UniCEUB - Centro Universitário de Brasília
FACETS – Faculdade de Tecnologia e Ciências Sociais Aplicadas
Curso de Engenharia de Computação

RECONHECIMENTO DE VOCÁBULOS UTILIZANDO REDES NEURAIS

MARCOS VINÍCIUS SANT'ANA DIAS

BRASÍLIA – DF
2008

MARCOS VINÍCIUS SANT'ANA DIAS

RECONHECIMENTO DE VOCÁBULOS UTILIZANDO REDES NEURAIS

Monografia apresentada ao
Centro Universitário de
Brasília, para obtenção do
título de Bacharel em
Engenharia de Computação.

Orientador: M.C. Aderlon Marcelino Queiroz

Brasília – DF
Julho de 2008

Agradecimentos

Dedico este trabalho a meus familiares, sabendo se tratar de diminuta retribuição diante do oceano de carinho e companheirismo no qual tenho sido generosamente banhado.

Também dedico a minha namorada, Juliana, que sempre me apoiou em todas as fases deste projeto.

O número de pessoas a quem devo prestar meus sinceros agradecimentos é enorme. Mas resumo aos mestres pela excelência do ensino; aos colegas porque importantes veículos de informação e camaradagem.

Resumo

Este trabalho apresenta a constituição de um sistema básico de reconhecimento de voz em língua portuguesa, utilizando como ferramenta as Redes Neurais Artificiais (RNA's).Na execução do projeto foi utilizado o Mapa de Kohonen, por se constituir em treinamento não-superviosionado.Como bases de dados para treinamento das RNA's foram selecionadas 20 (vinte) vocábulos que poderão ser reconhecidos pelo sistema. O pré-processamento de sinais foi realizado utilizando-se filtros e a Transformada Rápida de Fourier. Embora o treinamento possa ter alcançado resultado expressivo, não se pode negar a dificuldade enfrentada na realização do pré-processamento do sinal, decorrente do alto grau de complexidade referente à extração de fonemas de palavras faladas. Dada a grande importância do tema reconhecimento de voz, a utilização de redes neurais artificiais tem ainda um papel central, merecendo a atenção da comunidade de informática. Assim, este trabalho tem como objetivo sensibilizar essa comunidade e contribuir para dinamizar a atividade neste domínio.

Palavras-chave: redes neurais artificiais, programação em Matlab, reconhecimento de padrões, processamento digital de sinal

Abstract

This paper presents the establishment of a basic system of voice recognition in Portuguese, using Artificial Neural Networks (ANN). In implementing the project was used the map of Kohonen, because they provide a self-learning. As database for training of the ANN were selected 20 (twenty) words that could be recognized by the system. The pre-processing of signals was performed using the filters and Fast Fourier Transform. Although the training has achieved significant results, there is no denying the difficulty faced in implementing the pre-processing of the signal, due to the high degree of complexity relating to the extraction of phonemes of spoken words. Given the great importance of the subject of voice recognition, the use of artificial neural networks still has a central role, deserving the attention of the community of IT. Therefore, this paper aims to raise awareness that community and help stimulate activity in this area.

Keywords: Artificial Neural Networks, programming in Matlab, recognition of standard, digital signal processing

SUMÁRIO

Lista de figuras

Lista de tabelas

1. Introdução

Quando o computador conseguir entender e identificar a voz humana e obedecer a comandos verbais, mais um capítulo da revolução da informática será escrito pelo homem.

Com um programa com essa função, o usuário pode emitir um comando de voz para o computador e ditar um texto. Ou então, o usuário pode fazer restrições de uso a determinadas informações e através de sua voz, ter acesso a elas.

Essa tecnologia visa facilitar o uso do computador, como também auxiliar pessoas com distúrbio da fala, da linguagem ou audição. Os benefícios reais aparecem em lugares e situações que não se imaginava até então, podendo ser citado como exemplos:

a) Mãos livres para outras tarefas – Como não é mais necessário sentar na frente do computador para digitar, dá maior mobilidade ao usuário, pois fica livre para fazer outras tarefas, como por exemplo, realizar pesquisa ou ler um texto enquanto está ditando.

b) Facilitação para entrada de dados – O reconhecimento de voz permite uma taxa de escrita que calcula a média em torno de 45 a 65 palavras por minuto, sendo que alguns usuários alcançam pontuações tão altas quanto 90 palavras por minuto. Um usuário que dê a entrada de dados pode digitar 80 palavras por minuto, mas não pode fazer isso o dia todo sem pausar para caminhar e descansar, então as taxas de digitação e reconhecimento de voz pode ser quase as mesmas. Além disso, o reconhecimento de voz quase sempre acha a palavra exata, quando sua

soletração é certa.

c) Proteção contra danos físicos – Por não digitar ou utilizar um mouse, pelo menos não muito freqüentemente, o usuário será menos propenso a sofrer Lesão por Esforço Repetitivo (LER) ou Distúrbios Osteomusculares Relacionados ao Trabalho (DORT).

d) Permitir acesso ao computador a pessoas com limitações físicas – Trata-se de importante ferramenta na área de inclusão digital. Serve para aqueles usuários que não podem digitar, ou tenham problemas ao utilizar um teclado.

e) Identificação de locutores – Para ambientes que necessitem de acesso seguro.

Referente à aplicação final do processamento digital de sinais de voz, a figura 1.1 mostra variados setores de intenso interesse:

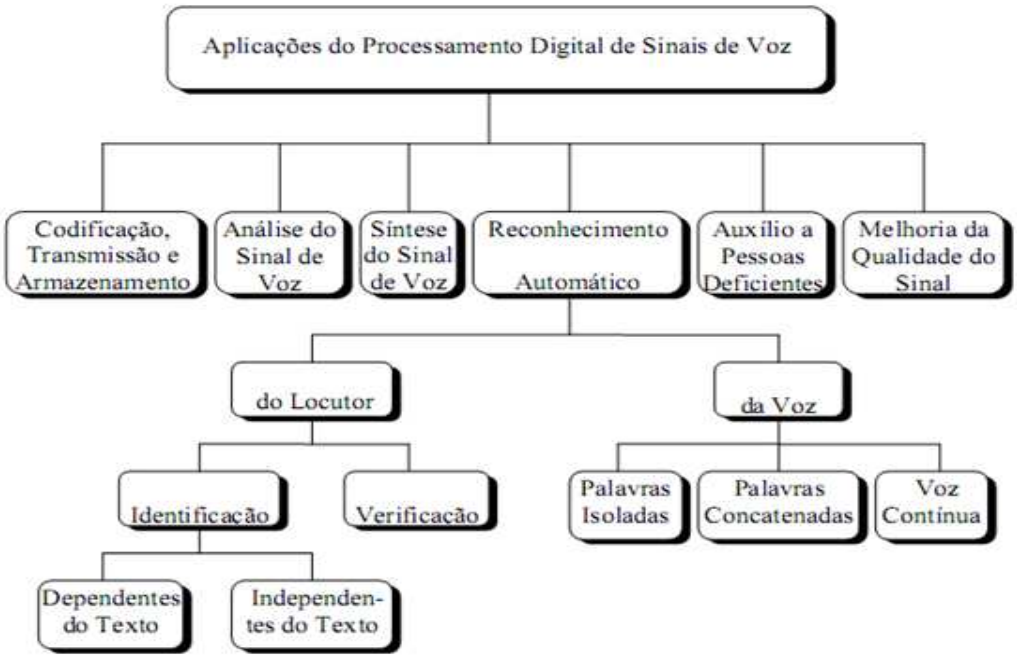

Figura 1.1 – Aplicações do processamento de voz

O projeto não apresenta uma ferramenta completa para reconhecimento de vocábulo/locutor, mas um estudo para auxiliar projetos futuros, até porque, apesar dos esforços e intensas pesquisas, não foi possível confirmar a existência de máquina capaz de compreender e identificar um discurso falado de uma pessoa dentro de um ambiente de vários falantes.

Delimitado o campo da experiência, o problema a ser enfrentado exige os seguintes procedimentos: aquisição do sinal de voz, extração de parâmetros e reconhecimento de padrões.

A aquisição objetiva apenas a captura do sinal sonoro através de uma interface analógica digital, que consiste basicamente em entrar com a informação (voz) de forma analógica e recolher na saída essa mesma informação de forma digital. Tudo isso ocorre utilizando-se o microfone e a placa de som do computador.

A extração abrange o processo digital do sinal, para que dele possam ser destacados parâmetros que servirão para uma futura comparação.

Por seu turno, o reconhecimento do padrão consiste em identificar os dados selecionados na fase de extração e fazer um treinamento. Em outras palavras, consiste em fazer uso das informações catalogadas sob a forma de padrões, os quais serão utilizados na construção de um modelo capaz de reconhecer novas ocorrências de um conjunto pré-determinado de palavras.

O objetivo inicial deste trabalho foi de desenvolver um sistema que permitisse reconhecer determinada palavra de um conjunto de 20(vinte), mediante utilização de técnicas de processamento digital de sinal e de padronização por modelos neurais. Entretanto, devido ao modo em que se ia desenvolvendo o projeto, acabou-se tendo como objetivo final um reconhecedor de locutores.

1.1. Perfil Histórico

Datam do final da década de 1950 as primeiras pesquisas tecnológicas para o reconhecimento de voz. Bell, Davis, Biddulph e Balashek construíram um sistema para reconhecimento de um dígito isolado falado por uma pessoa. (BARBISAN,2008)

Em 1964, a IBM apresenta um sintetizador de voz para a fala de dígitos; e em 1978, foi lançado pela Texas Instruments o primeiro chip dedicado à síntese de voz.

Contudo será em 1993 que a IBM lança o primeiro software comercial para reconhecimento de voz, que foi denominado de "IBM Personal Dictation System, para OS/2). Neste mesmo ano, a Apple apresenta conjunto de rotinas para Mac, para reconhecimento e síntese de voz. No Brasil, ainda nesse ano, a Universidade Federal do Rio de Janeiro desenvolve DOSVOX, com síntese de voz em português, para deficientes visuais usarem PC's com DOS.

O que se constatou na década de 70, no tocante ao reconhecimento de voz, é que a meta era reconhecer uma palavra independente; já na década 80, a meta era reconhecer a fala fluente de frases.

Das pesquisas efetuadas nos anos 80, resultaram as aproximações estatísticas, citando-se aqui o modelo de Hidden Markov.

Também é dessa década a aplicação de redes neurais em problemas de reconhecimento de voz. (BARBISAN,2008)

A partir da década de 90, pesquisas foram efetuadas na busca de um sistema de reconhecimento de voz contínuo. Dentre os sistemas conhecidos vale

citar o denominado DARPA (Defense Advance Reseach Projects Agency), que foi desenvolvido para reconhecer continuadamente sem erros palavras dentro de um arquivo de 1000 palavras.

Além disso, vale ser mencionado o sistema Voice da Dragon Systems (1994), que contém o Dragon Dictate para ditados, e o MedSpeak/Radiology, da IBM (1996), primeiro produto para reconhecimento da fala em tempo real, como também o VIA VOICE, para fala contínua, lançado em 1997. (BARBISAN,2008)

Em língua portuguesa, a IBM lança o VIA VOICE (1998); a MicroPower, o Delta Talk (sintetizador de voz em português); e a Philips, o FREE SPEECH 2000 (reconhecimento da fala em português).

1.2. Áreas de conhecimento necessárias

O uso de uma única classe de ferramenta ou metodologia específica não permite responder a todos os problemas que se apresentam durante a experiência.

As áreas de conhecimento aqui adotadas podem ser assim sistematizadas(BRAGA,2006):

a) Processamento de Sinais: a comunicação humana se desenvolve através do uso de sinais da fala. Formalmente um sinal é definido como uma função de uma ou mais variáveis, a qual veicula informações sobre a natureza de um fenômeno físico. Em um sinal, sempre há um sistema associado à sua geração, e outro associado à extração da informação do sinal.

b) Aspectos físicos (acústica): A compreensão da fala humana requer

conhecimentos de diversas áreas da ciência como, por exemplo, fonética, fonologia, lingüística, acústica, mecânica, matemática, computação, fisiologia e neurociência. Profissionais e estudantes que dominem uma ou algumas destas áreas devem almejar, pelo menos, compreender o conteúdo das demais áreas, para que possam analisar a comunicação falada de maneira ampla e, ao mesmo tempo, profunda.

c) Padrões de reconhecimento: é a área de pesquisa que tem por objetivo a classificação de objetos (padrões) em um número de categorias ou classes. Com o conjunto de algoritmos usados, podem ser agrupados dados e criados um ou mais padrões de conjunto de dados, que serão posteriormente comparados com um sinal qualquer para efeito de reconhecimento.

d) Análise de padrões: ao fazer análise, tentamos entender o domínio do problema, tendo como forma de resolução o conjunto de procedimentos para estimação de parâmetros de modelos estatísticos.

e) Lingüística: tendo-se em vista as relações entre os sons (fonologia), palavra de uma linguagem (sintaxe), significado das palavras faladas (semântica) e o sentido derivado do significado.

f) Fisiologia: para melhor compreensão do sistema nervoso central do ser humano incluindo a produção e percepção da fala.

