JANDERSON RODRIGO DE OLIVEIRA

UMA PROPOSTA DE UTILIZAÇÃO DE REDES NEURAIS PARA RECONHECIMENTO DE INDIVÍDUOS ATRAVÉS DA FALA

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

LAVRAS
MINAS GERAIS – BRASIL
2007

JANDERSON RODRIGO DE OLIVEIRA

UMA PROPOSTA DE UTILIZAÇÃO DE REDES NEURAIS PARA RECONHECIMENTO DE INDIVÍDUOS ATRAVÉS DA FALA

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Área de Concentração:

Redes Neurais

Orientador:

Prof. Dr. Wilian Soares Lacerda

LAVRAS

MINAS GERAIS – BRASIL

2007

Ficha Catalográfica preparada pela Divisão de Processos Técnico

da Biblioteca Central da UFLA

Oliveira, Janderson Rodrigo de

Uma Proposta de Utilização de Redes Neurais para Reconhecimento de Indivíduos através da Fala / Janderson Rodrigo de Oliveira. Lavras – Minas Gerais, 2007.

Monografia de Graduação – Universidade Federal de Lavras. Departamento de Ciência da Computação.

1. Computação. 2. Redes Neurais. 3. Reconhecimento de Fala. I. OLIVEIRA, J. R. II. Universidade Federal de Lavras. III. Título.

JANDERSON RODRIGO DE OLIVEIRA

UMA PROPOSTA DE UTILIZAÇÃO DE REDES NEURAIS PARA RECONHECIMENTO DE INDIVÍDUOS ATRAVÉS DA FALA

Monografia de graduação apresentada ao Departamento de Ciência da Computação da Universidade Federal de Lavras como parte das exigências do Curso de Ciência da Computação para obtenção do título de Bacharel em Ciência da Computação.

Aprovada em 16 de janeiro de 2008

_____
Prof. Dr. Wilian Soares Lacerda
(Orientador)

_____
Prof. Dr. Claudio Fabiano Motta Toledo

_____
Prof. Dr. Tadayuki Yanagi Junior

LAVRAS

MINAS GERAIS – BRASIL

Agradeço a todas as vozes que ecoaram neste trabalho.

Uma Proposta de Utilização de Redes Neurais para Reconhecimento de Indivíduos através da Fala

RESUMO

O objetivo desse trabalho foi desenvolver um sistema de reconhecimento de indivíduos através da fala, utilizando uma rede neural *multilayer perceptron* com um algoritmo de aprendizagem *backpropagation*. O sistema desenvolvido foi implementado na ferramenta Scilab, sendo constituído por dois módulos: um módulo responsável pelo tratamento do sinal sonoro e outro destinado ao treinamento e teste da rede. O sistema apresentou taxas de acerto consideráveis para os testes realizados. Porém, para a sua utilização em uma aplicação do mundo real deve-se levar em consideração alguns critérios importantes, como por exemplo: a segurança e a confiabilidade, visto que em nenhuma das implementações realizadas se conseguiu atingir uma taxa de erros nula.

Palavras chaves: Redes neurais artificiais, reconhecimento de padrões, reconhecimento de indivíduos através da fala

An Approach of Use of Neural Networks to Speaker Recognition

ABSTRACT

The aim of this work was to develop a speaker recognition system, using a neural network multilayer perceptron with an algorithm for learning backpropagation. The system was implemented in the Scilab tool, and consists of two modules: a module for the treatment of the sound and another for the training and testing of the network. The system presented considerable adjustment rates for the tests. But for its use in a real world application must take into account some important criteria, such as: safety and reliability, since in any of the implementations has been made to achieve a rate of zero errors.

Keywords: Artificial neural network, pattern recognition, speaker recognition

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Uma das diferenças entre máquinas e humanos é o tempo requerido para realizar tarefas complicadas, tal como o reconhecimento de padrões. Computadores são extremamente rápidos, contudo se torna difícil, por exemplo, implementar máquinas que possam reconhecer objetos tridimensionais em tempo real. Enquanto que os humanos, cujos cérebros são compostos por neurônios que são milhares de vezes mais lentos do que componentes eletrônicos, podem reconhecer velhos amigos quase que instantaneamente.

Sabe-se que os computadores executam suas computações seqüencialmente, ou seja, passo a passo, enquanto que o cérebro humano processa a informação em paralelo. Sendo assim, a capacidade que o ser humano possui de realizar tarefas complexas e principalmente de aprender se deve ao processamento paralelo e distribuído da rede de neurônios do cérebro. Tendo-se como preceito esta característica, as redes neurais artificiais surgiram como uma alternativa computacional para a resolução das mais diversas questões relacionadas ao reconhecimento de padrões, classificação, agrupamento e aproximação de dados.

Segundo Von Zuben (2003) a ampla utilização das redes neurais e a sua eficácia se devem a dois fatores importantes, o próprio amadurecimento da área de pesquisa e a evolução da natureza dos problemas que hoje desafiam a computação.

Tem-se que o paradigma neural se baseia no reconhecimento de regularidades e padrões de dados através da experiência, tornando-se capaz de realizar generalizações apoiadas no conhecimento acumulado previamente. Sendo assim, ele pode ser considerado uma poderosa ferramenta na resolução de questões cujas regras de solução são desconhecidas ou difíceis de formalizar.

De acordo com Bishop (1995) o termo “reconhecimento de padrões” se refere a uma ampla extensão de problemas de processamento de informação que possuem grande importância prática, tal como a classificação de caracteres manuscritos, a detecção de doenças e falhas em diagnósticos médicos e mecânicos e o reconhecimento de fala que será alvo de estudo e análise no presente trabalho.

O reconhecimento de padrões possui uma longa história, mas depois de 1960 foi amplamente utilizado em pesquisas na área de estatística. Com o advento dos computadores houve uma demanda crescente por aplicações práticas do reconhecimento de padrões, que determinou novas demandas por desenvolvimentos teóricos. Como nossa

sociedade evoluiu de uma fase industrial para uma fase pós industrial, a automação na produção industrial e a necessidade de manipulação e recuperação de informação tornaram-se de considerável importância. Essa tendência foi impulsionada pela utilização de reconhecimentos de padrões nas mais diversas aplicações e pesquisas na área de engenharia. Sendo assim, o reconhecimento de padrões tornou-se parte integral nos mais diversos sistemas inteligentes que auxiliam na tomada de decisões.

1.1. Motivação

Tem-se como motivação para o presente trabalho as seguintes considerações:

- Oferecer uma alternativa para a identificação de pessoas de forma única, podendo ser utilizada na identificação criminal e no controle de acesso, por exemplo.
- Ser uma proposta de baixo custo para o reconhecimento de indivíduos através da fala.
- Descobrir e explorar as potencialidades de extração de padrões/características das redes neurais no reconhecimento de palavras faladas.
- Analisar a efetividade da voz humana como instrumento de autenticação.

1.2. Objetivos

Este trabalho tem como principal objetivo desenvolver uma rede neural que seja capaz de reconhecer um indivíduo através da sua voz. Sendo assim, pode-se definir para este trabalho os seguintes objetivos específicos:

- Abordar os diversos conceitos sobre reconhecimento de padrões, redes neurais e as características do som e da fala humana;
- Definir uma estratégia de captura e tratamento do sinal de voz;
- Apresentar e desenvolver uma rede neural que seja capaz de realizar o reconhecimento proposto;
- Analisar a efetividade da rede implementada.

1.3. Estrutura do Trabalho

O presente trabalho se encontra dividido em 5 (cinco) capítulos, sendo que os

próximos estão descritos a seguir.

O Capítulo 2 elucida os principais aspectos do reconhecimento de padrões, principalmente no que diz respeito ao reconhecimento da fala humana. O capítulo aborda também questões relevantes sobre as redes neurais, o som e a voz humana.

No Capítulo 3 descreve-se sucintamente e objetivamente a metodologia empregada no desenvolvimento deste trabalho, apresentando o tipo de pesquisa realizada e os procedimentos e ferramentas utilizados na sua realização.

O Capítulo 4 oferece uma discussão sobre os resultados alcançados neste trabalho, e o Capítulo 5 apresenta as considerações finais desta pesquisa.

2. REFERENCIAL TEÓRICO

Kasabov (1998) define a inteligência artificial (IA) como o conjunto de métodos, ferramentas e sistemas utilizados para a resolução de problemas que normalmente requerem a inteligência humana. Sendo que o termo inteligência é descrito como a habilidade efetiva de aprendizagem, de reação adaptativa, de realizar tomadas de decisões apropriadas, de comunicação através de linguagens e imagens, e de compreensão.

Tem-se, portanto, como principal objetivo da IA o desenvolvimento de métodos e sistemas para a resolução de problemas que são solucionados usualmente pela atividade intelectual dos humanos. Neste contexto encontra-se, por exemplo, o reconhecimento de imagens, o processamento de linguagens e da fala e as atividades de planejamento e predição. A conseqüência deste paradigma é a evolução natural dos sistemas computacionais de informação e a conseqüente elaboração de modelos que simulam os organismos vivos e, em particular o cérebro humano.

Assim as redes neurais artificiais (RNA) surgiram como um campo da ciência da computação ligado à inteligência artificial que busca implementar modelos matemáticos inspirados nos neurônios biológicos e nos sistemas nervosos, tornando-se capazes de solucionar problemas através do próprio aprendizado (VEELENTURF, 1995; FERNEDA, 2006).

