=100 $\sigma^2 = 0.1$
12 MFCC	42%	48%	46%
12 MFCC+Ko	44%	48%	50%
12 MFCC+pp	54%	54%	54%
12 MFCC+Ko+pp	54%	60%	60%
36 MFCC	56%	60%	60%
36 MFCC+pp	62%	64%	66%
36 MFCC+Ko	62%	64%	66%
36 MFCC+Ko+pp	66%	66%	72%

Tabela 6.13: Sexta experiência com a base de dados de 30 locutores masculinos com $\sigma^2 = 1$ e variando C.

Atributos do vetor	C=1 $\sigma^2 = 1$	C=10 $\sigma^2 = 1$	C=100 $\sigma^2 = 1$
12 MFCC	52%	54%	50%
12 MFCC+Ko	56%	58%	56%
12 MFCC+pp	58%	58%	58%
12 MFCC+Ko+pp	64%	64%	64%
36 MFCC	58%	64%	60%
36 MFCC+pp	70%	70%	70%
36 MFCC+Ko	72%	72%	72%
36 MFCC+Ko+pp	82%	82%	82%

Tabela 6.14: Sexta experiência com a base de dados de 20 locutores masculinos e 10 locutores femininos com $\sigma^2 = 0.01$ e variando C.

Atributos do vetor	C=100 $\sigma^2 = 0.01$	C=150 $\sigma^2 = 0.01$	C=200 $\sigma^2 = 0.01$
12 MFCC	46%	46%	46%
12 MFCC+Ko	72%	66%	64%
12 MFCC+pp	72%	58%	56%
12 MFCC+Ko+pp	74%	68%	64%
36 MFCC	52%	56%	56%
36 MFCC+pp	68%	72%	70%
36 MFCC+Ko	58%	72%	70%
36 MFCC+Ko+pp	78%	78%	78%

Tabela 6.15: Sexta experiência com a base de dados de 20 locutores masculinos e 10 locutores femininos com $\sigma^2 = 0.1$ e variando C.

Atributos do vetor	C=1 $\sigma^2 = 0.1$	C=10 $\sigma^2 = 0.1$	C=100 $\sigma^2 = 0.1$
12 MFCC	44%	46%	46%
12 MFCC+Ko	68%	64%	66%
12 MFCC+pp	66%	62%	66%
12 MFCC+Ko+pp	66%	66%	66%
36 MFCC	52%	56%	58%
36 MFCC+pp	72%	70%	70%
36 MFCC+Ko	64%	66%	64%
36 MFCC+Ko+pp	74%	74%	74%

Tabela 6.16: Sexta experiência com a base de dados de 20 locutores masculinos e 10 locutores femininos com $\sigma^2 = 1$ e variando C.

Atributos do vetor	C=1 $\sigma^2 = 1$	C=10 $\sigma^2 = 1$	C=100 $\sigma^2 = 1$
12 MFCC	46%	46%	46%
12 MFCC+Ko	64%	64%	64%
12 MFCC+pp	56%	56%	56%
12 MFCC+Ko+pp	64%	64%	64%
36 MFCC	56%	56%	56%
36 MFCC+pp	72%	72%	72%
36 MFCC+Ko	72%	72%	72%
36 MFCC+Ko+pp	80%	80%	80%

Para valores muito pequenos de σ^2, a base consegue bons resultados, mas o modelo varia muito com as mudanças de C. Para σ^2 maior ou igual a 1 o modelo é muito mais estável para os diferentes valores C e mostra o melhor desempenho, para σ^2 maiores, o desempenho piora. Os resultados com a função Kernel RBF são melhores que com Kernel linear e o modelo é muito mais estável.

Capítulo 7

Conclusões e trabalhos futuros

7.1 Conclusões

-Neste trabalho foi proposta uma técnica para reconhecimento automático de locutor unindo em um único vetor os coeficientes MFC com os parâmetros extraídos da estimativa do sinal glotal. A superioridade do desempenho da técnica com o vetor híbrido, em comparação com um vetor de apenas coeficientes MFC, foi comprovada em várias experiências com bases de vozes e para diferentes combinações do vetor híbrido. Observamos que os parâmetros da estimativa do sinal glotal apresentam-se como bons discriminantes para o locutor, sendo um excelente complemento para a técnica dos coeficientes MFC.