Reiterando o que já foi dito acima, o objetivo deste trabalho é de programar um sistema de reconhecimento de locutores, englobando como uma das ferramentas significativas as redes neurais artificiais.

Este trabalho está disposto em oito capítulos. No primeiro capítulo são abordados os aspectos introdutórios do trabalho, como a apresentação do tema, o perfil histórico e considerações prévias a respeito da matéria. O capítulo 2 trata da voz humana e do mecanismo de fonação, demonstrando a estrutura e composição do aparelho fonador. No capítulo 3 examina-se o tema a interferências na fala, enquanto no capítulo 4 são examinados os parâmetros utilizados do sinal de voz: transformada de Fourier, predição linear, energia do sinal de voz, cruzamento zero e detecção dos pontos extremos No capítulo 5 é feita uma introdução sobre as redes neurais artificiais, dispondo sobre seus elementos formadores, tratando das funções de ativação, topologia de uma rede e tipos de treinamento, fazendo considerações ainda sobre as redes de Kohonen.

Quanto à implementação do modelo, a questão é tratada no capítulo 6, com a indicação do software escolhido e das vantagens do uso do Matlab. O capítulo 7 alinha os resultados obtidos e o capítulo 8 apresenta as considerações finais e conclusões do trabalho.

2. A voz humana

2.1. Aparelho fonador

O ser humano não possui um órgão específico para produzir a fala, assim como a visão e a respiração. São vários órgãos que trabalham em conjunto no que resulta a produção sonora. (MARTINS, 1988). Os órgãos que compõem o aparelho fonador são: os brônquios, a traquéia, a laringe, a faringe, as fossas nasais e a boca com língua, as bochechas, o palato duro, o palato mole, conforme a figura 2.1.

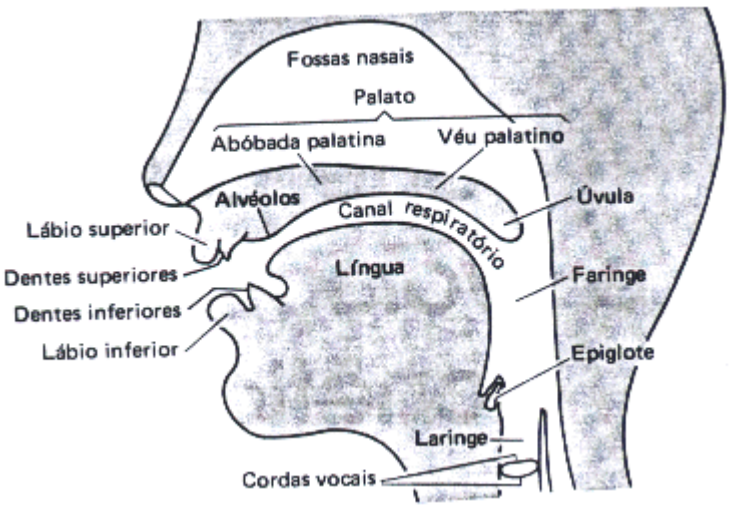

Figura 2.1 - Órgãos que compõem o aparelho fonador

No português, os sons falados são produzidos pelo ato de expiração, ou seja, a saída da corrente de ar dos pulmões. Existem algumas línguas em que os fonemas são produzidos com a inspiração (entrada de ar nos pulmões).

Os sons da fala não devem ser confundidos com os fonemas da língua portuguesa, uma vez que o som é entendido como uma complexa realidade físico-

acústica de cada unidade sonora da fala, enquanto que os fonemas correspondem à percepção eclética e interpretativa realizada pelo falante e ouvinte, respectivamente. (LUFT,89)

A corrente de ar originado pelos pulmões passa pela traquéia e chega à sua extremidade, que é a laringe (também conhecido como pomo-de-adão). Na laringe se encontram as cordas vocais (duas superiores e duas inferiores), que são formadas por uma prega mucosa elástica. Apenas as cordas vocais inferiores são consideradas para a reprodução do som. Entre as cordas vocais situa-se a glote. Nela ocorre a passagem da corrente de ar que chega à faringe. Deste ponto, podem-se seguir dois caminhos: sair totalmente pela boca ou parcialmente pela boca e parcialmente pelo nariz. (MARTINS, 1988).

Essa descrição permite deduzir que há uma fonte de som variável e um filtro acústico também variável que altera o som da fonte. Ao lado disso, quando é impulsionado o aparelho fonador pela fala, isso representa a emissão de sons articulados na linguagem oral enquanto que a produção da voz é a fonação.

2.2. O mecanismo da fonação

Inicialmente, é importante assinalar que o computador não entende a nossa linguagem. Em conseqüência, é preciso de alguma forma converter as palavras em algo que pode ser processado e compreendido pela máquina.

Quando o usuário fala, é feita a conversão do sinal analógico para digital por meio da placa de som do próprio computador, criando-se assim um fluxo de dados digitais que é interpretado pelo software.

A forma como o som é digitalizado depende do programa que foi

preparado e do modelo utilizado (lingüístico ou acústico).

No modelo lingüístico é analisado o conteúdo da fala para em seguida se proceder à comparação das combinações de fonemas com as palavras contidas na base de dados, utilizado pelo programa.

No modelo acústico é feita a análise dos sons da voz do locutor, os quais são convertidos em fonemas (elementos básicos da fala).

No caso da língua portuguesa, embora estejam catalogados 28 fonemas, no Brasil, devido às variações de pronúncia que existem entre regiões distintas, podem ser registrados até 36 fonemas.

A propósito é notável o uso de diversas linguagens para comunicação do ser humano, seja através de códigos, gestual, escrita, ou oral.

A linguagem escrita é bastante utilizada para preservação e difusão do conhecimento e da cultura. Atribui-se também à escrita um ponto importante para evolução do ser humano. Isso ocorre devido à importância dessa forma de linguagem nas sociedades bem estruturadas e num mundo globalizado, já que por ela podem ser descritos fatos e idéias, os quais podem ser transmitidos a longas distâncias. Apesar disso, é inegável a importância da linguagem oral. Embora se considere a escrita como recente em comparação com a utilização da fala na história humana, mesmo assim é através do discurso oral que o homem pode estabelecer um conectivo entre suas idéias, seu imaginário e o mundo real.

Dadas as diferenças de modelos, para que se compreenda o funcionamento do processo de voz, é necessário examinar alguns elementos da fonética.

Existem duas ciências que tratam do estudo da fala: a fonética e a fonologia. A fonética estuda os sons da fala humana e a fonologia estuda a representação mental de sons como parte do sistema cognitivo simbólico. (LUFT, 89)

Em face desses conceitos, pode-se afirmar que a fonética estuda a manifestação física da linguagem em ondas sonoras, como estes sons são articulados e percebidos, além disso, trata das propriedades físicas de sons da fala ao estudar a natureza física da produção e da percepção dos sons da fala. Por sua vez, a fonologia trata da organização de sons da fala, aspecto mental de língua e métodos da ciência humana. Será então pela fonologia que serão identificados os sons predizíveis e os sons não-predizíveis, assim como o ambiente que permita prever a ocorrência de alguns sons, como aqueles que afetam o significado das palavras.

Neste passo, torna-se necessário para compreensão da pesquisa, examinar alguns aspectos da fonética lingüística e da fonética articulatória, tendo em vista que os sons de nossa fala resultam quase todos da ação de certos órgãos sobre a corrente de ar vinda dos pulmões.

2.3. Fonema e letra

Fonema é o nome que se dá aos sons da fala e é a menor unidade de som da fonética (CALLOU, 1995).

As letras são representações gráficas dos fonemas. É importante não confundir letra com fonema

Nem sempre há uma correspondência numérica entra as letras e os fonemas. Por isso foi criado um sistema de signos no qual a cada fonema corresponde um único signo. É o alfabeto fonético.

Cada fonema da língua é produzido de maneira diferente. Essas diferenças determinam suas características acústicas, denominadas traços.

Em decorrência da tradição etimológica, podem-se observar, na representação dos fonemas portugueses, imperfeições da seguinte ordem (SACCONI, 1989):

a) a mesma letra pode representar fonemas diferentes: e**x**ame, **x**ale, pró**x**imo, se**x**o; **c**ola, **c**era;

b) o mesmo fonema pode ser figurado por letras diferentes: ca**s**a, e**x**ílio, co**z**inha, ti**g**ela, **laje**;

c) um fonema pode ser representado por um grupo de duas letras (dígrafo): ma**ch**ado, mu**lh**er, u**nh**a, mi**ss**a, ca**rr**o;

d) usam-se letras simplesmente decorativas: não representam fonemas nem funcionam com notações léxicas. Mantiveram-se em razão da etimologia: **h**otel, e**x**ceção, q**u**ina, di**s**cípulo;

f) Há fonemas que, em certos casos, não se representam graficamente: falam (fálãU).

Algumas vezes um único fonema pode ter mais de uma forma. Na linguagem falada no Brasil existem variações de fonemas de uma região para outra. A palavra dia pode ser diferente da região sul para região nordeste ou para sudeste. Enquanto no sul se pronuncia o ***d*** mais forte, na região nordeste este **d** se aproxima

ao som do **t** e no estado do Rio de Janeiro, este **d** pode soar um pouco chiado.

Nesse caso não existe a criação de um fonema e sim a variação dele. Essa variação é chamada de alofone.

2.4. Representação dos fonemas

Na língua escrita, os fonemas são representados por signos ou sinais gráficos chamados letras. O conjunto de letras denomina-se alfabeto ou abecedário. Como não há uma correspondência biunívoca entre letras e fonemas, foi criado um sistema de signos no qual a cada fonema corresponde um único signo, que passou a ser considerado o alfabeto fonético.

No português falado no Brasil ainda não existe uma solução específica para fazer a representação de fonemas, sendo ainda utilizado na transcrição o Alfabeto Fonético Internacional (IPA). (FARIA, 1996)

2.5. Classificação dos fonemas

Os fonemas são classificados em vogais, consoantes e semivogais. (SACCONI, 1989).

Vogais são os sons que chegam livremente ao meio externo sem a interferência dos órgãos (a-é-ê-i-ó-ô-u).

Consoantes são ruídos provenientes da resistência que os órgãos bucais opõem à corrente de ar: bola, copo, depósito.

Semivogais são os fonemas /i/ e /u/ átonos que se unem a uma vogal: vai – andei – ouro – água.

A importância dos fonemas no presente estudo é precisamente quanto a sua aplicação no modelo acústico, já que por ele, são analisados os sons de quem os emitiu e convertidos em fonemas dos elementos básicos da fala.

3. Tipos de interferências na fala

3.1. Introdução

A voz pode ser descrita por modelos parametrizados e o reconhecimento é executado levando-se em conta qual seqüência de palavras é a mais provável de produzir dados acústicos determinados de acordo com o modelo (BECCHETTI, 1999).

Os parâmetros de modelo são variáveis aleatórias cujos valores esperados devem ser estimados das amostras observadas, isto é, dos dados acústicos. Os dados de voz são caracterizados pela ampla variabilidade, devido à natureza da fala, ao tom, à expressão vocal, etc. Isso significa que o espaço do parâmetro deve refletir a maioria das variabilidades da voz e, além disso, deve possuir uma dimensão enorme.

Aplicações mínimas requerem aproximadamente 10^5 parâmetros para um reconhecimento de fala contínua que tenha múltiplos locutores. Sabe-se que a exatidão da avaliação de cada parâmetro pode ser considerada proporcional ao inverso do número de observações. Desta forma, um vasto número de amostras e de bancos de dados correspondentes é obrigatório para uma etapa precisa da avaliação de parâmetros que é chamado de “fase de treinamento”.

A fase de treinamento de um sistema de reconhecimento é similar ao processo de aprendizagem de um bebê. Uma criança deve experimentar um fenômeno diversas vezes e com uma ampla variabilidade antes de ser capaz de reconhecê-lo. A tecnologia atual dos sistemas automáticos de voz não permite uma implementação em tempo real de modelos comparáveis à complexidade humana.

Isso significa que a variabilidade de voz deve ser limitada para atingir resultados adequados. Assim, os sistemas de reconhecimento são desenvolvidos para trabalhar com aplicações específicas (e deste modo, limitadas).

3.2. Variabilidade da fala

As fontes diversas de variabilidade que podem afetar a voz determinam a maioria das dificuldades no reconhecimento de voz. Durante a produção da voz, os movimentos de diferentes articuladores sobrepõem o tempo por segmentos fonéticos consecutivos e interagem com cada um. Como conseqüência, a configuração do trato vocal em qualquer tempo é influenciada por mais um segmento fonético. Esse fenômeno é conhecido como co-articulação. O efeito principal da co-articulação é que o mesmo fonema pode ter características acústicas bastante diferentes, dependendo do contexto no qual elas são pronunciadas. (BECCHETTI, 1999)

A produção de voz é afetada não somente pelo fonema de co-articulação, mas também pelos números de fontes de variação, como regional, social, estilístico e individual. Uma série de variáveis “ambientais”, como barulhos ao fundo, eco e condições de gravação têm também que ser levado em conta, de como contribuem para produção de mudanças na propriedade do sinal de fala. Os resultados destas variabilidades afetam todos os níveis de comunicação de voz, desde sons até a gramática e ao léxico. Essencialmente, cada produção de fala é única e esta faz com que o processo natural e automático de decodificação seja quase difícil.