2.1. Reconhecimento de Padrões

Segundo Duda et al. (1973) a facilidade com que nós reconhecemos um rosto, compreendemos palavras faladas, lemos caracteres manuscritos, identificamos pelo tato as chaves do carro em nossos bolsos, e decidimos se uma maça está madura pelo cheiro é um processo completamente complexo que exemplifica as bases do reconhecimento de padrões. O reconhecimento de padrões pode ser descrito como o ato de capturar dados brutos e realizar uma ação baseada na categoria de seu padrão.

O reconhecimento de padrões é uma disciplina científica cujo objetivo é a classificação de objetos em um número de categorias ou classes. Dependendo da aplicação, estes objetos podem ser imagens ou sinais de formas de ondas, ou ainda, algum tipo de medida que precisa ser classificada. Para referenciar esses objetos usa-se o termo genérico de padrões.

Reis (2001) elucida algumas áreas de aplicação para o reconhecimento de padrões, dentre as quais destaca-se:

- Comunicação do homem com a máquina: reconhecimento automático da fala, reconhecimento da escrita, compreensão da fala, compreensão das imagens e processamento da linguagem natural.
- Defesa: reconhecimento, orientação e controle automático de alvos.
- Medicina: diagnose médica, análise de imagens e classificação de doenças.
- Veículos: controladores de automóveis, aviões, trens e barcos.
- Reconhecimento de identidade: identificação a partir da fala, escrita manual, impressões digitais e fotografias.
- Estudo e estimativa de recursos naturais: agricultura, extrativismo, geologia e ambiente.

Theodoridis & Koutroumbas (2003) afirmam que a fala é o meio mais natural para os humanos se comunicarem e trocarem informações. Assim, o objetivo de se contruir máquinas inteligentes que reconheçam uma informação falada foi uma meta buscada tanto pelos cientistas e engenheiros quanto pelos escritores de ficção. As aplicações de tais máquinas são numerosas, elas podem ser usadas, por exemplo, para melhorar a eficiência em um ambiente de manufatura, para controlar máquinas em ambientes hostis, e para ajudar pessoas a controlar máquinas falando diretamente com elas. O maior esforço, o qual já possui considerável sucesso, é a entrada de dados em um computador via microfone. Um software, construído sobre um sistema de reconhecimento de padrões (palavras faladas neste caso), reconhece um texto falado e o converte em caracteres ASCII, os quais são apresentados na tela e podem ser armazenados na memória.

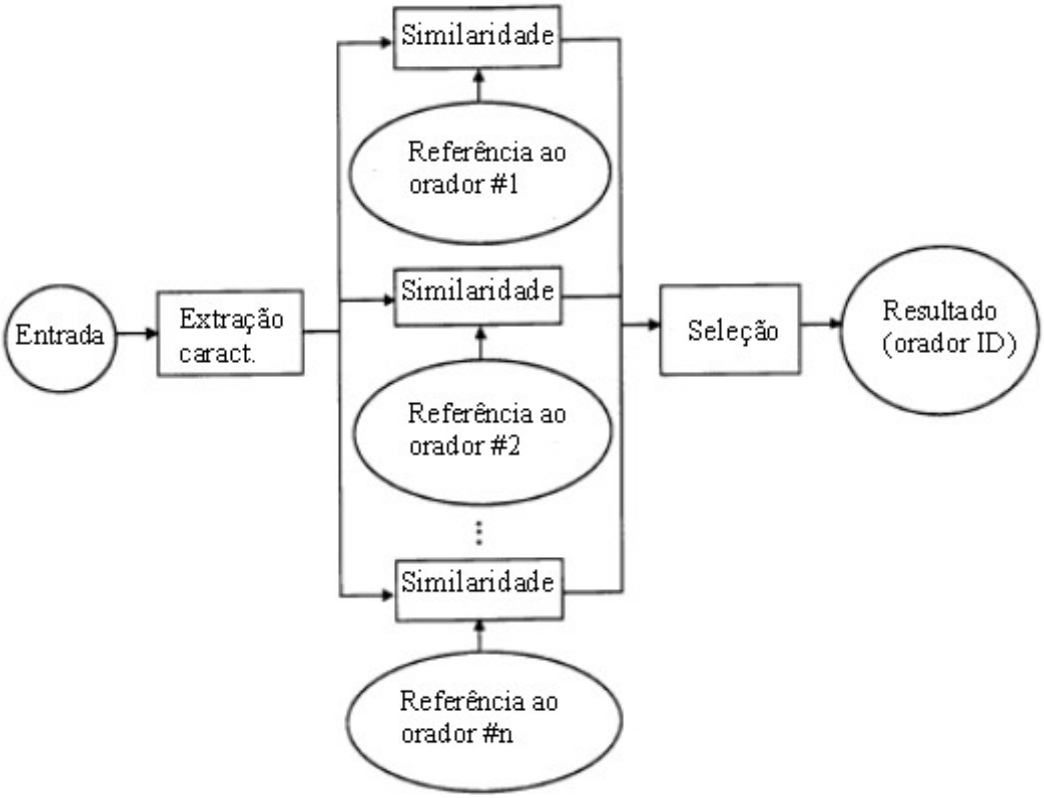
Enquanto que para Ben-Yacoub (1999) a área de reconhecimento de identidade tem recebido muita atenção nos últimos anos. Há uma crescente demanda de sistemas automáticos de identificação de usuários confiáveis para o controle de acesso à lugares e serviços. As técnicas clássicas baseadas em senhas e cartões possuem um certo número de inconvenientes. Senhas podem ser esquecidas ou comprometidas, cartões podem ser perdidos ou roubados e o sistema nestes casos não seria capaz de diferenciar um cliente de um impostor. Muitas outras técnicas tem sido sugeridas e investigadas para o reconhecimento de usuários, principalmente utilizando características que dificultem qualquer forma de falsificação. Sendo assim, surgiu uma nova área de pesquisa relacionada ao reconhecimento de pessoas, denominada biometria, que utiliza características

fisiológicas, tal como a impressão digital, a íris, a face e a voz como forma de identificação única de tais pessoas.

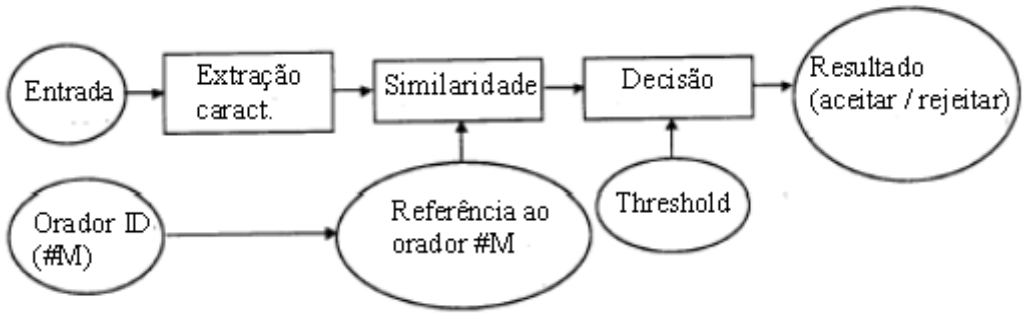
2.1.1. Reconhecimento de identidade através da fala

O reconhecimento de identidade por meio da voz, o qual pode ser classificado em identificação e verificação, é o processo de reconhecer automaticamente quem está falando com base na informação individual contida nas ondas sonoras. Essa técnica torna possível o uso da voz de uma pessoa para verificar sua identidade e para permitir e controlar o acesso à serviços, tais como discagem através da voz, realizar operações bancárias e compras pelo telefone, serviços de acesso a banco de dados, serviços de informações, correio de voz, controle seguro para áreas de informação confidencial, e acesso remoto a computadores. Portanto, a tecnologia de reconhecimento de indivíduos através da voz cria novos serviços que tornam a vida cotidiana mais cômoda, além de ser uma alternativa para propostas forenses.

A Figura 2.1 apresenta a estrutura básica dos sistemas de verificação e identificação de identidade através da fala. Furui (1997) elucida que a identificação de indivíduos através da voz é o processo de determinar um indivíduo registrado dado uma pronúncia. Enquanto que a verificação é o processo de aceitar ou rejeitar a reivindicação de identidade de uma determinada pessoa. A maioria das aplicações que utilizam a voz como forma de confirmar a identidade de um indivíduo são classificadas como sistemas de verificação de identidade.



(a) Identificação de identidade



(b)Verificação de identidade

Figura 2.1 – Estruturas básicas dos sistemas de reconhecimento de identidade através da voz

Os métodos de reconhecimento de identidade através da voz também podem ser divididos em métodos dependentes do texto e métodos independentes do texto. No primeiro caso o sistema requer que o indivíduo diga palavras chaves ou sentenças que possuam o mesmo texto tanto para o treinamento quanto para as tentativas de reconhecimento. Enquanto que para o segundo caso nenhum texto pré-definido é

estabelecido como padrão.

Contudo, ambos os métodos, dependente do texto ou independente do texto, compartilham de um problema. Os sistemas que utilizam tais métodos podem ser facilmente enganados, pois alguém pode gravar a voz de um indivíduo registrado pronunciando algumas palavras, ou sentenças chaves, e usá-la para ser reconhecido como a pessoa registrada. Para tentar resolver esse problema, existem métodos nos quais algumas palavras, tais como dígitos, são usadas como palavras chaves e cada usuário deve pronunciar uma sentença dada aleatoriamente, formada por essas palavras chaves, toda vez que o sistema é usado.

A identificação de indivíduos através da voz está diretamente relacionada as características fisiológicas e comportamentais da pessoa. Essas características residem tanto no invólucro spectral (característica da área vocal) quanto nas características supra-segmentárias (as características da fonte da voz e as características dinâmicas que medem seus vários segmentos).