-O vetor híbrido de características MFC e parâmetros da estimativa do sinal glotal, foram provados na tarefa de reconhecimento de voz obtendo resultados muito abaixo do esperado e apresentando mal desempenho comparados com os obtidos na tarefa de reconhecimento do locutor, o que leva

a concluir que os parâmetros glotais servem para reconhecimento de locutor mas não para reconhecimento de voz. Pode-se intuir que isso deve-se ao fato de os parâmetros do sinal glotal mostrarem características intrínsecas do locutor.

-Para a tarefa de classificação observaram-se melhores resultados quando foram usadas as redes neurais. Mas a rede neural precisa levar em conta muitos fatores para seu desenho e de muitas provas para construir uma configuração da rede que apresente um bom desempenho. Por outro lado, o SVM apresenta uma configuração mais consistente e os parâmetros que são determinados pelo usuário são poucos. Observou-se que a função Kernel que apresenta os melhores resultados ao ser usada foi a RBF, comprovando o que diz a literatura

-A melhor configuração do vetor híbrido para a tarefa de reconhecimento de locutor, como mostraram todas as experiências, foi, o formado por 12 coeficientes MFC e sua primeira e segunda derivada. Isso deve-se ao fato de o vetor ter maior número de características de cada locutor e ter maiores fatores discriminantes. Observa-se um melhor desempenho do parâmetro Ko em quase todas as experiências, sendo o maior parâmetro discriminante do sinal glotal.

-A estimativa do sinal glotal feita pelo método PSIAIF mostrada neste trabalho, teve um modelo com um pouco mais de ruído para os locutores femininos. Chegou-se a conclusão de que o modelo PSIAIF pode apresentar problemas para frequências altas. O desempenho da tarefa de reconhecimento de locutor também foi menor no caso da base que combinava locutores masculinos

e femininos comparada com a base solo para locutores masculinos.

7.2 Trabalhos futuros

Apresentamos, a seguir, algumas sugestões de trabalhos futuros para dar continuidade a pesquisa.

-Implementar a técnica SVM para reconhecimento de locutor, considerando um vetor hibrido como entrada, em caso de sinais de voz com ruído.

-Implementar um classificador híbrido envolvendo SVM ou redes neurais e modelos de Markov, tomando como entrada o vetor hibrido de características.

Referências Bibliográficas

[1] Majewski W., Basztura C.,"Integrated approach to speaker recognition in forensic applications", Forensic Linguistics 3 (1), pp.50-64, 1996.

[2] Davis K. H."Automatic Recognition of Spoken Digits", Journal of the Acoustical Society of America, vol. 24, no. 6, pp. 637-642, 1952.

[3] Koenig W., "The Sound Spectrograph", Journal of the Acoustical Society of America, vol. 17, pp. 19-49, 1946.

[4] Noemi D. B., " The glottal closure in diagnostic of minor structural alterations ", Rev. Bras. Otorrinolaringol. vol.70 no.4 São Paulo July/Aug. 2004.

[5] Gobl C. e Chasaide A. N., " The role of voice quality in communicating emotion, mood and attitude ", Speech Communication, vol. 40, no. 1-2, pp. 189-212, 2003.

[6] Fourcim A J., Maddieson I., "Laryngographic assessment of phonatory function",ASHA Rep. 11, pp. 116-124, 1981.

[7] Pedersen M. F., " Electroglottography compared with synchronized stroboscopy in normal persons ", Folia Phoniatr. 29, pp. 191-199, 1977.

[8] Mattos J. S. , Silva D. G., Apolinário J. A. e Cataldo Edson. ,"Incursionando pelos domínios da eletroglotografia: proposta de um corpus EGG", xxvi simpósio brasileiro de telecomunicações - SBrT 2008, 02-05 de setembro de 2008, Rio de Janeiro, RJ.

[9] Fabre P., "Sphygmographie par simple contact d électrodes cutanées, introduisant dans l arterè de faibles courants de haute fréquence détecteurs de ses variations volumétriques", Comptes Rendus Soc. Biol., vol. 133, pp. 639-641, 1940.

[10] Baken R. J., "Electroglottography ", Journal of Voice, vol. 6, no.2, pp. 98-110, 1992.

[11] Alku P., "Glottal wave analysis with Pitch Synchronous Adaptive Inverse Filtering", Speech Communication, vol. 11, pp. 109-118, 1992.

[12] Gajic B., Paliwal K., "Robust Speech Recognition Using Features Based On Zero Crossing With Peak Amplitudes.", ICASSP 2003, (2003), 64-67.