Classificações diferentes têm sido propostas na literatura para determinar algumas categorias largas para essas variações. Entretanto, é muito difícil esboçar uma linha entre uma categoria e outra, porque na maioria do tempo há uma interação entres as fontes de variação. Por isso, é proposta uma classificação principal baseada na diferenciação entre a variabilidade intralocutor e interlocutor.

3.3. Variabilidade do intralocutor

A fala individual não é uma constante em termos de que qualquer uma daquelas propriedades que resulta principalmente de mecanismos físicos da voz, possa ser repetida com precisão.

A voz de uma pessoa pode variar por causa dos momentos e das amplitudes diferentes de articuladores de fala envolvidos na sua produção. Então, a mesma palavra ou a mesma pronúncia, produzida pelo mesmo locutor em diferentes ocasiões e em diferentes condições, podem ser substancialmente diferentes. Além disso, o mecanismo físico da voz sobre mudanças, que podem afetar a ressonância da cavidade nasal e o modo de vibração das cordas vocais. Isso é evidente, por exemplo, como uma conseqüência de uma patologia da laringe, quando uma pessoa tem um resfriado. Menos óbvias, são as mudanças na freqüência fundamental e tipo de fonação, que ocorrem devido a fatores como cansaço e estresse, e em longo prazo devido ao envelhecimento. (BECCHETTI, 1999)

A voz difere também em escolhas feitas no uso dos recursos do mecanismo da linguagem. Essas escolhas variam de mudanças de barulho e tom, para mudanças do sistema fonológico segmentário. Mudanças no barulho, escala de

tom e tipo de fonação são aleatórios, por exemplo, em caso de grito sobre uma conexão fraca de telefone ou de um aumento no tom de voz em um ambiente ruidoso, ou ainda para que possa ocorrer a transmissão de estados emocionais, como: raiva, alegria, medo, etc.

O sistema de fala está também sujeito à variação como têm mostrado estudos sociolingüísticos. A maioria dos indivíduos é capaz de controlar uma escala de pronuncia ao longo de uma duração de clareza fonética. Também são capazes de adaptar sua pronúncia para uma situação comunicativa na qual uma interação possa acontecer. A variabilidade da voz é evidenciada pela experiência do dia-a-dia. O estilo de falar pode ser facilmente mudado de formal para informal. Mudanças na velocidade de elocução, no tom da voz, na atitude desembaraçada ou tímida, de acordo com a situação contextual, são freqüentes e aumentam a variabilidade.

3.4. Variabilidade do interlocutor

As diferenças entre locutores encontram-se na anatomia do trato vocal e cordas vocais, nos dialetos e /ou pronunciações regionais, na classe social, no nível de educação, na idiossincrasia da fala. Pode ser útil esboçar uma distinção explícita entre duas fontes maiores na variação da produção de voz, no qual tem sido classificada como fonético versus orgânico.

Variações fonéticas resultam de diferenças no modo que uma pessoa usa o aparato vocal, enquanto diferenças orgânicas dependem da diferença individual da forma e proporção dos órgãos utilizados na produção vocal. Uma das maiores

diferenciações orgânicas evidentes é entre a voz feminina e a masculina. (BECHETTI, 1999)

As mulheres geralmente têm um tom maior que os homens devidos aos tamanhos diferentes das cordas vocais (23-25 mm nos homens, 15-17 mm nas mulheres). O volume alto pode afetar técnicas de análise espectral; de fato, volume e extração formante são geralmente menos estimados em vozes de volumes altos. Portanto, vozes femininas e infantis têm sido menos estudadas que as vozes masculinas, devido às enormes dificuldades encontradas na extração dos parâmetros acústicos, em particular nas freqüências formantes. Essa dificuldade atrasa a geração do reconhecimento automático de mulheres e crianças.

O ambiente influencia também na produção, percepção e representação acústica da voz de forma tangível. Elementos ambientais podem ser divididos em duas categorias:

- Estático (com relação à sessão de reconhecimento): sala acústica, tempo de reverberação, dispositivos de gravação;
- Dinâmico: ruídos, a posição do microfone, etc.

É freqüentemente difícil modelar os elementos ambientais estáticos e dinâmicos devido a suas imensas variabilidades. Isso porque existe uma grande diferença no desempenho entre sistemas de laboratório e sistemas usados em situações reais.

Finalmente, deve-se levar em conta que as características regionais e sócios-lingüistas do locutor podem ter um grande efeito. Pessoas falam diferencialmente de acordo a sua procedência (sua tonicidade ou dialeto) e

conforme alguns fatores, como características lingüísticas da família, status social, processo educacional. Pronúncias, acentos ou regionais diferem principalmente nas realizações de fonemas e no modelo de entonação, enquanto que o dialeto difere também na sintaxe e no vocabulário.

Essas variações lingüistas podem interferir na velocidade, na escolha de vogais e no espectro de muitos outros fonemas, assim como no modelo de entonação da fala. A variabilidade intralocutores também interage com outras fontes variáveis tais como dependência contextual na variabilidade fonética, variabilidade interlocutores, condições de gravação, ruídos do ambiente, que todos em conjunto faz da voz um sinal enriquecido e complexo.

Todas essas fontes de variabilidade resultam em mais implementações exigentes de reconhecedores de voz. Três questões básicas de aplicação devem ser levadas em conta: os locutores (treinados ou não treinados, acentuados ou não-acentuados, etc.), o ambiente (nível de ruído, largura da banda, distorção), o vocabulário (complexidade, tamanho e sintaxe). (BECCHETTI, 1999)

No sistema automático de voz, essas são geralmente questões fundamentais de aplicação que determinam a tecnologia a ser usada. São elas a segui citadas:

a) ***Dependência de locutor:*** Duas estratégias podem ser consideradas no projeto de um sistema automático de voz: o sistema pode ser dependente ou não de locutor. No primeiro caso, o sistema é projetado para reconhecer a voz de somente um locutor e ele pode ser ajustado para um locutor especifico diretamente ou depois do treinamento de locutores genéricos. No segundo caso, o sistema deve ser capaz de

reconhecer a voz de diferentes locutores. Tal sistema requer um grande treinamento específico já que tem que lidar com grande variabilidade de interlocutores.

b) ***Adaptação de locutores:*** O sistema pode aprender características de locutores atuais aumentando sua precisão durante o uso. O treinamento pode ser executado também enquanto o sistema está em uso.

c) ***Características da voz:*** A magnitude da voz pode variar de sons isolados até conversações completas; pode ser composto de palavras isoladas (isto é,quando pausas são introduzidas na pronunciação de duas palavras consecutivas), ou pode ser feito de voz continua. Neste segundo caso, os efeitos de co-articulação podem ser muito mais tangíveis. Correlacionado com modo de produção é o estilo de produção, que varia de formal a informal: quando se lê um texto, as pessoas tendem a falar mais formalmente e a articular mais cuidadosamente do que quem conversa espontaneamente.

d) ***Taxa de voz:*** A velocidade da pronúncia da palavra varia de locutor para locutor, e também do mesmo locutor, dependendo de suas condições físicas e psicológicas. A voz pode ser, portanto, rápida, normal ou vagarosa. A velocidade é normalmente mensurada levando-se em conta a média numérica dos frames de voz de uma locução vocal. Um locutor pode ser requisitado a não exceder o limite de velocidade.

e) ***Sons extralingüísticos:*** Durante a pronúncia, sinais não-lingüisticos relevantes, tais como tosses, espirros, estalos da língua, etc., são freqüentemente produzidos. Esses fenômenos podem se tratados explicitamente como partes da voz ou tratados a nível lingüístico. O projeto do sistema deve considerar a manipulação de tais aspectos.

f) ***Tamanho do vocabulário:*** O número de palavras que o sistema pode reconhecer pode variar de poucas a milhares.

g) ***Dados de treinamento:*** A fase de treinamento é um passo critico que envolve tipos diferentes de dados e fontes de informação. Os dados podem incluir formas de onda de palavras isoladas ou frases classificadas foneticamente e podem também incluir fenômenos extralingüísticos.

4. Parâmetros para o reconhecimento

4.1. Introdução

Utilizando-se de ferramentas de processamento digital de sinal, é possível formar um conjunto de parâmetros do sinal de fala. Com a escolha correta de modelos paramétricos, eliminam-se informações desnecessárias do sinal de fala, apenas enfatizando os aspectos que realmente contribuem para o projeto de reconhecimento de fala.

4.2. Transformada de Fourier.

A transformada discreta de Fourier ou DFT (Discrete Fourier Transform) é definida como:

$$X(k) = \sum_{k=0}^{N-1} x[n]e^{-j2\pi kn/n}, \qquad k = 0,1,2,\dots,N-1 \quad (4.1)$$

Onde ***N*** é o número de amostras de janelas que vão ser analisadas. E ***n*** é o número de amostras dentro de cada frame. São valores complexos, entretanto são somente considerados os valores absolutos, ou seja, sua magnitude.

A complexidade da transformada discreta está relacionada com os números de operações. Como o objetivo é de acelerar o cálculo deste procedimento, é utilizada no lugar transformada discreta de Fourier (DFT), a transformada rápida de Fourier FFT (Fast Fourier Transform). A transformada de Fourier neste trabalho é utilizada na função `specsubm`. A comparação da DFT e da FFT é mostrada na figura 4.1.

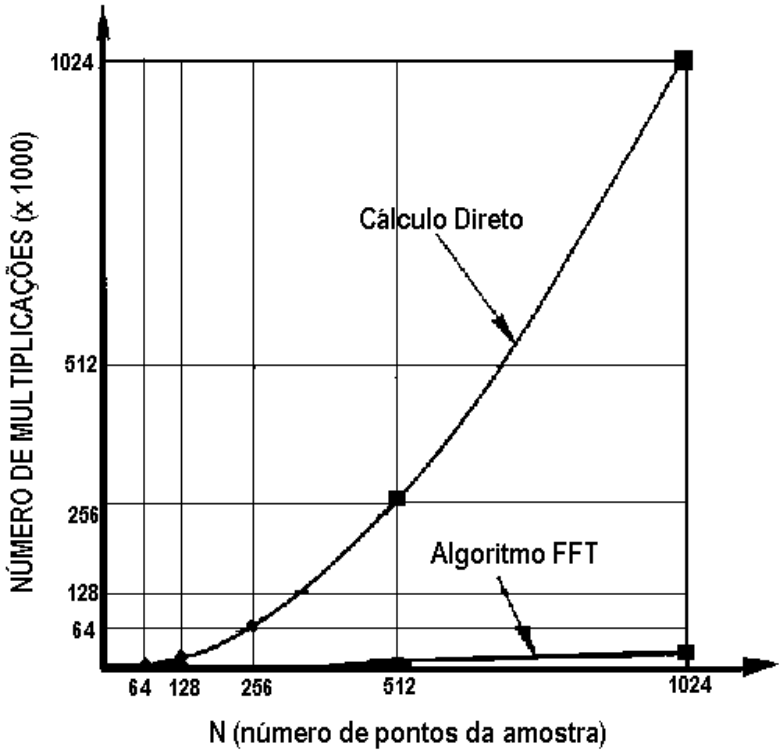

Figura 4.1 – Comparação entre a DFT e a FFT

O motivo do uso da transformada de Fourier parte do fato da utilidade que se tem em decompor o sinal de voz saída em seus componentes de freqüência. Um aspecto importante de se utilizar a transformada de Fourier com sinais de voz é que em períodos curtos de tempo se assume pelo menos que o sinal é estacionário. Na verdade isso não é deste modo que ocorre, daí se faz uma aproximação. A solução consiste em multiplicar o sinal por uma função de janelamento que seja fora de determinada faixa e reproduzir o resultado de forma que se tenha número de blocos iguais.

Janelamento

Pode-se dividir a voz em pequenas porções, que corresponderiam aos *frames* do sinal, que é feito através do janelamento. Com o uso do janelamento,

consegue-se aumentar as informações adquiridas para o projeto de reconhecimento de fala. A seguir são listados dois tipos de janelamento. (HAYKIN,2001)

O janelamento retangular é dado por

$$W_n = \begin{cases} 1 & 0 \leq n \leq N \\ 0 & resto \end{cases} \qquad (4.2)$$

Entretanto a utilização deste janelamento faz com que no inicio e no fim se tenha uma grave descontinuidade. Para reduzir o efeito da descontinuidade o mínimo que deve ser feito é empregar janelas que tendem a reduzir o 0 dos valores das amostras nos extremos.

Ainda que existam vários tipos de janelas, a mais comum para análise de voz é a de Hamming, por apresentar características espectrais interessantes e suavidades nas bordas. (BECCHETTI, 1999) Sua fórmula é:

$$W_n = \begin{cases} 0.54 - 0.46\, cos\left(\frac{2\pi n}{N-1}\right) & 0 \leq n \leq N \\ 0 & resto \end{cases} \qquad (4.3)$$

4.3. Predição linear

O método de predição linear ou LP (Linear Prediction) é historicamente um dos métodos mais importantes para a análise da voz. A idéia básica por detrás da predição linear é a de que o valor de uma amostra pode ser aproximado por

combinação linear dos valores da amostras anteriores, tirando partido da correlação entre elas.