Campbell (1998) propõe que muitos fatores podem contribuir para erros durante a verificação e a identificação. A Tabela 2.1 lista alguns dos fatores humanos e ambientais que contribuem para esses erros. Esses fatores geralmente saem do escopo dos algoritmos, ou são melhores corrigidos por meios diferentes do que uma alteração/correção do próprio código (por exemplo, um microfone melhor).

Tabela 2.1 – Fontes de erro de verificação

Fontes de erro de verificação
Erros de leitura ou pronúncia de frases
Estado de emoção extrema (ex.: stress ou raiva)
Tempo de variação de alocação do microfone
Acústica do ambiente pobre ou inconsistente (ex.: ruídos)
Canais diferentes (ex.: microfones diferentes para cadastro e para verificação)
Doenças (ex.: resfriados podem alterar as cordas vocais)
Envelhecimento (ex.: as codas vocais mudam com o passar dos anos)

(Fonte: CAMPBELL, 1998)

2.2. Redes Neurais

As redes neurais surgiram como um mecanismo de simular em um computador tanto a estrutura quanto a funcionalidade do cérebro. Para isso buscou-se alternativas para

modelar o neurônio biológico em sua estrutura, funcionalidade, conectividade, interatividade entre os neurônios e, principalmente, na dinâmica do sistema biológico.

Segundo Arbib (2003) um neurônio biológico é uma célula formada por basicamente três estruturas com funções bem específicas e complementares: corpo celular, dendritos e axônio. Os dendritos são responsáveis por captar os estímulos recebidos pelas extremidades do neurônio e os transmitir ao corpo celular, onde são processados. Quando tais estímulos atingem determinado limite, o corpo da célula envia novo impulso que se propaga pelo axônio e é transmitido às células vizinhas por meio de sinapses. Este processo pode se repetir por várias camadas de neurônios e como resultado, a informação de entrada é processada. A Figura 2.2 ilustra de forma simplificada as partes de um neurônio biológico.

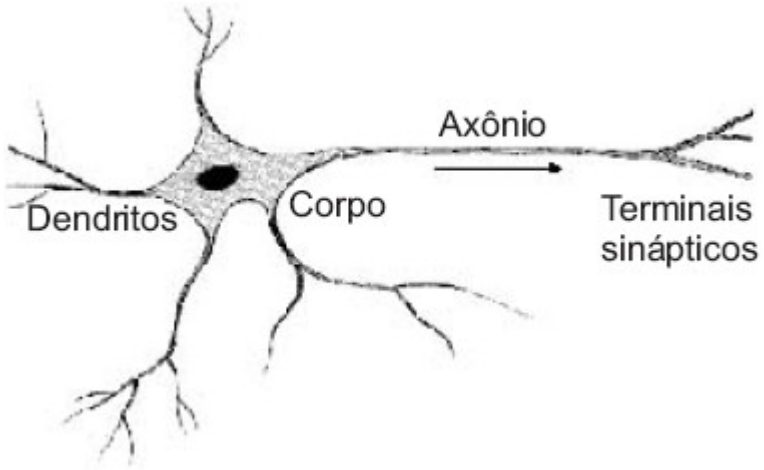


Figura 2.2 – Neurônio biológico

A estrutura do neurônio artificial, inicialmente idealizada por McCulloch e W. H. Pitts em 1943 citado por Rabuñal & Dorado (2006), Dreyfus (2005), Barreto (2002), Jain & Martin (1998), pode ser definida sumariamente como um conjunto de conexões simulando os dendritos, pesos simulando as sinapses, uma função de mapeamento simulando o corpo celular, e uma saída emulando o axônio, conforme ilustrado abaixo.

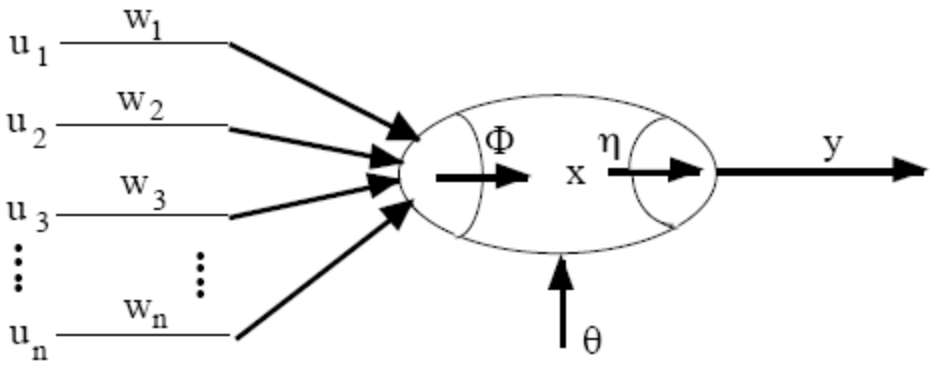


Figura 2.3 – Neurônio artificial

De acordo com a Figura 2.3 tem-se que os valores $<u_1,u_2,u_3,..., u_n>$ representam as entradas/conexões do neurônio e os valores $<w_1,w_2,w_3,..., w_n>$ são os respectivos pesos sinápticos. A função de mapeamento se encontra dividida em duas funções, uma função de ativação (Φ) que vai estabelecer o estado de ativação do neurônio e sua capacidade de disparar um sinal para outro neurônio e uma função de propagação(η), que irá gerar o sinal de saída do neurônio (y). Um valor auxiliar θ é geralmente usado para representar uma polarização, ou seja, um valor abaixo do qual a saída é nula.

Usualmente a função de ativação é a soma das entradas, podendo ser em algumas situações o produto. A função de propagação é que definirá o sinal de saída do neurônio, sendo que as mais usadas são degrau, degrau simétrico, linear e logística sigmoidal.

Uma rede neural artificial é um sistema composto por vários neurônios dispostos em camadas consecutivas. Kröse & Van der Smagt (1996) propõem que o padrão de conexões destas unidades de processamento pode ser categorizado da seguinte maneira:

- Redes *feed-forward,* nas quais o fluxo de dados das unidades de entrada para as unidades de saída é estritamente unilateral. Deve-se notar que o processamento dos dados pode passar por múltiplas camadas de neurônios, porém nenhuma conexão de retorno poderá acontecer, isto é, nenhuma conexão em sentido da saída das unidades de uma camada para a entrada das unidades desta mesma camada ou das camadas anteriores pode ocorrer.
- Redes recorrentes, que podem conter conexões de retorno. Ao contrário das redes *feed-forward*, as propriedades dinâmicas da rede são importantes. Em alguns casos, os valores de ativação das unidades de processamento se submetem a um processo de relaxação no qual a rede evoluirá para um estado estável no qual estas ativações

não mudarão mais. Em outras aplicações, a mudança dos valores de ativação dos neurônios de saída é significante, tal que este comportamento dinâmico constitui a saída da rede.

2.2.1. Perceptron

As redes de uma única camada, e com função de ativação limiar, foram estudadas por um pesquisador chamado Rosenblatt em 1962 e receberam o nome de perceptrons. De acordo com Hu & Hwang (2002) o perceptron é um único neurônio que possui uma função linear ponderada sob a forma de uma função de ativação de limiar. A entrada para este neurônio é um vetor de características de n dimensões na forma $x = (x_1,x_2,...,x_n)$. A função linear, denominada u(x), é a soma ponderada das entradas, como se segue:

$$u(x) = w_0 + \sum_{i=1}^{n} w_i x_i \tag{2.1}$$

Sendo assim, a saída do perceptron, denominada y(x), pode ser obtida diretamente através da função de ativação, como pode ser observado abaixo:

$$y(x) = \begin{cases} 1 & u(x) \geq 0 \\ 0 & u(x) < 0 \end{cases} \tag{2.2}$$

O perceptron pode ser usado tanto para problemas de detecção quanto para problemas de classificação. Por exemplo, o vetor de pesos $w = (w_1,w_2,...,w_n)$ pode representar um padrão para um certo problema. Se o vetor de características de entrada x combinar inteiramente com w tal que seu produto interno exceda um determinado limiar $-w_0$, a saída será +1, indicando a detecção de um padrão.

O vetor de pesos w necessita ser determinado para se aplicar o modelo do perceptron. Para isso, frequentemente, um conjunto de exemplos de treinamento $\{x(i),d(i); i \in I_r\}$ e um conjunto de exemplos de teste $\{x(i),d(i); i \in I_t\}$ são dados. Aqui, $d(i)(\in \{0,1\})$ é o valor da saída desejada de y(x(i)) se o vetor de pesos w for escolhido corretamente, enquanto que I_r e I_t representam os conjuntos disjuntos de treinamento e teste. Através da apresentação dos exemplos de treinamento ao perceptron um algoritmo de aprendizagem pode ser aplicado iterativamente para estimar o valor correto de w. Assim, a formulação do algoritmo de aprendizagem é a seguinte:

$$(2.3) \quad \mathbf{w(k+1) = w(k) + \eta * (d(k) - y(k)) * x(k)}$$

Em que y(k) é computado usando as equações (2.1) e (2.2). Na equação (2.3), η é a taxa de aprendizagem, sendo que é um parâmetro escolhido pelo implementador no intervalo compreendido entre ($0 < \eta < 1/|x(k)|_{max}$), sendo que $|x(k)|_{max}$ é a máxima magnitude dos exemplos de treinamento {x(k)}. O índice k é usado para indicar que os exemplos de treinamento são aplicados seqüencialmente ao perceptron. Cada vez que um exemplo de treinamento é usado, a saída correspondente y(k) ao perceptron é então comparada com a saída desejada d(k). Se elas são aproximadamente idênticas, significa que o vetor de pesos w está correto para aquele exemplo de treinamento, assim, os pesos do vetor não mudarão. Contudo, se y(k) for diferente de d(k), então w será atualizado com uma pequena alteração na direção do vetor de entrada x(k). Isto prova que se os exemplos de treinamento são linearmente separáveis, o algoritmo de aprendizagem do perceptron convergirá para uma possível solução do vetor de pesos dentro de um número finito de iterações. Porém, se os exemplos de treinamento não são linearmente separáveis, então o algoritmo não irá convergir, tendo-se um valor de η fixado e diferente de zero.