[13] Cuadros C. R., "Comparacão entre as tecnicas de MFCC e ZCPA para reconhecimento robusto de locutor em ambientes ruidosos ", Dissertação de Mestrado em Engenharia de Telecomunicações, UFF 2007.

[14] Haykin S., "Redes Neurais Princípios e pratica",2a ediçao Porto Alegre Bookman 2001.

[15] Juliano S. M., "Um estudo comparativo entre o sinal electroglotográfico e o sinal de voz", Dissertação de mestrado em Engenharia de Telecomunicações, UFF 2008.

[16] Rabiner L. R., Juang B., “Fundamentals of Speech Recognition”, Prentice Hall, p. 493, 1993.

[17] Kent R., Read C., “The Acoustic Analysis of Speech”, Singular Publishing Group, 1992.

[18] Fant, G., Acoustic Theory of Speech Production, Mouton, The Hague, 1960.

[19] Fabiana S., “configuração do trato vogal supraglotico na produção das vogais de portugues brasileiro”, Dissertação de mestrado em linguistica aplicada, PUC-SP 2006.

[20] Juang B. H.; Rabiner L. ;Wilpon J., “On the use of bandpass filtering in speech recognition”, Acoustic speech and signal processing IEEE transactions on V. 35, p 947-954 jul 1987 .

[21] Stevens S. S., Volkman J., “The relation of pitch to frequency”, American Journal of Psychology, vol. 53, p. 329, 1940.

[22] Oppenheim A. V., Schafer R. W., “Discrete-Time Signal Processing”. Englewood. Cliffs, NJ: Prentice Hall, p. 796, 1989.

[23] Sierra A. B. “Aprendizaje automatico: conceptos basicos y avanzados: aspectos practicos utilizando el software Weka ”,1a edição Prentice Hall 2006.

[24] Young S., Evermann G., Gales M., “The HTK Book (for HTK Version 3.3)”, Cambridge University Engineering Department, p. 354, 2005.

[25] Cataldo E., Rubens S., Nicolato L., “Uma Discussão sobre Modelos Mecânicos de Laringe para Síntese de Vogais”, ENGEVISTA, v. 6, n. 1, p. 47-57, abr. 2004

[26] Van den Berg, J.,"Myoelasticaerodynamic theory of voice production", Journal of Speech and Hearing Research, vol.1, pp. 227- 244, 1958.

[27] Titze, I. R., "Comments on the myoelastic-aerodynamic theory of phonation", The Journal of the Acoustical Society of America, vol. 23, pp. 495-510, 1980.

[28] Gauffin J., Hertegard S., Lindestad A., "A comparison of subglottal and intraoral pressure measurements during phonation", Journal of Voice, vol. 9, pp. 149-155, 1995.

[29] Hertegard S., Gauffin J. e Karlsson, I., "Physiological correlates of the inverse filtered flow waveform", Journal of Voice, vol. 6, no. 3, pp. 224-234, 1992.

[30] Sodersten, M., Hakansson, A. e Hammarberg, B., "Comparison between automatic and manual inverse filtering procedures for healty female voices", Logopedics Phoniatrics Vocology, vol. 24, pp. 26-38, 1999.

[31] Pulakka H.," Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography", Helsinki University of Technology, Dept. of Computer Science and Engineering, 2005.

[32] Arbib M. A., "The Handbookof brain theory and neural networks ", Cambridge MA: The MIT Press,1985 1118p.

[33] http://www.ee.ic.ac.uk/hp/staff/ dmb/voicebox/voicebox.html

[34] Deller J. R., Hansen J. H., ProakisJ. G., "The relation of pitch to frequency", IEEE Press, p.936,2000."

[35] http://audacity.sourceforge.net/

[36] www.cs.waikato.ac.nz/ml/weka/

[37] http://aparat.sourceforge.net/index.php/

[38] Zebulum R., Vellasco M.“ a comparison of different spectral analysis models for speech recognition using neural networks”0-7803-3636-4197 1997 IEEE

[39] Smola A. J., Schölkopf B. Learning with Kernels. The MIT Press, Cambridge, MA, 2002.

[40] Vapnik V. N. The nature of Statistical learning theory. Springer-Verlag, New York, 1995.

[41] Lorena A. C., Carvalho A. Uma Introdução às Support Vector Machines. Universidade Federal do ABC, São Pablo, 2007.

[42] http://www.isis.ecs.soton.ac.uk/resources/svminfo/

[43] Al-Jaroudi A., Makhoul J. , Discrete all-pole modeling IEEE transaccion on signal processing vol. 39 no2 pp 411-423.