Os coeficientes de predição linear ou coeficientes LPC (*Linear Predictive Coding*) são estimados por minimização do erro quadrático entre a amostra atual e a sua predição.

Com um número suficiente de parâmetros modelo de predição linear, pode-se constituir uma aproximação adequada à estrutura espectral de todo tipo de sons. O método de predição linear pretende extrapolar o valor da amostra seguinte de voz ***x(n)*** como a soma ponderada de amostra passadas, sendo ***n*** o número de amostras de cada *frame*:

$$\mathfrak{x}(n) = \sum_{i=1}^{k} a_i x\,[n-i] \qquad (4.4)$$

4.4. Energia do sinal de voz

Têm-se notado que a amplitude do sinal de voz varia apreciavelmente com o tempo. Em particular, a amplitude de segmentos não pronunciados é geralmente muito mais baixa que a amplitude de segmentos sonoros. A energia de tempo curto do sinal de voz fornece uma representação conveniente que reflete estas variações da amplitude. Em geral, pode-se definir a energia de tempo curto como:

$$E = \sum_{m=n-N+1}^{n} x^2\,[n] \qquad (4.5)$$

Onde ***n*** é o número de amostras em cada *frame*.

4.5. Taxa de cruzamento-zero

No contexto de sinais discretos, um cruzamento-zero acontece se amostras sucessivas tiverem sinais algébricos diferentes. É calculada conforme:

$$Z_t = \frac{1}{2}\sum_{n=1}^{N} |\, sign(x[n]) - \ sign(x[n-1])| \quad (4.6)$$

Onde ***x[n]*** é o sinal no domínio do tempo e a função ***sign*** é 1 para argumentos positivos e 0 para argumentos negativos. Resumidamente, pode-se dizer que a taxa de cruzamento-zero é o número de vezes que o sinal passa pela amplitude zero.

4.6. Detecção dos extremos

Outro fator de grande importância em um reconhecedor de fala é a necessidade de se determinar de forma eficiente e precisa o início e o final de uma locução, com a finalidade de excluir os silêncios que não trazem nenhuma informação adicional sobre a locução a ser reconhecida, evitando carga computacional e economizando tempo, além de servir como marco de início e fim de um segmento de fala. Estes fatores são de grande importância, pois minimizam a carga do reconhecedor, visto que o mesmo não terá que processar atributos de reconhecimento de trechos sem informação de fala.

A detecção dos extremos é determinada pelo primeiro quadro onde o sinal de fala realmente se inicia e pelo último quadro do sinal de fala. Deve ser feita de forma cuidadosa, pois os mínimos erros nesta estimação podem degradar o processo de reconhecimento.

5. Redes Neurais Artificiais

5.1. Introdução

É comum afirmar-se que o cérebro humano é um potente computador. Contudo, o que se sabe sobre a arquitetura do cérebro humano e o conhecimento que se tem sobre a arquitetura dos computadores, evidencia a vasta diferença entre ambas. Em relação à sensibilidade do conhecimento adquirido dos seres humanos, nota-se que é muito fácil para o homem reconhecer uma imagem ou então entender uma língua. Já com referência aos computadores, essas tarefas se tornam complexas e muitas vezes não produzem o resultado esperado. A complexidade das estruturas das redes neurais biológicas é muito maior que a dos modelos matemáticos utilizados em redes neurais, mostrando assim os obstáculos encontrados para se tentar imitar o funcionamento do sistema nervoso humano. (HAYKIN, 1999)

As redes neurais artificiais surgiram a partir do momento em que se quis que um computador simulasse a estrutura e a funcionalidade do neurônio biológico. As redes neurais propõem-se a solucionar problemas de reconhecimento de padrões, resultando em aprendizagem.

As redes neurais se constituem de um sistema altamente integrado, que trabalha de modo paralelo e distribuído com vários dados interconectados, constituídos de uma alta capacidade de processamento que processa vários algoritmos matemáticos, a fim de gravar conhecimento e utilizá-lo.

Algumas características são: (HAYKIN, 1999)

- Capacidade de aprender por meio de exemplos e de generalização; entende-se por generalização a habilidade de reconhecer elementos iguais, sendo que estes ainda não foram definidos;

- Tolerância a falhas; com a estrutura paralela, caso um neurônio falhe, os efeitos na totalidade da rede não serão sentidos no desempenho do sistema, já que por ser em paralelo, outro caminho poderá ser seguido;

- Possuir alta capacidade de adaptação;

- Robustez diante de informações falsas.

As redes neurais são boas para tarefas que exijam: (LUDERMIR, 2000) (HAYKIN, 1999)

- Reconhecimento, classificação e a associação de padrões;

- Resistência a ruído;

- Robótica;

- Processamento de sinais e imagem.

5.2. O neurônio biológico

A capacidade do cérebro humano de processar informações de forma não-linear e paralela permite solucionar problemas complexos. Esse poder de processamento ocorre por causa dos neurônios e suas conexões. Devido aos diversos estudos da área neural, a compreensão do cérebro se tornou melhor,

introduzindo o conceito de neurônios. (KOVACS, 1996)

O neurônio é composto de três partes: corpo celular e seus prolongamentos celulares, os dendritos e os axônios. (CARLSON, 2002) O corpo celular (ou soma) contém o núcleo e os processos vitais do neurônio. Os dendritos possuem ramificações parecendo “árvores”. Os neurônios comunicam entre si e seus dendritos funcionam como receptores importantes das informações, que passam de neurônio para neurônio através da sinapse. O axônio é um tubo longo, esguio, quase sempre coberto por uma bainha de mielina. Transporta as informações do corpo celular. A estrutura do neurônio é apresentada na figura 5.1.

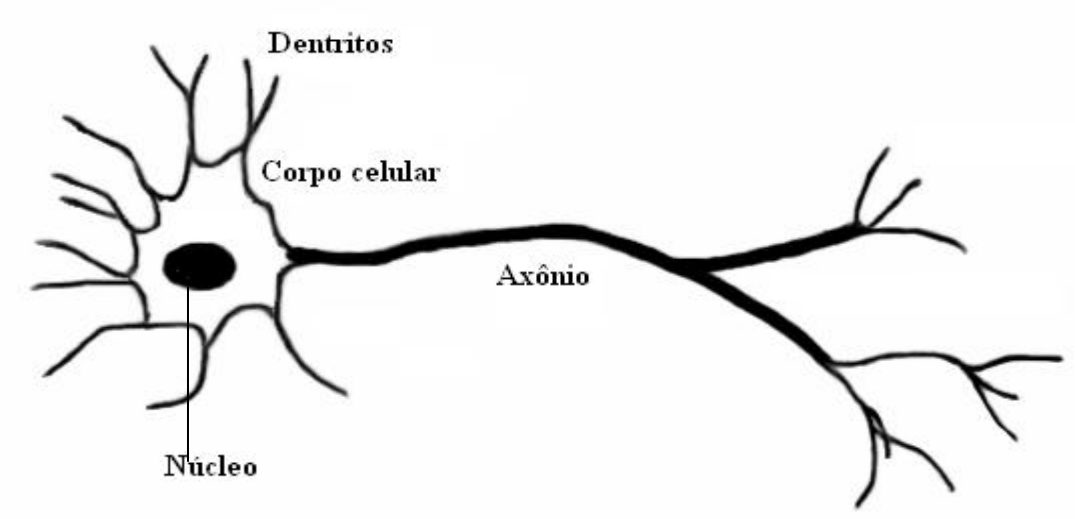

Figura 5.1 – Estrutura do neurônio.

Cada neurônio, basicamente, recebe os estímulos nos dendritos, os processa em seu corpo celular e, dependendo do seu estado de ativação resultante dos estímulos recebidos, gera e transmite um estímulo pelos axônios, para que atinja outros neurônios ou outros tipos de células, num processo altamente paralelo.

O axônio, que é considerado uma fibra nervosa, é envolvido por uma bainha de lipídica denominada de bainha de mielina. Ela funciona como um isolante elétrico e determina a condução mais rápida do impulso elétrico. Os axônios podem

variar bastante em comprimento, sendo que alguns se estendem por mais de um metro além do corpo celular. Nas extremidades dos dendritos se encontram botões terminais que contêm mediadores químicos utilizados na comunicação celular.

A sinapse é a comunicação estabelecida entre um neurônio com outros neurônios ou com outros tecidos. Como mostra a figura 5.2.

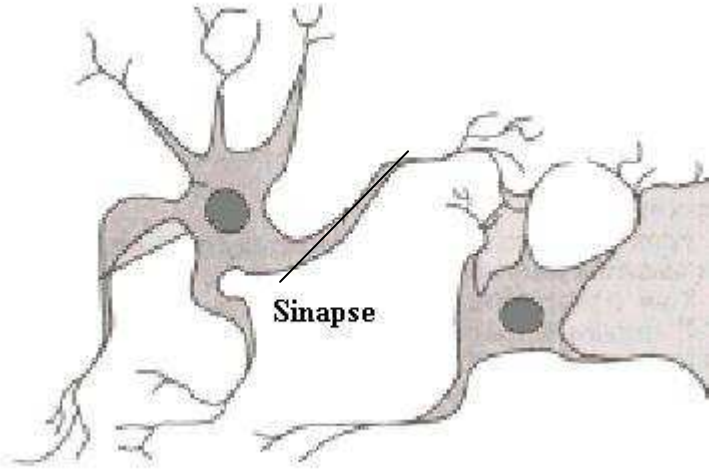

Figura 5.2 – Sinapse entre neurônios

O mecanismo de aprendizado está relacionado às sinapses que acontecem nos neurônios. A direção de um estímulo elétrico pela membrana celular de um neurônio é unidirecional. O elemento que fica antes da sinapse celular é denominado pré-sináptico e o que fica depois de pós-sináptico. O espaço entre o elemento pré-sináptico e o pós-sináptico é chamado de fenda sináptica e é onde são liberados os mediadores químicos inibidores ou excitadores da membrana.

Os sinais que chegam através dos axônios são pulsos elétricos com potenciais de ação, e constituem a informação que o neurônio processará de alguma forma para produzir, como saída, um impulso nervoso no seu axônio. A formação de um potencial de ação no axônio ocorre quando a membrana axonial sofre um

estímulo suficientemente forte para cruzar um determinado valor conhecido como limiar de disparo, ou limiar de ação.

O objetivo da sinapse é estabelecer ao neurônio receptivo uma condição de excitação ou inibição. Uma conexão sináptica poderá ser excitatória quando provocar alteração no potencial da membrana, o que contribui para a formação de um impulso nervoso no axônio de saída. Ou pode ser uma conexão inibitória, agindo no sentido oposto. Um neurônio é capaz de produzir até 10.000 sinapses com neurônios adjacentes. (CARLSON, 2002)

5.3. O neurônio artificial

A rede neural artificial é composta por elementos que simulam a rede neural biológica. É composta por conexões e pesos de entrada, correspondendo aos dendritos e às sinapses, e uma saída, correspondendo ao axônio.

Podem-se representar as entradas das redes neurais por um vetor $X, (x_1, x_2,...,x_n)$, onde n é o numero total de entradas. As entradas representam os estímulos que chegam ao neurônio em determinados instantes de tempo. Os pesos sinápticos, que representam as sinapses, ajustam esses estímulos a serem processados. Estes pesos são geralmente representados por um vetor W (do inglês weight). Este foi o modelo de neurônio proposto por McCulloch e Pitts (LUDERMIR, 2000), conforme mostrada na figura 5.3.

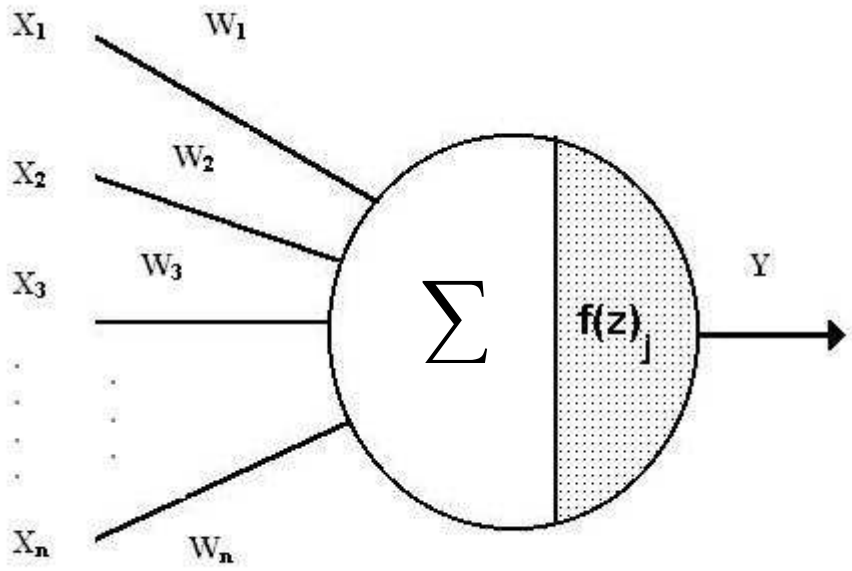

Figura 5.3 – Modelo McCulloch-Pitts

O valor do peso é alterado em função da importância do sinal de entrada, e com isso, o peso muda o seu valor representativo para a rede, criando assim um processo de aprendizagem. Um peso sináptico pode receber vários estímulos de entrada, caracterizados aqui pelo vetor X. Forma-se o produto $W_n X_n$, e logo é feito o somatório dos n-itens deste produto. Este somatório serve então como argumento para a função de ativação v. (HAYKIN, 1999)

$$v_k = \sum_{j=1}^{n} w_{kj} x_j + b_k \qquad (5.1)$$

5.4. Funções de ativação

A função de ativação pode assumir diversas formas matemáticas. A função pode ser vista como uma exigência de um limiar mínimo de ativação

(chamado de *threshold*) para que o neurônio resulte uma saída. O neurônio atua quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de ativação

Um modelo do neurônio artificial, que facilita o estudo de rede neural artificial, é apresentado na figura 5.4.