Há várias dificuldades em se aplicar o modelo do perceptron para resolver problemas de detecção de sinal e classificação de padrão no mundo real, dentre os quais se destaca:

- A transformação não-linear que extrai o vetor de características x apropriado não é específica.
- O algoritmo de aprendizagem do perceptron não irá convergir (dado uma taxa de aprendizagem fixa) se o padrões de características de treinamento não forem linearmente separáveis.
- Se os padrões de características forem linearmente separáveis, não se pode saber em quanto tempo o algoritmo irá convergir para um vetor de pesos que corresponda à um hiperplano que separe esses padrões de características.

Uma alternativa para tentar resolver estas dificuldades é o emprego de redes MLP (*Multilayer Perceptron*).

2.2.2. Multilayer Perceptron

O modelo de uma rede neural *multilayer perceptron* (MLP) consiste em uma rede *feed-forward* em camadas de neurônios de McCulloch e Pitt's. Cada neurônio na MLP tem uma função de ativação não-linear que é diferenciável, sendo que as funções de ativação mais frequentemente usadas para a MLP são: função de limiar, função sigmoidal e a função tangente hiperbólica.

Uma configuração típica de MLP pode ser observada na Figura 2.4. Cada círculo representa um neurônio individual. Estes neurônios são organizados em camadas, denominadas por camada escondida #1, camada escondida #2, e camada de saída. Enquanto que as entradas também são organizadas em uma camada de entrada, porém, usualmente, nesta camada não há efetivamente nenhum neurônio. O nome de camada escondida se refere ao fato de que as saídas destes neurônios irão alimentar os neurônios da camada superior e, portanto, ficarão escondidas do usuário que apenas observa a saída dos neurônios da camada de saída. A Figura 2.4 ilustra uma configuração popular de MLP onde as interconexões são realizadas apenas entre neurônios de camadas sucessivas da rede, entretanto, na prática, interconexões cíclicas entre neurônios também são permitidas.

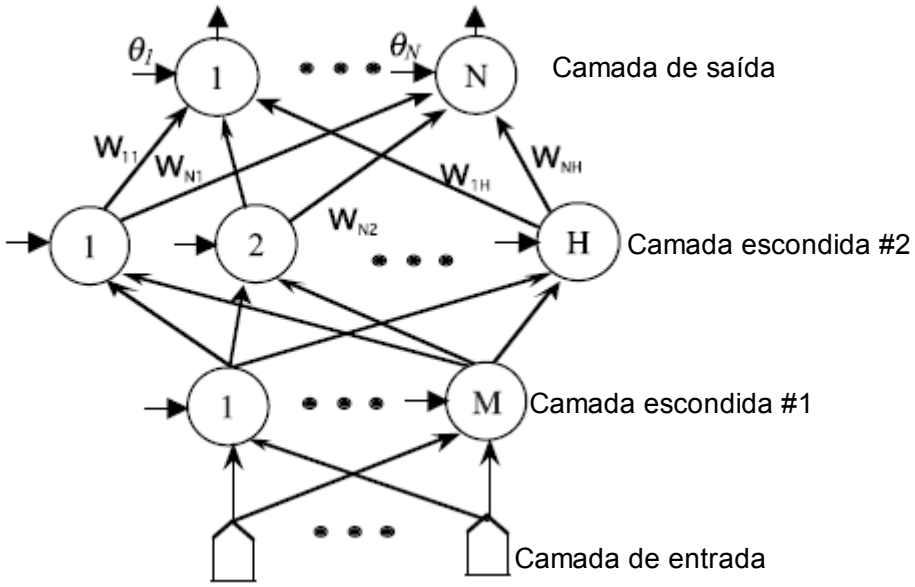


Figura 2.4 – Configuração de uma *multilayer perceptron* de três camadas

(Fonte: HU & HWANG, 2002)

Uma MLP providencia um mapeamento não-linear entre entradas e saídas. É provado que com um número suficiente de neurônios escondidos, uma MLP com aproximadamente duas camadas escondidas é capaz de aproximar um mapeamento

complexo dentro de um intervalo de tempo finito e exeqüível (KRÖSE & VAN DER SMAGT,1996).

2.2.3. O treinamento de redes neurais

Treinamento é o procedimento algorítmico segundo o qual os parâmetros dos neurônios da rede são estimados, para que a rede neural cumpra, tão exata quanto possível, a tarefa que lhe foi atribuída. Dentro deste contexto, Dreyfus (2005) considera dois tipos de categorias de treinamento: treinamento supervisionado e treinamento não-supervisionado.

Uma rede neural *feed-forward* computa uma função não-linear de acordo com suas entradas. Portanto, pode ser atribuído à uma rede a tarefa de computar uma função não-linear específica. Sendo assim, duas situações podem ocorrer:

- A função não-linear é analiticamente conhecida: logo a rede executa a tarefa de aproximação da função.
- A função não-linear não é analiticamente conhecida, mas um número finito de valores numéricos da função são conhecidos, em muitas aplicações, esses valores não são conhecidos exatamente porque eles são obtidos através de experimentos executados em processos físicos, químicos, financeiros, econômicos, biológicos, etc.: nesse caso, a tarefa que é atribuída para a rede é a de aproximação da função de regressão dos dados disponíveis, conseqüentemente se torna um modelo estático do processo.

Na maioria das aplicações, redes neurais *feed-forward* com treinamento supervisionado são usadas nas situações do segundo caso. O treinamento pode ser considerado "supervisionado" quando a função que a rede deve implementar é conhecida em alguns, ou todos, dos seus pontos: um "professor" providencia "exemplos" de valores de entrada e os valores correspondentes de saída, ou seja, o resultado da tarefa que a rede deveria executar. Caso a saída da rede não coincida com os valores de saída passados, modifica-se os valores das conexões sinápticas no sentido de se fazer a saída se aproximar da desejada. Como a cada exemplo apresentado uma correção é introduzida depois de se observar a saída da rede este é o caso de treinamento supervisionado.

Também pode ser atribuída a uma rede neural *feed-forward* uma tarefa de análise de dados ou visualização, onde um conjunto de dados, descritos por um vetor com um

grande número de componentes, é disponível. Deseja-se agrupar esses dado, de acordo com algum critério de similaridade que não é conhecido a priori. Métodos de agrupamento são bem conhecidos na área de estatística, e redes neurais *feed-forward* podem ser utilizadas para executar esta tarefa: a partir de uma representação dos dados de alta dimensão, encontrar a representação de menor dimensão (usualmente 2 dimensões) que preserve a similaridades. Assim, nenhum "professor" está presente nessa tarefa e, portanto, o treinamento da rede deveria "descobrir" as similaridades entre os elementos do banco de dados. O treinamento não-supervisionado é caracterizado pelo fato de que quando é necessário realizar modificações nos valores das conexões sinápticas não se usa informações sobre se a resposta da rede foi correta ou não. Usa-se apenas a descoberta de padrões de similaridades.

2.2.4. Regras de aprendizagem

Basicamente o conceito de treinar, ou aprender, das redes neurais consiste em atribuir valores às conexões sinápticas. Em alguns casos esses valores são colocados representando um certo conhecimento, em outros usa-se um algoritmo para encontrá-los, e a este algoritmo chama-se algoritmo de aprendizagem (BARRETO,2002). Dentre os vários algoritmos de aprendizagem serão abordados os seguintes: Aprendizado Hebbiano, Regra Delta e *Backpropagation*.

Aprendizado Hebbiano. A lei de Hebb pode ser considerada uma das mais antigas regras de aprendizagem usada, uma extensão dessa lei é:

> *A intensidade de uma conexão entre dois neurônios A e B deve ser ajustada uma quantidade proporcional ao valor da ativação simultânea dos dois neurônios. Se no entanto A tentar excitar B e não conseguir a conexão deverá ser enfraquecida.*

Uma característica relevante da lei de Hebb é sua propriedade de localidade, ou seja, para se alterar o valor de uma conexão sináptica apenas as informações locais à sinápse em questão serão usadas, proporcionando plausibilidade biológica ao algoritmo. Assim tem-se:

$$\Delta w_{ij} = \eta x_i o_j \quad (2.4)$$

Em que:

- w_{ij} : intensidade da conexão entre os neurônios i e j
- Δw_{ij} : acréscimo da intensidade da conexão entre os neurônios i e j
- η : taxa de aprendizagem
- x_i : estado de ativação do neurônio i
- o_j : saída do neurônio j

Apesar da plausibilidade biológica o uso da lei de Hebb nesta forma apresenta alguns inconvenientes. Sendo assim, utiliza-se usualmente versões mais sofisticadas (como por exemplo a regra delta).

Regra Delta. Tem-se que a expressão empregada na lei de Hebb é muito simplificada, pois considerando-se uma sinapse real conclui-se que:

- O valor da modificação da intensidade da conexão sináptica para as mesmas excitações dos neurônios envolvidos podem variar com o tempo.
- A modificação Δw_{ij} pode depender de w_{ij}, o que seria um efeito não-linear. Isto ocorre como um efeito de saturação do valor da conexão sináptica.
- Pode-se inferir que a modificação da intensidade da conexão sináptica dependa também dos neurônios vizinhos.