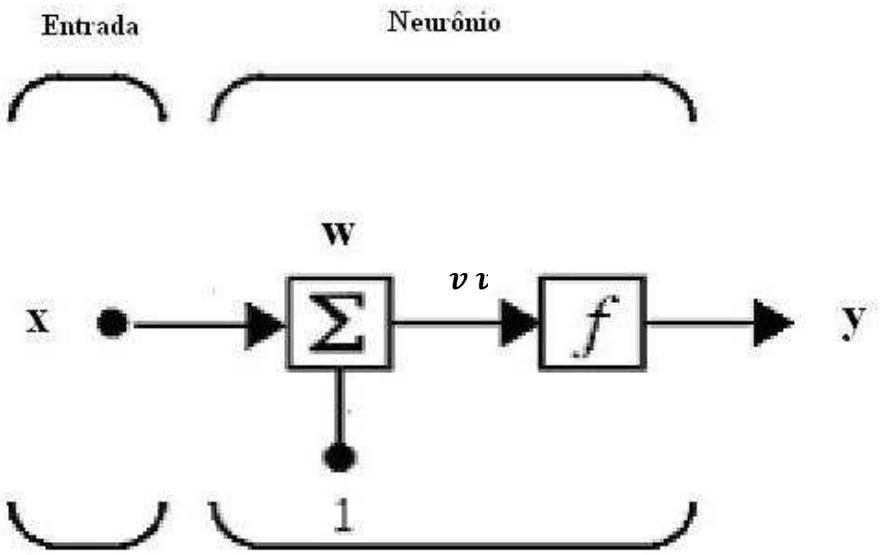

Figura 5.4 – Modelo do Matlab do neurônio artificial(MATHWORKS,2008)

As entradas da rede serão agora apresentadas pelo vetor X, neste caso exemplificado, a entrada do neurônio possui somente um elemento. W representa o peso, e a nova entrada bias – b – é apenas um adicional (pode-se entender o bias como um peso cujo valor de entrada é 1, como foi tomado na figura 5.4.), que reforça a saída da soma. A saída Y é determinada pela função de ativação v, que pode ser tanto linear ou não linear, e é escolhida dependendo das especificações do problema que o neurônio tem que resolver.

Ainda que os neurônios artificiais sejam moldados a partir de modelos biológicos, não existe nenhuma limitação para realizar modificações nas funções de

saída. Assim, serão vistos modelos artificiais que nada têm haver com as características do sistema biológico. Existem três tipos básicos de função de ativação (HAYKIN, 1999)

a) Função degrau ou função de limiar: Esta função de transferência chega à saída da rede como zero, se o argumento da função é menor que zero e se torna um se o argumento é maior que um. Seu gráfico é mostrado na figura 5.5.

$$y_k = \begin{cases} 1 \ \ se \ v_k \geq 0 \\ 0 \ \ se \ v_k < 0 \end{cases} \quad (5.2)$$

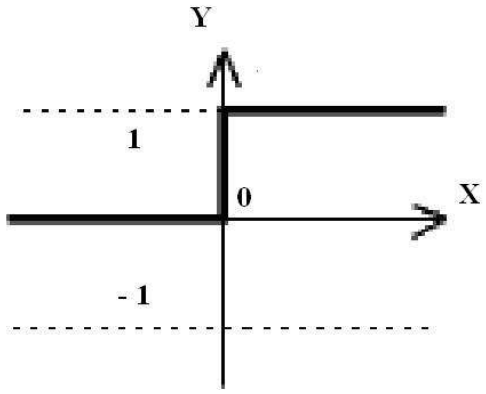

Figura 5.5 – Função degrau (MATHWORKS,2008)

b) Função rampa ou função de transferência linear: A saída de uma função de transferência linear é igual a sua entrada. Sua figura é mostrada na figura 5.6.

$$y = v \quad (5.3)$$

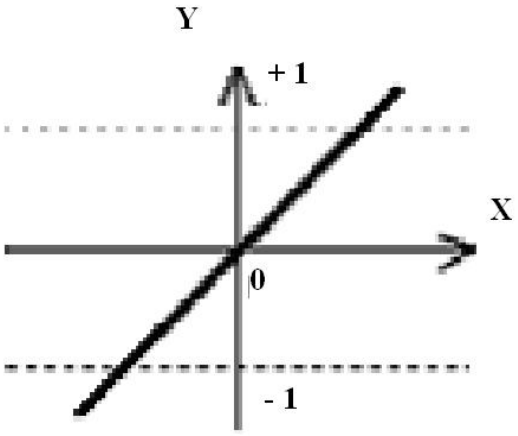

Figura 5.6 – Função rampa (MATHWORKS,2008)

c) Função de transferência sigmoidal: Esta função pega os valores de entrada, os quais podem oscilar entre mais e menos infinitos, e restringe a saída a valores entre zero e um, conforme a expressão:

$$y = \frac{1}{1+ e^{(-av)}} \qquad (5.4)$$

Onde a é a variação da inclinação. A figura 5.7 mostra seu gráfico

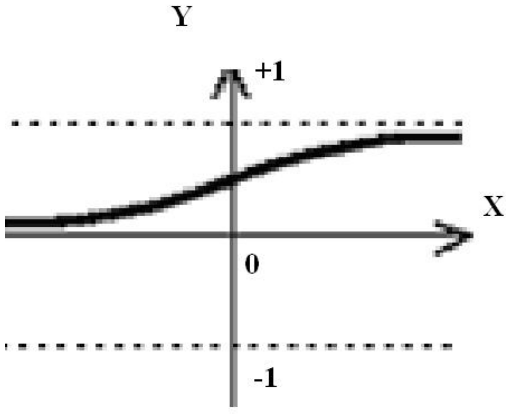

Figura 5.7 – Função sigmóide (MATHWORKS,2008)

5.5. Topologia de uma rede

Geralmente, um neurônio possui mais de uma entrada; na figura abaixo se observa um neurônio com ***n*** entradas. As entradas individuais $X_1, X_2, \ldots, X_n$ são multiplicadas pelos pesos correspondentes $W_1, W_2, \ldots, W_i$ pertencentes à matriz de pesos W.

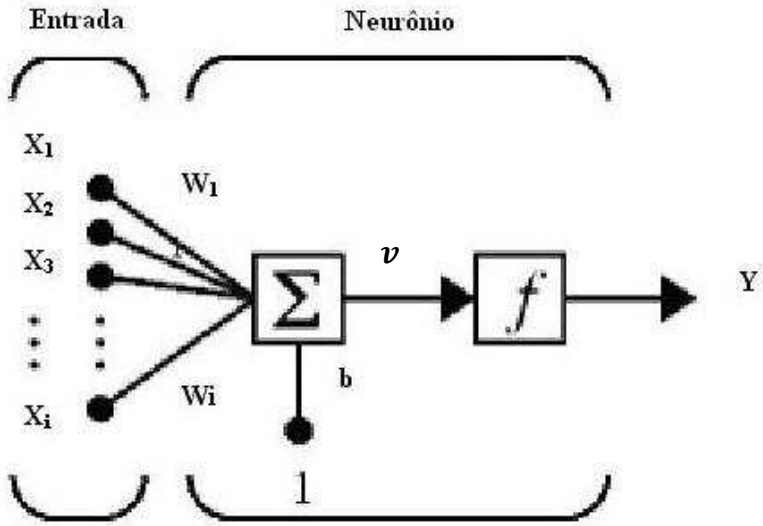

Figura 5.8 – Várias entradas no neurônio (MATHWORKS,2008)

Conforme mostrado na figura 5.8., o neurônio tem um ganho b, que chega ao mesmo somador das entradas multiplicadas pelos pesos, para a função de ativação.

$$v = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_ix_i + b \quad (5.5)$$

Dentro de una rede neural, os elementos de processamento se encontram agrupados por camadas. Uma camada é a coleção de neurônios. De acordo com a localização da camada na rede neural, acaba recebendo diferentes nomes:

a) Camada de entrada: Recebe os sinais de entrada da rede, (alguns

autores não consideram o vetor de entrada como uma camada, pois não carrega nenhum processo);

b) Camadas ocultas: Estas camadas são aquelas que não têm contato com o meio exterior, seus elementos podem ter diferentes conexões e são elas que determinam as diferentes topologias de rede;

c) Camada de saída: Recebe a informação da camada oculta e transmite a resposta para o exterior.

Uma rede com vários neurônios é mostrada na figura 5.9. Ela tem uma só camada com um número ***m*** de neurônios, Cada uma das entradas ***i*** é conectada a cada um dos neurônios, a matriz de pesos tem agora ***i***-linha e ***j***-colunas.

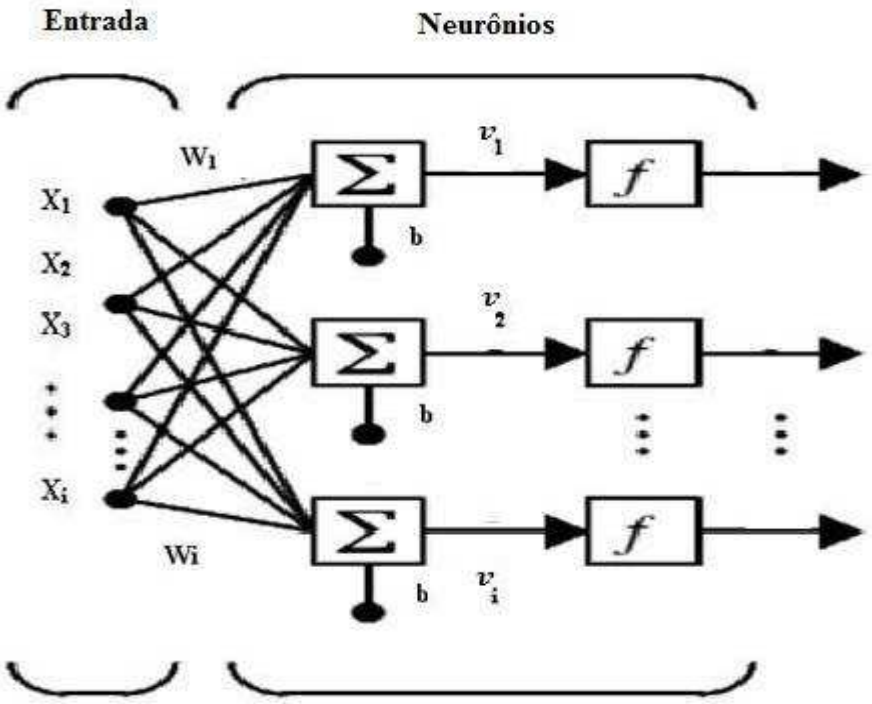

Figura 5.9 – Vários neurônios (MATHWORKS,2008)

5.6. Tipos de aprendizado

Aprendizado Supervisionado

Tem como objetivo ajustar os pesos da rede, de acordo com a resposta a ser encontrada, para encontrar uma ligação entre os pares de entrada e saída fornecidos. Neste caso, o "professor" indica explicitamente um comportamento bom ou ruim. Nos aprendizados supervisionados, é utilizada a diferença entre a resposta apresentada pela rede e a resposta desejada.

Aprendizado Não-Supervisionado

É o tipo de aprendizado que para fazer modificações nos valores das conexões sinápticas, não se usa as informações sobre a resposta da rede, isto é, não há um “professor” para acompanhar o processo de aprendizagem.

5.7. Rede de Kohonen

Os mapas Auto-Organizáveis de Kohonen (Self-Organizing Map – SOM) são umas das redes neurais artificiais mais populares na categoria de aprendizado não-supervisionado. Os mapas são redes competitivas que possuem a habilidade de formar mapeamentos que preservam a topologia entre os espaços de entrada e de saída. Os mapas resolvem problemas não-lineares de alta dimensionalidade, como por exemplo, a extração de características e classificação de padrões acústicos. O Mapa de Kohonen é um dos modelos representativos mais realísticos da função

biológica do cérebro. (KOVACS, 1996)

No cérebro existem neurônios que se organizam em zonas distintas, de modo que as informações captadas através dos órgãos sensoriais se apresentam em forma de mapas bidimensionais, conforme a figura 5.10. Ainda que a organização dos neurônios seja pré-determinada, é razoável que parte desta organização se origine do processo de aprendizagem, partindo do sentido de que o cérebro possui a capacidade inerente de formar mapas topográficos das informações extra-sensoriais.