Um modelo mais completo seria a Regra de Widrow-Hoff ou Regra Delta que pode ser expressa como:

$$\Delta w_{ij} = \eta(x_i - d_i)o_j \quad (2.5)$$

Em que d_i representa o estado de ativação desejado para o neurônio i. A Regra Delta é biologicamente plausível pois usa apenas a informação local à sinapse para o aprendizado. Como é uma generalização da lei de Hebb ela pode efetuar uma otimização que pode ser interpretada como o modelo matemático de um mecanismo de seleção (BARRETO, 2002).

Backpropagation. O método *Backpropagation* pode ser considerado como a

generalização da Regra Delta para redes *feed-forward* com mais de duas camadas. Neste caso, pelo menos uma camada de neurônios não está envolvida diretamente com a camada de entrada ou a camada de saída e é portanto interna à rede. Essa camada e suas conexões quando aprendem a efetuar determinada função agem como se houvesse uma representação interna da solução do problema. De acordo com Rao (1995) o *Backpropagation* é uma regra de treinamento supervisionado, apresenta-se à rede um exemplo e verifica-se a saída da mesma, saída esta que é comparada à saída desejada gerando-se um erro relacionado. Calcula-se o gradiente deste erro com relação aos valores sinápticos da camada de saída, que é atualizada conseqüentemente e, então, esse erro é utilizado como base para o ajuste dos pesos das conexões entre a camada de entrada e as camadas escondidas. Ajustando o conjunto de pesos de pares de camadas e recalculando a saída em um processo iterativo que executa até que seja alcançado um erro que seja menor que uma tolerância desejada.

A maioria dos programas que utilizam RNA dispõem de uma implementação do *backpropagation*, ou na sua forma original (usando gradiente) ou em uma forma modificada que melhore a performance da regra. Contudo, é importante destacar que nem toda rede *feed-forward* pode ser treinada usando *backpropagation*, pois é necessário que a função de ativação da rede seja derivável.

Para Freeman & Skapura (1991) o procedimento básico para treinar uma rede se baseia na seguinte descrição:

1. Aplicar um vetor de entrada para a rede e calcular os valores de saída correspondentes.
2. Comparar a saída atual com a saída correta e determinar o valor do erro.
3. Determinar em qual direção (+ ou -) cada peso deverá ser alterado para reduzir o erro.
4. Determinar o valor da alteração para cada peso.
5. Aplicar as correções para os pesos.
6. Repetir os passos de 1 a 5 com todos os dados de treinamento até o erro de todos os dados no conjunto de treinamento seja reduzido para um valor aceitável.

Tem-se portanto que dada uma taxa de erro, os pesos poderão ser atualizados de

acordo com a equação:

$$(2.6) \quad \mathbf{w_{ij}\,(k+1) = w_{ij}\,(k) + \eta * e_i(k) * x_i(k)}$$

Em que:

- k: iteração corrente do algoritmo de aprendizagem
- $w_{ij}(n)$: peso da conexão sináptica entre os neurônios i e j na iteração n
- $e_i(k)$: taxa de erro do neurônio i na iteração k
- $x_i(k)$: vetor dos valores de entrada na iteração k do neurônio i
- η: taxa de aprendizagem

Uma outra possível maneira de se atualizar os pesos é através de uma reformulação da equação acima, inserindo o conceito de *momentum*. O *momentum* nada mais é do que uma memória (incremento anterior) do algoritmo proposto, com a finalidade de aumentar a velocidade e estabilizar a convergência da rede. Como pode ser observado pela equação abaixo:

$$(2.7) \quad \mathbf{w_{ij}\,(k+1) = w_{ij}\,(k) + \eta * e_i(k) * x_i(k) + \beta[w_{ij}\,(k) - w_{ij}\,(k-1)]}$$

Onde β é a taxa de *momentum*. Vale destacar que há portanto dois parâmetros que precisam ser escolhidos: a taxa de aprendizagem, e a constante de *momentum*. Ambos os parâmetros deveriam ser escolhidos no intervalo de [0,1], mas na prática, η frequentemente assume um pequeno valor, normalmente entre $0 < \eta < 0{,}3$, e β assume usualmente um grande valor, basicamente entre $0{,}6 < \beta < 0{,}9$.

2.3. O Som e a Voz Humana

Segundo Sears et al. (1996) uma onda é uma perturbação sobre um meio em equilíbrio, que se propaga através do tempo de uma região do espaço para outra, ou ainda, como sendo qualquer sinal que se transmite de um ponto a outro do meio com velocidade determinada, sem que haja transporte direto de matéria de um desses pontos ao outro. Este é um dos conceitos mais bem difundidos pela Física e na vida cotidiana existem vários exemplos de movimentos ondulatórios: as ondas produzidas pela colisão de uma pedra

sobre a superfície de um lago, a luz, ondas em molas e o som exemplificam este tipo de fenômeno.

Tem-se que as ondas podem ser classificadas em mecânicas e eletromagnéticas. As ondas mecânicas se caracterizam pela necessidade de um meio material pelo qual elas irão se propagar à medida que o estado de equilíbrio é deslocado. Considera-se que um meio é constituído por um grande número de partículas materiais (gasosas, líqüidas ou sólidas), sendo que cada uma é ligada ou acoplada à partícula adjacente por um material elástico. Nota-se que se uma das extremidades do meio for perturbada, o deslocamento resultante não ocorrerá simultaneamente em todos os pontos do meio. Pois devido às propriedade elásticas do meio, o distúrbio será transmitido de uma partícula à seguinte, progredindo conseqüentemente através do meio. Em contrapartida, as ondas eletromagnéticas não necessitam de nenhum meio material para se propagar, pois são constituídas pela oscilação dos campos elétricos e magnéticos, como é o caso das ondas de luz.

Citar as ondas de água, as ondas luminosas e as ondas sonoras como exemplos de movimento ondulatório é uma forma de classificação que considera suas propriedades físicas gerais, porém pode-se classificá-las analisando-se a direção de propagação da própria onda. Uma onda é dita transversal se o movimento das partículas materiais que transmitem a onda é perpendicular à sua direção de propagação (exemplo: ondas em uma corda), Figura 2.5(a). Enquanto que uma onda longitudinal é aquela em que o movimento das partículas possui a mesma direção de propagação da onda, como é o caso das ondas sonoras, alvo de estudo do presente trabalho, Figura 2.5(b). Porém, nem todas ondas são exclusivamente transversais ou exclusivamente longitudinais. Algumas ondas, como as que se propagam na superfície da água, descrevem trajetórias elípticas enquanto se propagam e podem ser caracterizadas como transversais e como longitudinais (RESNICK & HALLIDAY, 1986).

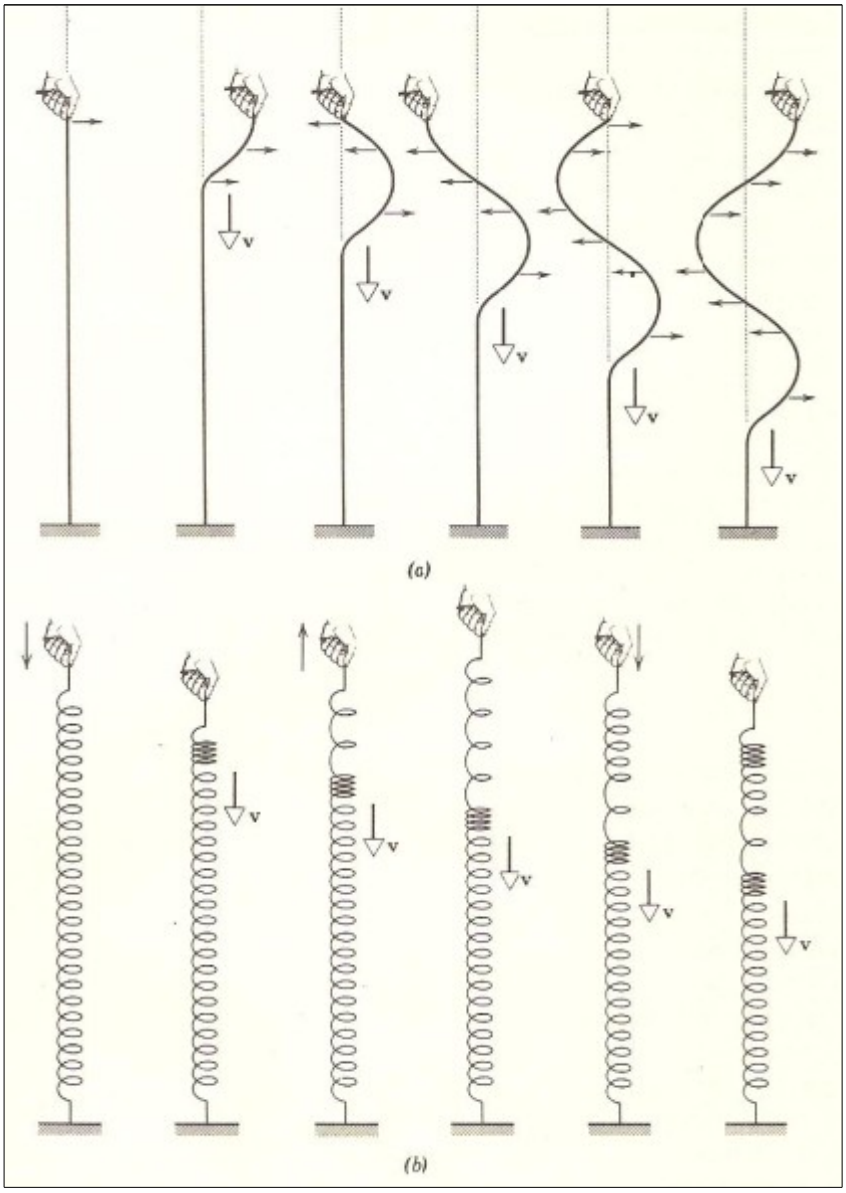


Figura 2.5

(a) Onda transversal. (b) Onda longitudinal.