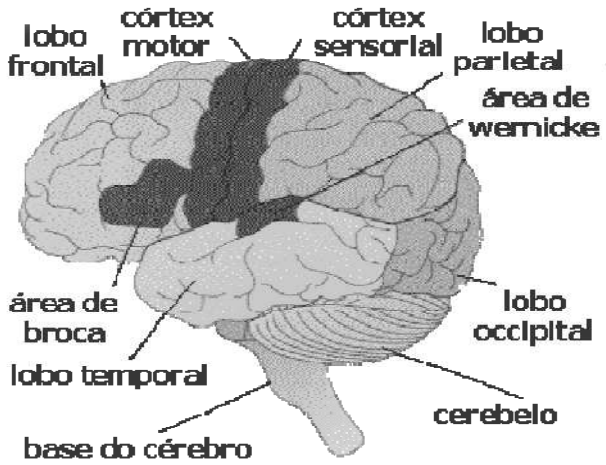

Figura 5.10 – Exemplo de mapeamento do cérebro

O desenvolvimento da rede de Kohonen adveio desta característica do cérebro humano. Com isso Tuevo Kohonen apresentou uma rede neural que possui a habilidade de construir mapas característicos de forma similar ao do cérebro, tendo como principal objetivo transformar um sinal padrão de entrada de qualquer dimensão em um mapa discreto, e desempenhar esta transformação de forma adaptativa e ordenada. (LUDERMIR, 2000)

A aprendizagem de Kohonen é off-line, que apresenta uma fase de

aprendizagem e outra (fase) de treinamento. Na fase de aprendizagem, os valores das ligações se fixam entre a camada de entrada e saída. A rede é do tipo de aprendizado não-supervisionado, de forma competitiva, sendo que os neurônios da camada de saída competem entre si e apenas um neurônio saí como vencedor. Posteriormente, os pesos sinápticos são ajustados de acordo com o neurônio vencedor.

Um conjunto de informações é apresentado à rede que servirá durante a fase de treinamento para realizar classificações de novos dados. Os valores finais dos pesos das conexões entre cada neurônio da camada de saída com a camada de entrada irão corresponder aos valores do vetor da aprendizagem que ativirá o neurônio. Em caso de existir mais padrões de treinamento do que neurônios de saída, mais um neurônio irá se associar com o neurônio vencedor, que pertence à mesma classe.

Os Mapas Auto-Organizáveis de Kohonen utilizam conceito de vizinhança, em que os neurônios próximos ao neurônio ativo terão seus pesos atualizados junto com o peso do neurônio vencedor, surgindo daí um exemplo de aprendizagem cooperativa.

O algoritmo funciona da seguinte maneira:

1. Inicializam-se os pesos aleatoriamente.

2. Aplica-se um vetor de entrada e determina-se o neurônio vencedor. Para um vetor de entrada $\boldsymbol{X}$, o neurônio vencedor é $||X-Wc|| = min\{||X-Wi||\}$, onde o índice $\boldsymbol{c}$ se refere ao neurônio vencedor e $\boldsymbol{i}$ é o neurônio comparado. O neurônio vencedor é aquele “mais próximo” do vetor de entrada.

3. Atualizam-se não só os pesos do neurônio vencedor, mas também os pesos que estão na vizinhança do neurônio vencedor. Cada vetor de peso que participa do processo de aprendizado tende um pouco à direção do vetor de entrada $\boldsymbol{X}$.

A vizinhança começa com um valor grande para que um grande número de neurônios participe do processo de aprendizado. À medida que o treinamento vai prosseguindo, o tamanho da vizinhança vai diminuindo até englobar apenas o próprio neurônio vencedor. (LUDEMIR, 2000)

5.8. Seres Humanos x Sistemas de reconhecimento utilizando redes neurais

O sistema de reconhecimento de voz/locutor que faz uso das redes neurais, assim como os humanos, conta com modelos de parâmetros com os quais o conhecimento é adquirido por afinação de parâmetro. Esses parâmetros são fixados de maneira que o modelo represente fenômenos para serem reconhecidos de forma mais precisa. Seres humanos processam as informações com redes neurais enquanto que o sistema de reconhecimento de locutor as programa utilizando diversos tipos de aprendizado.. As semelhanças entre o reconhecimento de voz/locutor e a capacidade de reconhecimento dos seres humanos estão detalhadas na tabela 5.1

Para os seres humanos o reconhecimento consiste em associar a voz do locutor com conceitos relacionados, como, por exemplo, associar as palavras pronunciadas a um objeto ou então a uma sensação, já para o sistema automático é feito através do treinamento da rede e a adaptação dos pesos sinápticos. A precisão

do reconhecimento depende da quantidade do treinamento e da capacidade de aprender o conhecimento. O sistema de reconhecimento de voz/ locutor é limitado em relação à capacidade de aprendizado do ser humano.

Tabela 5.1 – Comparação entre humanos e redes neurais (BECCHETTI, 1999)

Assuntos	Seres humanos	Sistemas de reconhecimento de voz/locutor
Modelo	Rede Neural vasta.	Redes Neurais limitadas.
Parâmetros dos modelos	Os parâmetros ajustáveis são as conexões entre as células neurais.	Os parâmetros ajustáveis são os pesos sinápticos.
Início	O bebê nasce com aproximadamente nenhum conhecimento: A rede neural está próxima do estado limpo	As redes neurais estão sem nenhum conhecimento.
Reconhecimento no começo	O bebê não reconhece as palavras, mas com algum treinamento, é bem capaz de aprender a linguagem.	Do mesmo modo.
Treinamento	O treinamento é realizado escutando os locutores e a associação da fala com o seu significado.	O treinamento é realizado escutando os locutores e a associação da fala é feita através dos ajustes dos pesos sinápticos.
Efeito do treinamento	Fixa e força as conexões das redes neurais. Quanto maior for o treinamento, maior é a precisão de reconhecimento.	Do mesmo modo.
Análise da voz	Os ouvidos processam a voz integrando em faixas de freqüência e percebendo o som com logaritmos dinâmicos	O sistema de reconhecimentos de voz faz aproximadamente o mesmo
Adaptação a novos locutores	A precisão de reconhecimento de novos locutores cresce após um tempo.	Se os algoritmos de adaptação a novos locutores estão disponíveis, a precisão aumenta depois de certo tempo.
Modelo de linguagem	Uma grande ajuda é obtida no reconhecimento considerando os aspectos semânticos, sintáticos e pragmáticos.	Os aspectos semânticos e pragmáticos ainda não são totalmente explorados. O uso de um corretor ortográfico poderia reduzir a quantidade de erros.
Processo de reconhecimento	Pode-se assumir que os seres humanos efetuam inconscientemente o mesmo procedimento do reconhecimento automático de voz.	Encontra o sinal que melhor se encaixa com a rede neural treinada.

6. Implementação do modelo

6.1. O software escolhido

O sistema construído foi feito em MatLab (MATrix LABoratory). O Matlab é um software de ambiente matemático e também é uma linguagem de programação. A figura 6.1 mostra o ambiente do Matlab.

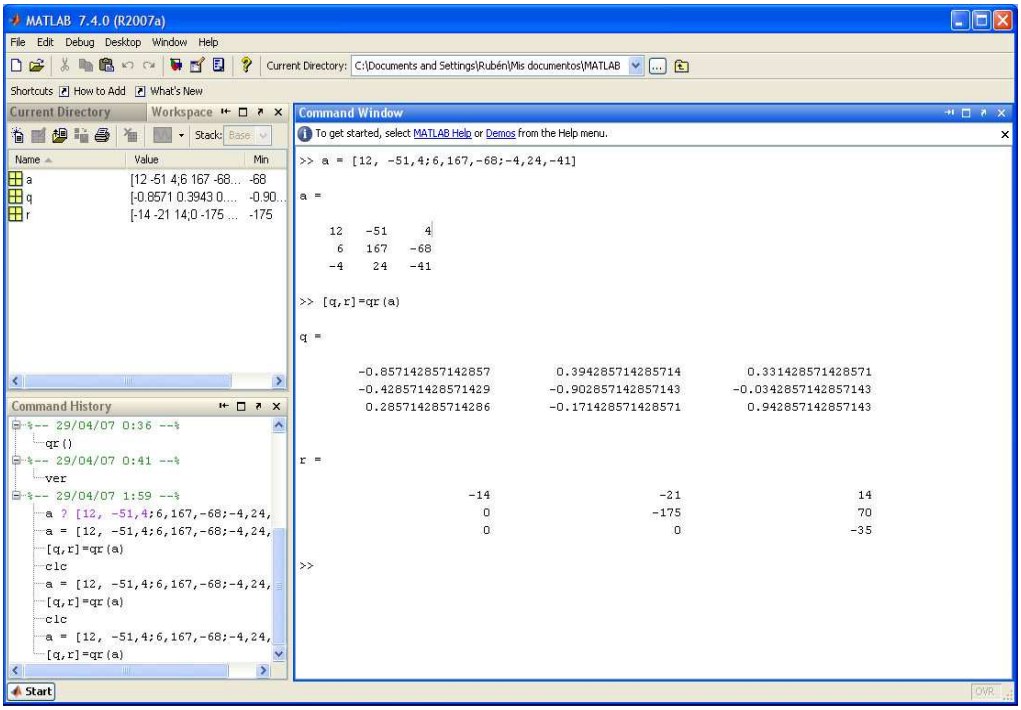

Figura 6.1 – Tela do Matlab

Ele é otimizado para executar cálculos científicos e de engenharia. As características que influenciaram na escolha do software são mostradas a seguir.

O Matlab apresenta muitas vantagens, para resolução de problemas técnicos. Pode-se destacar a facilidade de uso, é uma linguagem intuitiva, ideal para uso educacional. Além disso, é independente de plataforma, e deste modo, uma função que seja feita como, por exemplo, no *Linux,* pode ser interpretada e no

ambiente *Windows.* (CHAPMAN, 2002)

Outra vantagem que o Matlab apresenta está relacionado à *Toolbox*, que são bibliotecas de rotina aplicadas à áreas específicas, como por exemplo, redes neurais, processamento de sinais, estatística, controle, etc. (MATSUMOTO, 2004)

Por ser uma linguagem interpretada, a desvantagem que o Matlab apresenta é se tornar lenta em comparação a linguagens compiladas, tais como *Java* e C.

Foram usadas duas versões do Matlab, a versão 5.3.0.10183 (R11) e a versão 7. O motivo de ter-se usado ambas versões é que uma delas apresentou incompatibilidade para instalação no notebook, enquanto que no desktop utilizou-se outra.

6.2. Especificações do projeto

A freqüência em que foi realizada foi de 11025 Hertz e o processo de transformação de analógico para digital foi feito pela própria placa de som.

As aquisições dos sinais de voz foram feitas através do programa Audacity. Optou-se por este programa por ser de fácil manuseio e pela opção de trabalhar com o sinal capturado. Contudo, foi feita no Matlab, uma função que faz a aquisição da fala, caso seja necessário adicionar um novo usuário para treinamento. Foram salvos em formato .WAV, para serem trabalhadas no Matlab. A tela do Audacity é mostrada na figura 6.2.

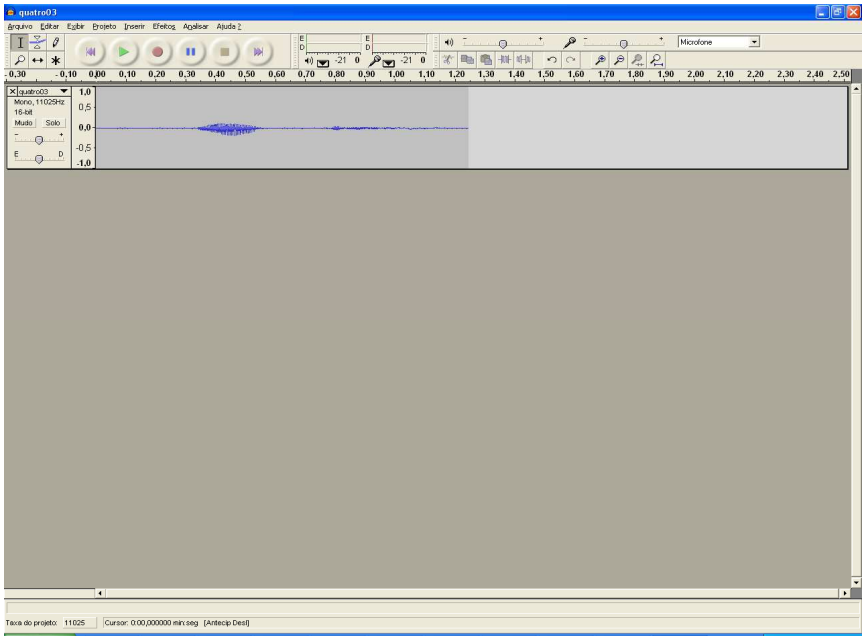

Figura 6.2 – Janela do Audacity

Foram gravadas as seguintes palavras: ***sobe, desce, um, dois, três, quatro, cinco, direita, esquerda, aplicativo, fecha, aumenta, diminui, iniciar, desligar, internet, texto, matlab, documentos, correio*.** Cada uma foi repetida três vezes por cada pessoa. Foram escolhidas seis pessoas para fazer a gravação, sendo quatro do sexo masculino e dois do sexo feminino, totalizando no final 360 palavras gravadas. Pediu-se aos locutores que falassem as palavras em seqüência para depois fazer a divisão das palavras utilizando o Audacity

Metodologia

Primeiro é feita aquisição do sinal de voz, Como exemplo ilustrativo, escolheu-se a palavra **aplicativo**. A aquisição do sinal de voz ocorre através do comando:

```
[Yentrada,fs,NBITS]=WAVREAD('C:\MATLABR11\work\Sons\Voz...
...Bruno\aplicativo01.wav');
```

Onde *Yentrada* se trata do vetor de dados que representa o sinal, *fs* é a freqüência em que se foi gravado e *NBITS* é o número de bits.