(Fonte: RESNICK & HALLIDAY, 1986)

As ondas podem ser ainda classificadas quanto ao número de dimensões em que elas se propagam. As ondas que se movem em uma única dimensão são denominadas unidimensionais, caso das ondas em uma corda ou mola. Denomina-se ondas bidimensionais àquelas que se propagam em duas dimensões (exemplo: ondas na superfície da água). E por fim, as ondas tridimensionais são as que se propagam em três dimensões, como a luz e o som.

Pode-se também classificar as ondas de acordo com o comportamento da partícula

do meio que transporta a onda, durante o tempo no qual ela se propaga. Chama-se pulso a onda única que se propaga no meio durante um período de tempo, voltando ao seu estado de equilíbrio logo em seguida. Enquanto que o movimento contínuo, periódico ou não, de pulsos sucessivos é denominado de trem de ondas.

2.3.1. Elementos de uma onda

Para se compreender a natureza das ondas é necessário definir os principais elementos que as compõem, dentre os quais se destaca:

- Amplitude, deslocamento máximo alcançado por uma partícula do meio ao oscilar, ou ainda, a distância máxima entre a posição de equilíbrio e a posição extrema ocupada pela partícula que oscila.
- Comprimento de onda, representa a distância entre dois pontos consecutivos da onda que possuem a mesma fase, ou seja, a distância entre duas cristas ou dois vales sucessivos de uma onda.
- Período, tempo necessário para que a onda percorra a distância de um comprimento de onda.
- Freqüência, representa o número de vibrações completas que a partícula do meio efetua por unidade de tempo. Sendo que o valor desta freqüência é sempre igual à freqüência da fonte que deu origem à onda.
- Velocidade de propagação, é a velocidade que a partícula alcança ao percorrer a distância de um comprimento de onda no intervalo de tempo de um período.

Como pode ser observado na Figura 2.6 alguns dos elementos mencionados acima foram ilustrados em uma onda senoidal.

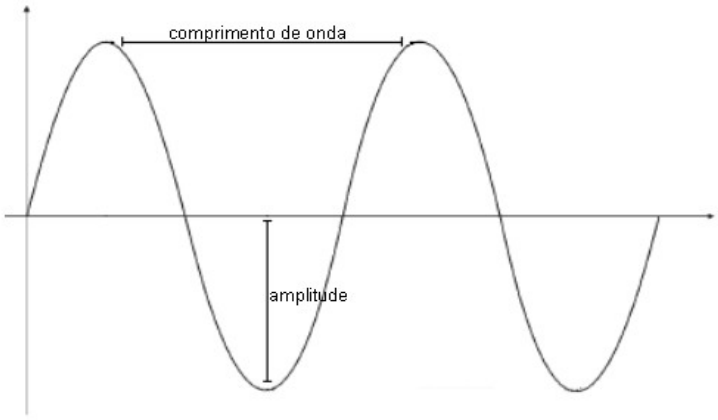


Figura 2.6 – Elementos de onda

Além destes elementos de ondas tem-se também outra importante característica das ondas a ser definida, denominada frente de onda. Uma frente de onda é a superfície por onde passam todas as partículas do meio em um dado instante do tempo, quando estas estão sendo perturbadas. Conseqüentemente a frente de onda sempre divide a região perturbada do meio da que ainda não foi perturbada (Figura 2.7).

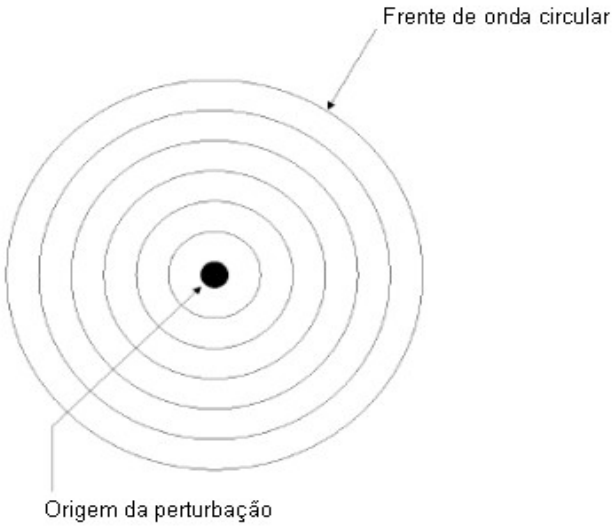


Figura 2.7 – Frente de onda

2.3.2. O som

Por definição tem-se que o som é uma onda longitudinal que se propaga em um meio gasoso, líqüido ou sólido. As ondas sonoras são provocadas quando um corpo vibra e ocasiona uma perturbação no meio. Como abordado anteriormente esta propagação se deve

às interações das partículas materiais que compõem o meio, onde estas partículas apenas vibram em torno das respectivas posições de equilíbrio (NUSSENZVEIG,1997; TIPLER,1991). As ondas sonoras normalmente se propagam em todas as direções a partir da fonte, com amplitudes que dependem da direção e da distância entre o ponto receptor e a fonte emissora.

A audição humana é sensível a ondas na faixa de freqüência entre aproximadamente 20 Hz e 20 000 Hz. As características que se podem distinguir em uma onda sonora são sua intensidade, altura e timbre. A intensidade se refere a transferência de energia associada a propagação da onda, ou seja, é a taxa média da energia transmitida por unidade de tempo e de área. Tem-se que para uma dada freqüência, quanto maior for a pressão máxima causada pela flutuação de pressão da onda sonora, mais elevada será a intensidade sonora. Contudo, a relação entre esta pressão máxima e esta intensidade sonora varia de uma pessoa para outra. Isto se deve principalmente ao fato de que o ouvido não possui a mesma sensibilidade para todas as freqüências do intervalo audível.

A unidade de nível de intensidade é o Bel (nome dado em homenagem a Alexander Graham Bell): duas ondas sonoras diferem de 1 Bel quando a intensidade de uma é 10 (dez) vezes maior que a da outra. Na prática usa-se o decibel (dB), que corresponde a 0,1 Bel. A Tabela 2.2 a seguir quantifica alguns exemplos de ondas sonoras em seus respectivos valores em decibéis.

Tabela 2.2 – Exemplos de níveis de intensidade sonora

Fonte ou descrição de ruído	*Nível de ruído, db*
Limiar da dor	120
Rebitamento	95
Trem elevado	90
Tráfego urbano pesado	70
Conversação ordinária	65
Automóvel silencioso	50
Rádio moderado em casa	40
Sussurro médio	20
Roçar de folhas	10
Limite de audição	0

Fonte: (NUSSENZVEIG,1997)

A altura de uma onda sonora corresponde à sensação que permite distinguir entre sons graves e sons agudos. A altura relaciona-se diretamente com a freqüência, para uma onda de intensidade constante quanto maior a freqüência, maior a altura (som agudo) e

quanto menor a freqüência, menor a altura (som grave).

O timbre é a característica que permite diferenciar duas ondas sonoras, de mesma freqüência e de mesma intensidade, desde que suas formas de ondas sejam diferentes. Desta forma, o ouvido é capaz de distinguir claramente a diferença entre uma mesma nota emitida por um piano, violino, flauta ou pela voz humana, por exemplo. O timbre representa uma espécie de "coloração" do som (NUSSENZVEIG,1997).

Uma outra importante definição a ser realizada é o conceito de ruído, que é a combinação de todas as freqüências, mas não apenas de freqüências harmônicas, ou seja, freqüências múltiplas de uma freqüência fundamental. Exemplos de ruídos são o som do vento e o silvo que ocorre ao se pronunciar a consoante "s".

Tem-se também a definição da envoltória de um sinal, que consiste em uma forma de onda f(t) deslocada por uma constante C. Assim sendo, para se recuperar um sinal f(t) é necessário simplesmente detectar sua envoltória.

As ondas audíveis podem se originar a partir de cordas vibrantes (violino, cordas vocais humana), colunas de ar em vibração (órgão, clarineta), placas e membranas vibrantes (alto-falante, tambor) e hastes vibrantes (diapasão). Todos estes elementos vibrantes comprimem o ar em sua volta, em seu movimento para frente, rarefazendo-o em seu movimento de volta. O ar transmite estas perturbações em forma de onda que se propaga a partir da fonte. Ao chegar no ouvido, estas ondas originam a sensação sonora.

2.3.3. Voz Humana

Segundo Paula (2000) a fala é uma das capacidades ou aptidões que os seres humanos possuem de comunicação, manifestando seus pensamentos, opiniões e sentimentos através de vocábulos, que podem ser transladados textualmente quando necessário. Sendo assim, consiste no principal sinal dentre os abordados pela linguagem natural, tal como os ideogramas, gestos, gritos, trejeitos e tantos outros tipos de linguagem corporal.

Tem-se que qualquer emissão humana – falada, cantada ou até mesmo uma simples exclamação – apresenta 3 funções:

- Função de representação: a voz comunica alguma coisa, ou seja, seu uso está relacionado ao conteúdo da mensagem verbal.
- Função de expressão – a voz revela alguma coisa do falante, como por exemplo sua idade, seu nível sócio-econômico-cultural, seu estado emocional, etc.