A figura 6.3 mostra a forma de onda do sinal, *amplitude x tempo*.

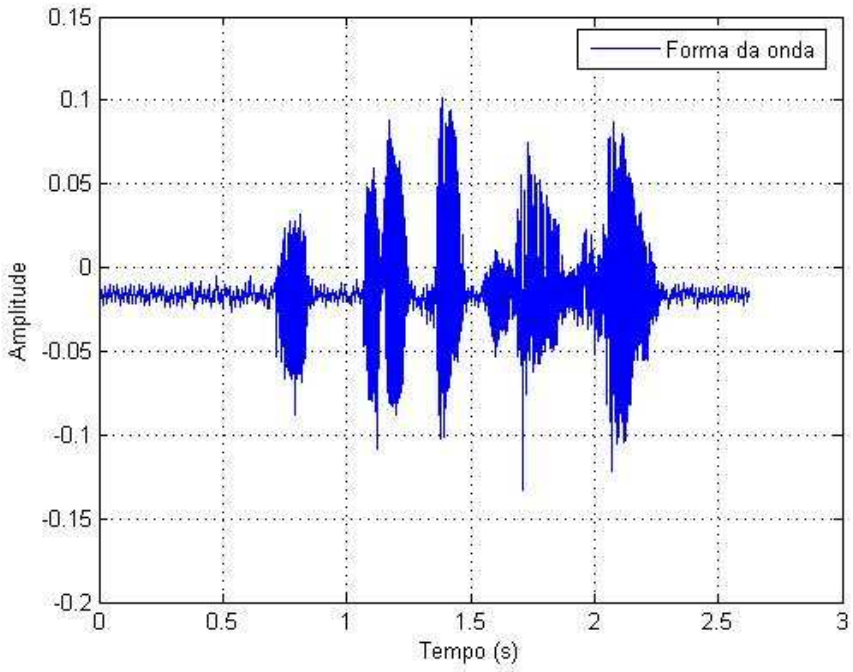

Figura 6.3 – Sinal original

Logo após, utilizando a função *endpoint2*, são detectados os extremos do sinal de fala, por meio dos seguintes comandos:

```
intervalo = endpoint2(Yentrada, fs, NBITS);
Yentrada = Yentrada(intervalo(1):intervalo(2));
```

O segundo comando aplica os intervalos encontrados ao sinal de voz. A figura 6.4 mostra o sinal de voz com os extremos detectados. A importância para se trabalhar somente com a palavra, retirando-se os silêncios anteriores e posteriores, é porque as variações de suas características retiradas diminuem, influenciando no resultado da comparação.

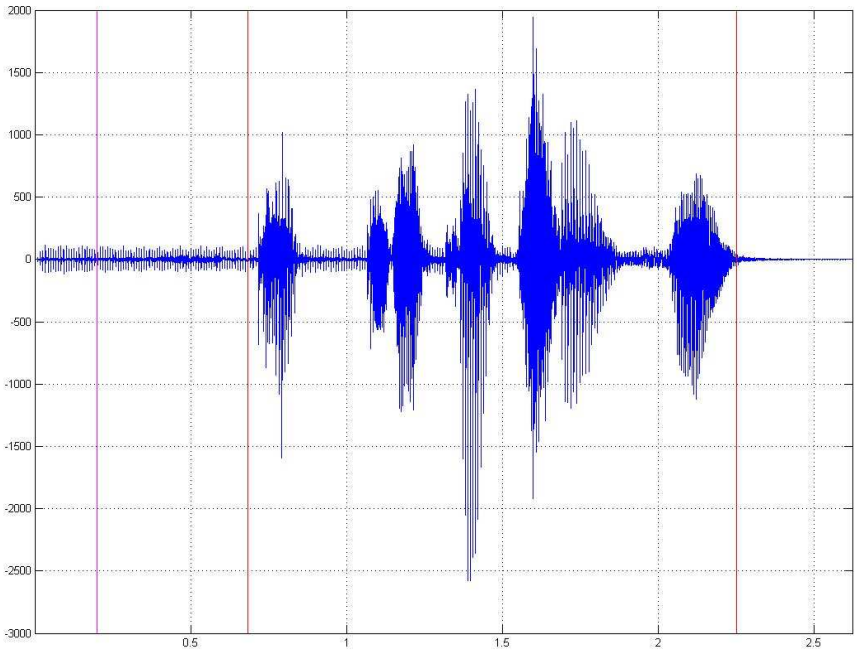

Figura 6.4 – Detecção do extremos da palavra

Logo após este processo, é feito um tratamento para se retirar o ruído como mostrado na figura 6.5, por meio do comando:

[Yentrada,po]=specsubm(Yentrada,fs);

A questão da retirada do ruído está relacionada ao tratamento do sinal.

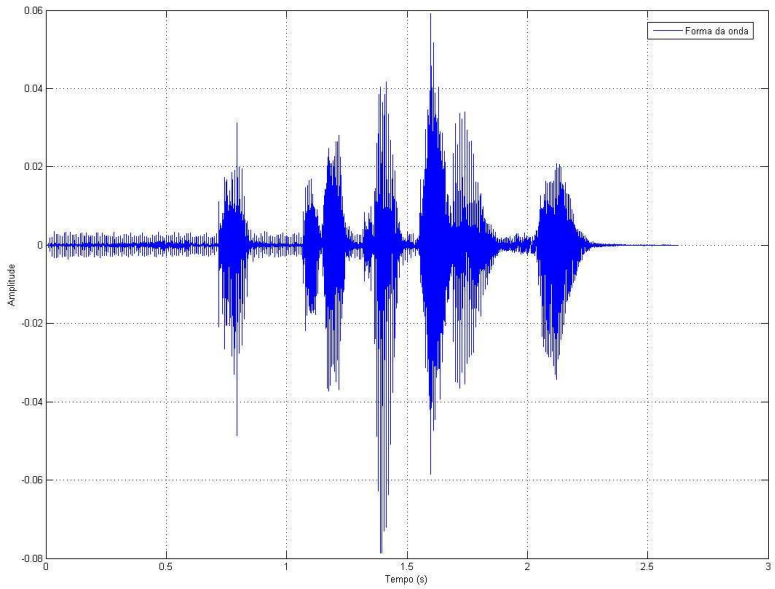

Figura 6.5 – Sinal sem ruído

Após esta estas etapas, faz-se a divisão do sinal em 256 frames. A figura 6.6 ilustra o sinal dividido em vários frames. Isso ocorre através da função buffer2 e aplicado da seguinte maneira

framedy = buffer2(Yentrada,256);

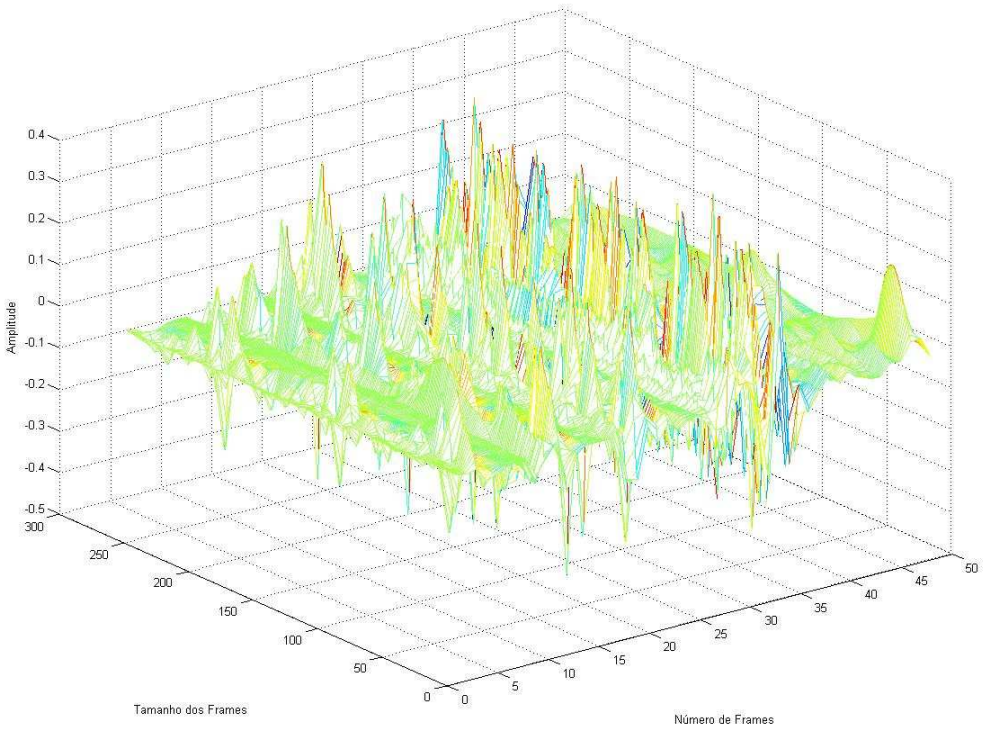

Figura 6.6 –Sinal dividido em vários frames

Depois, de cada frame, são retirados a energia, a taxa de cruzamento zero e o coeficiente LPC.

```
yin = Yentrada;
x = Yentrada;
zcnt=zeros(1,frameNum);
k=1;
for t=1: frameNum
  yin=x(k:k+frameSize-1);
 eng(t)=10*log10(norm(yin,2));
  zc=0;
 x2=zeros(frameSize,1);
```

```
x2(1: frameSize -1)=yin(2:frameSize);
zc=length(find((yin>0 & x2<0) | (yin<0 & x2>0))); %-- Zero Crossings
zcnt(t)=zc;
coefLPC(t) = mean(lpc(real(x2), 15));
k=k+frameSize;
end
```

Este tratamento é feito com todos os sinais de voz antes de serem inseridos na rede neural. O mínimo e o máximo de cada parâmetro são utilizados como pesos para a rede de Kohonen. Com os comandos

```
[mine,maxe,minz,maxz, maxlpc, minlpc] = peso(eng,zcnt,coefLPC);
```

A rede é inicializada com todas as características e ajustada para 20.000 épocas. Aqui, o número de épocas se trata de número de vezes que a rede neural é treinada.

```
net1 = newc([minegeral maxegeral; minzgeral maxzgeral; minlpcgeral ...
...maxlpcgeral], 3);
net1.trainParam.epochs = 20000;
```

São criados vários padrões para cada locutor da seguinte forma

```
Padrao = [eng' zcnt' coefLPC'];
Padrao1 = [eng1' zcnt1' coefLPC1'];
.
.
.
Padrao17 = [eng17' zcnt17' coefLPC17'];
```

Depois é feito o treinamento de diversos locutores, em uma única rede, caracterizando-a como rede padrão da palavra aplicativo.

net1 = train(net1,Padrao');

net1 = train(net1,Padrao1');

.

.

.

net1 = train(net1,Padrao17');

Logo após, retira-se o vetores de saída e é realizada sua organização.

Ysinalnet1 = sim(net1, Padrao');

Ysinal1net1 = sim(net1, Padrao1');

.

.

Ysinal17net1 = sim(net1, Padrao17');

Yo = vec2ind(Ysinalnet1);

Yo1 = vec2ind(Ysinal1net1);

.

.

.

Yo17 = vec2ind(Ysinal17net1);

Após estes processos, a rede e seus vetores de saída são salvas para posterior comparação. O formato que o Matlab salva dados é .MAT, que é entendido pelo próprio programa.

save aplicativo net1 Yo Yo1 Yo2 Yo3 Yo4 Yo5 Yo6 Yo7 Yo8 Yo9 Yo10 Yo11…

…Yo12 Yo13 Yo14 Yo15 Yo16 Yo17;

Com a rede já treinada, podem-se apresentar quaisquer palavras – podendo ser a mesma ou outras do banco de dados –, que sofrem o mesmo processo de extração das características. Por meio da função *comparando*. Nesta função são especificadas as palavras que serão comparadas, mudando-as no próprio código-fone. Faz-se a comparação destas palavras com cada um dos vetores já treinados.

Com saída, é apresentado a média de verdadeiro e falso e respectivos desvios padrões. Alguns resultados são exibidos no próximo capítulo.

7. Resultados obtidos

Como exemplo, pode-se mostrar o resultado do treinamento da palavra ***quatro***. Após a rede ser simulada, quando é apresentada a mesma palavra é reconhecida com um grau aceitável, independente de locutor. Conforme a tabela 7.1 a seguir.