- Função de apelo – a voz deseja e provoca uma reação no ouvinte, o que significa que existe sempre uma intenção, frequentemente inconsciente, no tipo de voz que se utiliza no discurso.

O processo pelo qual os seres humanos produzem palavras e orações audíveis para se comunicar possibilita a obtenção de informações a respeito do ambiente no qual o indivíduo está inserido. Grande parte das espécies animais possuem algum nível de comunicação, mas o homem, em virtude de sua complexidade social, adquiriu o mais alto grau de comunicação conhecido, dentre as quais, a voz tem uma extrema relevância.

A voz é uma onda sonora, e como tal possui todas as características das ondas sonoras. Sobre a voz humana pode-se dizer que consiste em um som, ou um conjunto de sons, emitido pelo aparelho fonador.

Young (2003) afirma que quando uma onda sonora entra no ouvido, ela produz vibrações do tímpano que, por sua vez, produzem oscilações nos minúsculos ossos do ouvido médio chamados de ossículos. Estas oscilações são então transmitidas ao ouvido interno, que está cheio de líqüido; o movimento deste fluido perturba as células capilares no ouvido interno, as quais transmitem impulsos ao nervo que se liga ao cérebro transportando a informação de que existe um som.

Segundo Bloch (1958) citado por Tafner (1996) a intensidade sonora é resultado da amplitude das vibrações das cordas vocais e da freqüência da onda sonora, pois os sons graves possuem maior amplitude do que os agudos. Ainda sobre a voz humana, a qualidade vocal, termo antes chamado de timbre (usualmente usado para instrumentos musicais), é o termo atualmente usado para designar o conjunto de características que a identificam. A qualidade vocal se relaciona à composição dos harmônicos da onda sonora. Isso ocorre porque, o que de fato se ouve, é o som resultante da superposição de vários sons de freqüências diferentes. No entanto, a freqüência do som ouvido é igual à do som de menor freqüência, denominada de freqüência fundamental. O que a diferencia é a presença dos sons harmônicos.

Aspectos como a freqüência fundamental também tem sido alvo de estudos, sendo que as freqüências fundamentais das vozes masculinas podem variar de 80 a 150 Hz, as femininas de 150 a 250 Hz, e as infantis encontram-se acima de 250 Hz. Ao analisar a intensidade sonora tem-se que em uma conversa normal a intensidade oscila entre 40 e 50 dB, sendo que a intensidade máxima da voz humana pode variar entre 60 e 120 dB.

Para Brunelli & Falavigna (1995) o sinal de voz contém basicamente dois tipos de informação: a individual e a fonética. Elas possuem efeitos mútuos que são difíceis de se separar, e isto representa um dos principais problemas nos desenvolvimento de sistemas de reconhecimento automático de pessoas e de fala. A conseqüência é que sistemas de reconhecimento de pessoas executam melhor sobre segmentos da fala cujos conteúdos fonéticos são específicos, enquanto que sistemas de reconhecimento de fala possuem alta precisão quando tratam da voz de uma pessoa em particular.

3. METODOLOGIA

O desenvolvimento do presente trabalho apresentou etapas bem delimitáveis, podendo-se descrevê-las nas seguintes categorias:

- Levantamento de técnicas de implementação de redes neurais, assim como suas características e limitações.
- Estudo das particularidades do sinal sonoro, principalmente no que se refere à voz humana.
- Captura e tratamento dos dados de treinamento e teste.
- Implementação da rede neural.
- Testes e análises do sistema proposto.

3.1. Tipo de Pesquisa

Conforme Zambalde (2005) e analisando-se os aspectos referentes ao método científico pode-se constatar que o presente trabalho é de natureza aplicada, com objetivos exploratórios e utiliza procedimentos de caráter experimental fundamentados em referências bibliográficas e documentais.

Tem-se como pesquisa aplicada por ter como principal finalidade a obtenção de uma rede neural que seja capaz de reconhecer um indivíduo através de sua voz.

O trabalho possui objetivos de natureza exploratória por buscar encontrar teorias e práticas que possibilitem a construção de uma ferramenta de reconhecimento.

Através de experimentos baseados na codificação/implementação de técnicas abordadas em bibliografias sobre redes neurais observa-se que será possível construir o sistema proposto.

3.2. Ferramentas Utilizadas

Para a consecução dos objetivos deste trabalho utilizou-se o Scilab versão 4.1.1 (Toolbox ANN) como ferramenta para o desenvolvimento/implementação da rede neural. Software este também responsável pela execução dos casos de teste e pelo tratamento, depois de capturado, do sinal sonoro.

Foi ainda utilizado o aplicativo de gravação de som padrão encontrado na instalação do sistema operacional Microsoft Windows 2000 Professional.

Todo o sistema e a realização dos casos de teste foram desenvolvidos em uma máquina AMD 2800+ com 768 Mb (Megabytes) de memória RAM e 80 Gb (Gigabytes) de disco rígido. Para a captura do som foi usado um microfone PH50401, disponível no momento de execução da pesquisa, de baixa impedância, com sensibilidade de 58 dB ± 2dB e cuja freqüência de resposta é de 30 Hz – 16.000 Hz.

3.3. Procedimentos Metodológicos

Este trabalho foi realizado no período de junho de 2007 a dezembro de 2007, iniciando-se por um levantamento bibliográfico e documental sobre o tema proposto.

Em seguida tem-se a etapa de captura do sinal sonoro, durante a qual se realizou a criação de uma base de dados que foi utilizada no treinamento da rede neural e nos testes do sistema. Para a elaboração desta base de dados foi definido um conjunto de palavras chaves válidas que especificou quais palavras deveriam ser capturadas.

Nesta fase pessoas diferentes tiveram suas vozes gravadas pronunciando separadamente cada palavra válida uma quantidade de vezes pré-determinada. É importante notar que o sistema só é capaz de reconhecer um indivíduo se este pronunciar isoladamente a palavra adequada.

Esta restrição se deve ao fato da RNA ser treinada utilizando uma única mesma palavra ou expressão sendo pronunciada por diversas pessoas distintas. Portanto, torna-se de fundamental importância que o indivíduo a ser identificado, assim como as demais pessoas cujas vozes compõem a base de dados, utilize a mesma palavra tanto para o conjunto de treinamento quanto para a execução dos testes da efetividade da rede.

Para o tratamento do sinal optou-se por seguir a metodologia proposta por Tafner (1996) que será descrita posteriormente.

Durante a etapa de implementação foi desenvolvido uma rede neural *feed-forward* com um algoritmo de aprendizagem *backpropagation*.

3.3.1. Tratamento do sinal sonoro

O som quantificado, e por assim dizer, registrado, possui características “cruas” que não são apropriadas para construir um conjunto de treinamento para uma rede neural artificial. Dessa maneira o processo de análise é importante pois faz parte de um mecanismo de refinamento da informação (TAFNER, 1996). Logo, torna-se necessário que seja realizado um tratamento do som com o objetivo de capturar de seu sinal apenas a informação relevante para o treinamento da rede.

Antes de abordar o tratamento propriamente dito torna-se necessário definir um importante conceito relacionado à intensidade do sinal sonoro denominado nível de quantização. A quantização se refere ao domínio da amplitude de um sinal analógico contínuo amostrado em um certo período de tempo, ou seja, trata-se da medida discreta da intensidade do sinal. A discretização da amplitude é definida em termos de bits, por exemplo, uma conversão de 8 bits proporciona uma representação de 2^8 estados, ou 256 níveis de quantização. Sendo que cada um destes estados está representando uma faixa de valores da amplitude.

Assim sendo, o refinamento do sinal foi dividido em 4 (quatro) etapas e estas estão descritas a seguir:

- Eliminação do ciclo negativo do sinal amostrado.
- Redução do sinal amostrado detectando a forma de onda (envoltória).
- Mediação do sinal reduzido.
- Normalização do sinal mediado.

Uma vez que o sinal sonoro esteja amostrado e digitalizado, ele estará sob a forma de um vetor de números. Considerando-se uma digitalização programada para operar com 256 níveis de quantização a uma taxa de amostragem de 8000 Hz, em um segundo de leitura, haverá 8000 valores compreendidos no intervalo de 0 a 255 , que organizados em ordem de leitura, representam o sinal original. Nota-se que como o tempo gasto para pronunciar cada palavra é diferente, assim este vetor possuirá um tamanho diretamente proporcional à palavra falada.

Quando o sinal capturado é lido pelo Scilab os valores correspondentes aos níveis de quantização (de 0 a 255) são convertidos automaticamente, sem perda de suas características, para os valores correspondentes no intervalo compreendido entre [-1,1]. A Figura 3.1 elucida um exemplo de sinal amostrado no qual um indivíduo pronuncia a palavra “teste”.