Tabela 7.1 – Comparação com palavra quatro

Palavra Treinada – Quatro Tendo locutor 1 como padrão	Porcentagem de Verdadeiro (%)	Porcentagem de Falso (%)
Locutor 1 – Gravação um	**52,126**	47,873
Locutor 1 – Gravação dois	**63,173**	36,827
Locutor 1 – Gravação três	**55,729**	44,271
Locutor 2 – Gravação um	**54,231**	45,768
Locutor 2 – Gravação dois	**63,672**	36,328
Locutor 2 – Gravação três	42,405	**57,595**
Locutor 3 – Gravação um	**55,339**	44,661
Locutor 3 – Gravação dois	**50,846**	49,154
Locutor 3 – Gravação três	**59,223**	40,777
Locutor 4 – Gravação um	**59,549**	40,451
Locutor 4 – Gravação dois	**57,140**	42,860
Locutor 4 – Gravação três	**63,824**	36,176
Locutor 5 – Gravação um	36,762	**63,238**
Locutor 5 – Gravação dois	48,893	**51,107**
Locutor 5 – Gravação três	41,341	**58,659**
Locutor 6 – Gravação um	**61,502**	38,498
Locutor 6 – Gravação dois	**53,581**	46,419
Locutor 6 – Gravação três	**64,562**	35,438

Já quando é apresentada uma palavra diferente a sua porcentagem de verdadeiro decai, conforme verificado na tabela 7.2.

Tabela 7.2 - Comparação com a palavra quatro e outras palavras

Palavra Treinada - Quatro	Porcentagem de Verdadeiro (%)	Porcentagem de Falso (%)
Locutor 1 - Esquerda	43,663	**56,337**
Locutor 1 - Cinco	44,271	**55,729**
Locutor 1 - Correio	29,622	**70,378**
Locutor 2 - Esquerda	37,630	**62,370**
Locutor 2 - Cinco	46,202	**53,798**
Locutor 2 - Correio	40,777	**59,223**
Locutor 3 - Esquerda	**58,746**	41,254
Locutor 3 - Cinco	**62,218**	37,782
Locutor 3 - Correio	**53,472**	46,528
Locutor 4 - Esquerda	**52,582**	47,418
Locutor 4 - Sobe	**59,744**	40,256
Locutor 4 - Correio	42,057	**57,943**
Locutor 5 - Esquerda	46,137	**53,863**
Locutor 5 - Cinco	**63,216**	36,784
Locutor 5 - Correio	48,524	**51,476**
Locutor 5 - Esquerda	**62,348**	37,652
Locutor 4 - Cinco	**58,203**	41,797
Locutor 4 - Correio	33,898	**66,102**

Outro teste foi feito com a palavra **um,** tabela 7.3 e tabela 7.4.

Tabela 7.3 - Comparação com a palavra um

Palavra Treinada – Um Tendo locutor 1 como padrão	Porcentagem de Verdadeiro (%)	Porcentagem de Falso (%)
Locutor 1 – Gravação um	**57,856**	42,144
Locutor 1 – Gravação dois	**59,549**	40,451
Locutor 1 – Gravação três	43,316	**56,684**
Locutor 2 – Gravação um	**54,926**	45,074
Locutor 2 – Gravação dois	**56,272**	43,728
Locutor 2 – Gravação três	49,957	**50,043**
Locutor 3 – Gravação um	46,701	**53,299**
Locutor 3 – Gravação dois	44,748	**55,252**
Locutor 3 – Gravação três	44,748	**55,252**
Locutor 4 – Gravação um	**54,297**	45,703
Locutor 4 – Gravação dois	**58,268**	41,732
Locutor 4 – Gravação três	**59,592**	40,408
Locutor 5 – Gravação um	**63,542**	36,458
Locutor 5 – Gravação dois	**57,574**	42,426
Locutor 5 – Gravação três	**63,455**	36,545
Locutor 6 – Gravação um	**60,764**	39,236
Locutor 6 – Gravação dois	**62,174**	37,826
Locutor 6 – Gravação três	**62,999**	37,001

Tabela 7.4 - Comparação com a palavra um e outras palavras

Palavra Treinada - Um	Porcentagem de Verdadeiro (%)	Porcentagem de Falso (%)
Locutor 1 - Esquerda	49,566	**50,434**
Locutor 1 - Cinco	**55,512**	44,488
Locutor 1 - Correio	27,083	**72,917**
Locutor 2 - Esquerda	29,666	**70,334**
Locutor 2 - Cinco	37,977	**62,023**
Locutor 2 - Correio	35,938	**64,063**
Locutor 3 - Esquerda	41,949	**58,051**
Locutor 3 - Cinco	30,816	**69,184**
Locutor 3 - Correio	26,324	**73,676**
Locutor 4 - Esquerda	27,170	**72,830**
Locutor 4 - Sobe	44,119	**55,881**
Locutor 4 - Correio	20,595	**79,405**
Locutor 5 - Esquerda	35,829	**64,171**
Locutor 5 - Cinco	**52,148**	47,852
Locutor 5 - Correio	19,271	**80,729**
Locutor 5 - Esquerda	50,760	**49,240**
Locutor 4 - Cinco	46,723	**53,277**
Locutor 4 - Correio	31,814	**68,186**

Outro teste, cujo resultado de porcentagem foi totalmente verdadeiro, ocorreu com a palavra **esquerda**, tabela 7.5.

Tabela 7.5 - Comparação com a palavra esquerda

Palavra Treinada – Esquerda Tendo locutor 1 como padrão	Porcentagem de Verdadeiro (%)	Porcentagem de Falso (%)
Locutor 1 – Gravação um	**62,196**	37,804
Locutor 1 – Gravação dois	**66,081**	33,919
Locutor 1 – Gravação três	**69,010**	30,990
Locutor 2 – Gravação um	**76,150**	23,850
Locutor 2 – Gravação dois	**78,472**	21,528
Locutor 2 – Gravação três	**78,190**	21,810
Locutor 3 – Gravação um	**55,469**	44,531
Locutor 3 – Gravação dois	**54,449**	45,551
Locutor 3 – Gravação três	**50,933**	49,067
Locutor 4 – Gravação um	**78,668**	21,332
Locutor 4 – Gravação dois	**71,875**	28,125
Locutor 4 – Gravação três	**77,886**	22,114
Locutor 5 – Gravação um	**70,378**	29,622
Locutor 5 – Gravação dois	**73,134**	26,866
Locutor 5 – Gravação três	**71,528**	28,472
Locutor 6 – Gravação um	**78,168**	21,832
Locutor 6 – Gravação dois	**78,906**	21,094
Locutor 6 – Gravação três	**78,255**	21,745

Discussão dos resultados

Conforme apreciação dos resultados, com relação ao reconhecimento de

vocábulos, quando se trata do reconhecimento da mesma palavra, houve alto grau de aceitação. Já para se rejeitar uma palavra diferente, aconteceu uma variação entre aceitação e rejeição.

Vale ressaltar que não está se fazendo distinções dos locutores masculinos e femininos, e como também se faz o uso de apenas três parâmetros: a taxa de cruzamento-zero, a energia e coeficientes LPC.

Inicialmente, tinha-se como objetivo fazer o reconhecimento de palavras isoladas. Entretanto não houve um grau elevado para o reconhecimento das mesmas. Outro fato a se comentar, é o grau de reconhecimento de locutor que está bastante alto, conforme as tabelas mostradas, motivo pelo qual foi mudado o enfoque do trabalho em questão para reconhecimento de vocábulos.

8. Conclusão

A proposta que aqui se apresenta não foi de fazer um modelo definitivo de reconhecimento de fala, mas sim de focar os processos básicos que as cerca.

Devido às várias áreas que um programa como este abrange, tendeu-se mais ou para áreas que foram vistas durante o curso – análise de sinais – ou para áreas de interesses do autor – redes neurais artificiais.

Dificuldades encontradas

Conforme mostrado nos resultados, o enfoque do projeto foi mudado de reconhecimento de voz para reconhecimento de vocábulo, e com isto não houve prejuízo em relação ao conhecimento adquirido.

Além disso, em alguns momentos fez-se necessário fazer a regravação de outras palavras, por vários motivos: má qualidade do áudio, erro ao detectar os extremos da palavra, padronização da amplitude das palavras.

Propostas futuras

Como foi feito apenas um protótipo, e devido à falta de tempo, ainda resta questão de melhoramento do programa para o usuário final, com utilização de janelas e botões. O Matlab facilita bastante para estas criações.

No processo de extração de parâmetros, é possível utilizar-se de coeficientes mel-cepstrais, que são obtidos pela representação da freqüência em outro tipo de escala. E logo após fazer a aplicação de um banco de filtros. (BRAGA,

2006)

Com relação ao processo de comparação, além das redes de Kohonen, é possível fazê-lo por meio das redes MLP – Multilayer Perceptron, ou redes Radial Basis. Outros tipos de redes neurais também são adaptáveis.

Além dos métodos de Redes Neurais Artificiais, podem-se usar os Modelos de Mistura Gaussiana e Modelos Ocultos de Markov. (BRAGA, 2006)

Referência Bibliográfica

AUDACITY. Disponível em: <http://audacity.sourceforge.net/>>. Acesso em: 04 jul 2008.

BARBISAN, Márcio. *Processador de Voz*. Disponível em: <http://www.marciobarbisan.hpg.ig.com.br/pcq.htm>. Acesso em: 04 jul 2008.

BECCHETTI, Claudio; RICOTTI, Lucio Prina. *Speech Recognition – Theory and C++ Implementation*. Inglaterra: Wiley, 1999.

BECHARA, Evanildo. *Moderna Gramática Portuguesa.* Rio de Janeiro: Editora Lucerna, 2003.

BRAGA, Petrônio. *Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais – Trabalho de conclusão de curso.* Pernambuco: Escola Politécnica de Pernambuco, 2006.

CALLOU, Dinah; LEITE, Yonne. *Iniciação à Fonética e à Fonologia.* Rio de Janeiro: Coleção Letras, Jorge Zahar Editor, 1995.

CARLSON, Neil R. *Fisiologia do Comportamento*. São Paulo: Mamole,2002.

CHAPMAN, J. Stephen. *Programação em MATLAB para engenheiros*. São Paulo: Thomsom, 2002.

FARIA, Isabel Hub; PEDRO, Emilia Ribeiro; DUARTE, Inês. *Introdução à Lingüística – Geral e Portuguesa.* Lisboa: Caminho, 1996.

GILAT, Amos. *MATLAB com aplicações em engenharia*. Porto Alegre: Bookman, 2006.

HAYKIN, Simon. *Redes Neurais – princípios e prática*. Porto Alegre: Bookman, 1999.

HAYKIN, Simon; VEEN, Barry Van. *Sinais e Sistemas*. Porto Alegre: Bookman, 2001.

HSU, Hwei P. *Sinais e Sistemas – Coleção Schaum*. Porto Alegre: Bookman, 2004.

INGLE, Vinay K.; PROAKIS, John G. *Digital Signal Processing using MATLAB*. Canadá: Brooks/Cole Thomson Learing, 2000.

KOVACS, Zsolt Laszlo. *Redes Neurais Artificiais: Fundamentos e Aplicações.* São Paulo: Livraria da Física, 1996.

LITTLEFIELD, Hanselman. *MATLAB 6 – Curso Completo*. São Paulo: Prentice Hall, 2003.

LUDERMIR, Teresa B; CARVALHO, André Carlos P.L.F; BRAGA, Antônio P. *Redes Neurais Artificiais: Teoria e Aplicações.* Rio de Janeiro: LTC, 2000.

LUFT, Pedro Celso. *Novo Manual de Português*, 5ª ed. Rio de Janeiro:Globo, 1989.

MARTIN, Rabiner. *Spectral Subtraction Based on Minimum Statistics.* Disponível em: <http://www.ruhr-uni-bochum.de/ika/ika/forschung/gruppe_martin/hum_mach_interf/ eusipco1994_martin.pdf>>. Acesso em: 04 jul 2008.

MARTINS, Maria Raquel. *Ouvir Falar – Introdução à Fonética do Português***.** Lisboa: Caminho, 1988.

MATHWORKS. Disponível em: < http://www.mathworks.com >. Acesso em: 04 jul 2008.

MATSUMOTO, Élia Yathie. *MATLAB 7 – fundamentos*. São Paulo: Editora Érica, 2004.

RIBEIRO, Carlos Eduardo de Meneses. *Curso de processamento digital de fala*.

Disponível em: <http://www.deetc.isel.ipl.pt/comunicacoesep/disciplinas/pdf/>. Acesso em: 04 jul 2008.

ROSETTI, A. Tradução: BUESCU, Maria Leonor Carvalhão. *Introdução à Fonética.*. Lisboa: Colecção Saber, Publicações Europa-América, 1974.

SACCONI, Luis Antonio. *Nossa Gramática – Teoria.* São Paulo: Atual Editora LTDA. 1989.

VOICEBOX – *SPEECH PROCESSING TOOLBOX FOR MATLAB*. Disponível em: <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>. Acesso em: 04 jul 2008.

Apêndice Código Fonte em Matlab