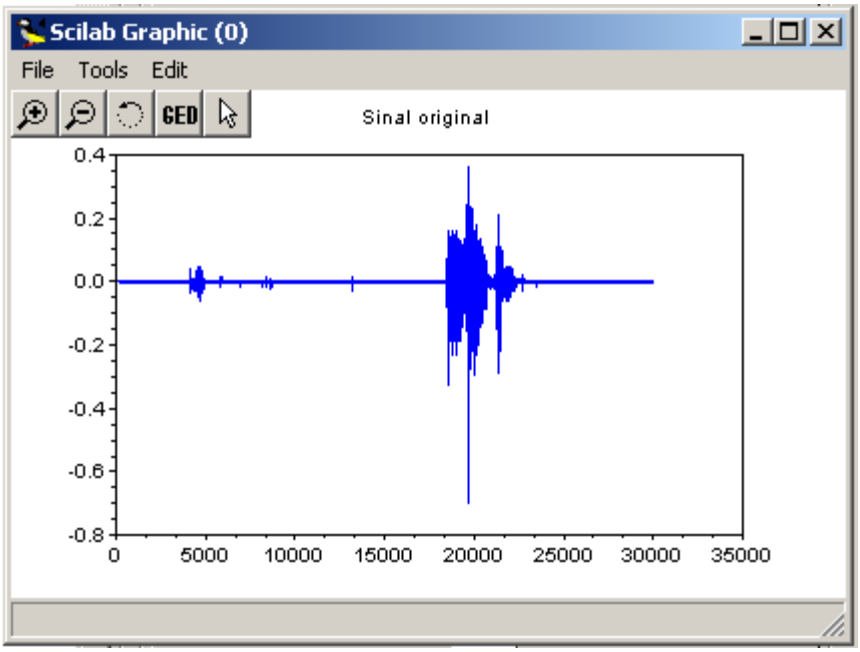


Figura 3.1 – Sinal amostrado original

É importante destacar que antes de prosseguir com os procedimentos de tratamento do sinal sonoro definidos anteriormente é necessário realizar um outro refinamento no sinal amostrado original, eliminando as partes do sinal situadas nas chamadas zonas de silêncio (TAFNER, 1996). As zonas de silêncio são os intervalos inicial e final do sinal amostrado no qual o indivíduo não está pronunciando efetivamente nenhum som. No caso da Figura 3.1 estas zonas de silêncio correspondem aproximadamente aos intervalos [0,15000] e [25000,35000], sendo assim o sinal correspondente à Figura 3.1 sem as zonas de silêncio pode ser observado pela Figura 3.2. Pode-se realizar este refinamento de forma manual ou automatizada, optando-se pela automação deve-se definir um intervalo de amplitude limite, eliminando todos os sinais iniciais e finais compreendidos entre este intervalo (por exemplo, entre -0,1 e 0,1).

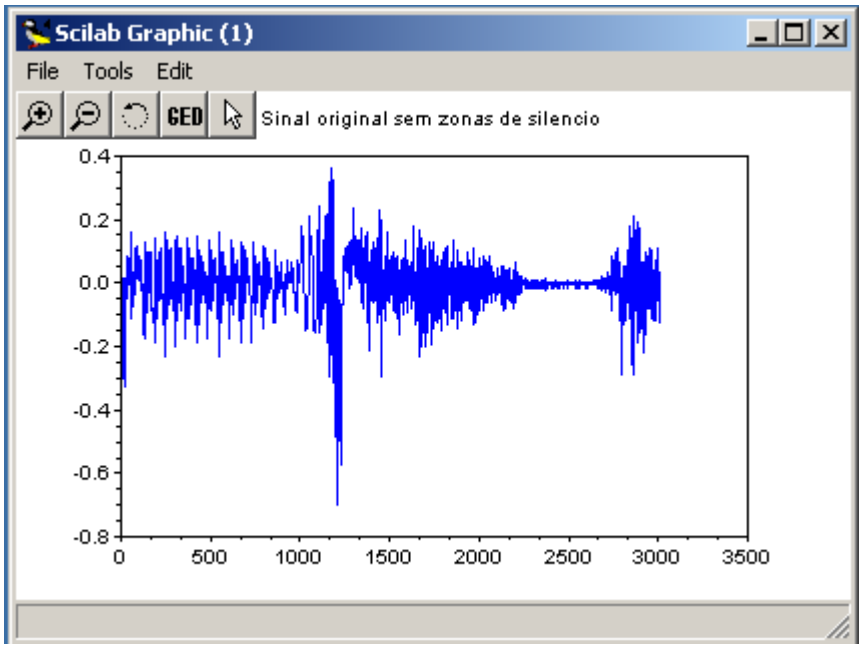


Figura 3.2 – Sinal original sem zonas de silêncio

A etapa de eliminação do ciclo negativo do sinal amostrado determina que todo o sinal negativo da onda sonora deverá ser eliminado. Sendo que essa eliminação consiste em atribuir o valor absoluto respectivo para todos os sinais abaixo da linha de silêncio (valor 0), veja Figura 3.3. Isto deve ser realizado sem perturbar a ordem original dos sinais escolhidos ou, ainda, eliminar a posição em que o sinal foi silenciado.

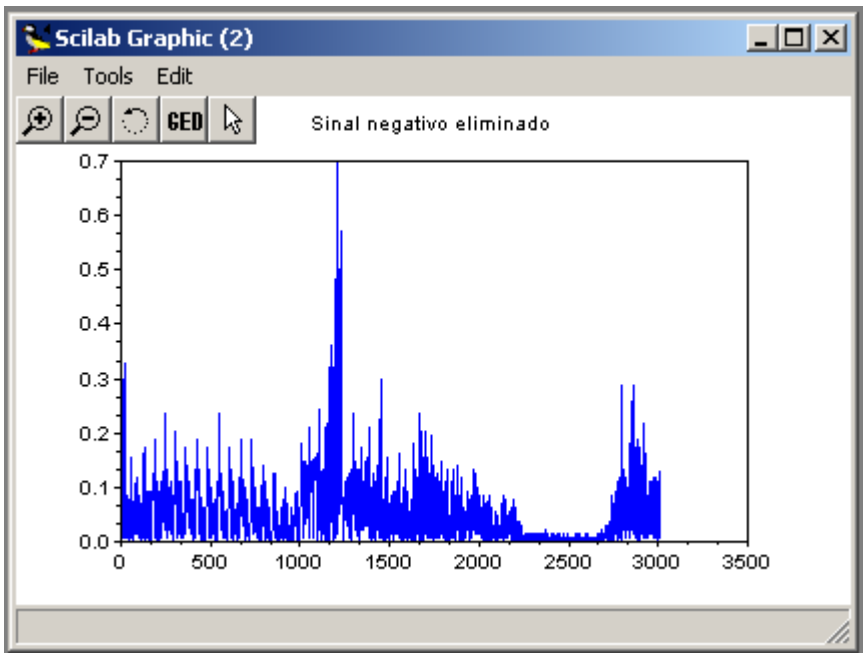


Figura 3.3 – Sinal negativo eliminado

A fase de redução do sinal amostrado se torna necessária para reduzir a quantidade

de valores gerados por uma leitura de uma palavra falada, pois a quantidade de valores capturados é extremamente elevada. Por exemplo, a uma taxa de 8000 Hz com níveis de quantização representados por 8 bits serão gerados em um segundo 8000 valores compreendidos entre 0 e 255. Levando-se em conta que precisa-se apenas da forma de onda para identificar o sinal (NUSSENZVEIG, 1997), a quantidade de 8000 valores se torna abundante. Considerando-se que a rede neural desenvolvida permite o uso de apenas 100 entradas, a redução deverá diminuir o número de sinais amostrados para 100 valores. Sendo que, essa diminuição deve ser realizada detectando-se a forma de onda, ou seja, considerando-se as amplitudes significativas, ou ainda, considerando-se os maiores valores de amplitude para compor o sinal amostrado, vide Figura 3.4. Quanto ao cálculo empregado para a redução, deve-se seguir a seguinte fórmula:

$$\text{(3.1)} \quad \text{taxa de redução} = \frac{\text{quantidade de sinais amostrados}}{100}$$

Por essa fórmula, se o sinal amostrado possui 6000 valores, por exemplo, 1 em cada 60 deverá ser escolhido para compor o sinal processado, sendo que o valor escolhido deverá ser o mais significativo.

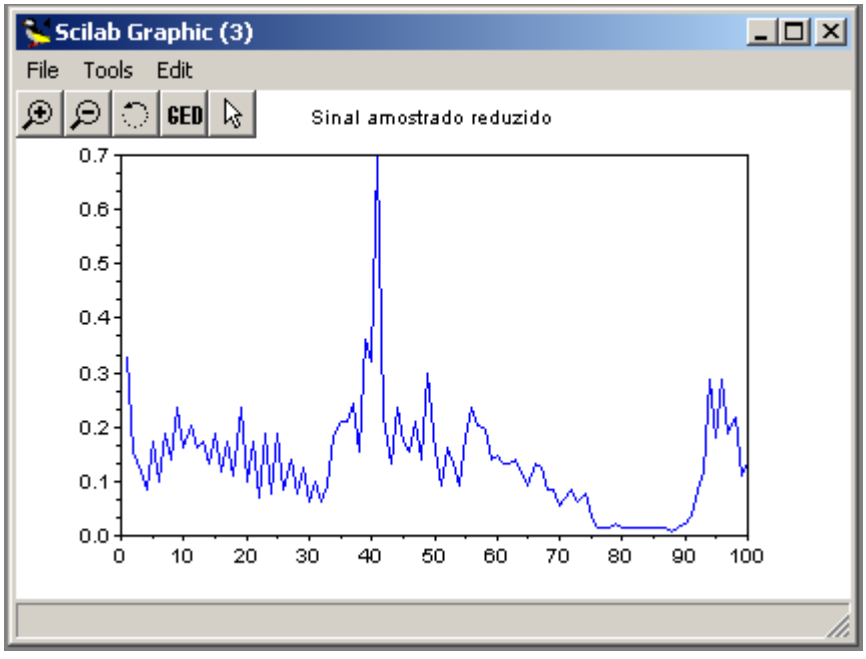


Figura 3.4 – Sinal amostrado reduzido

A mediação do sinal reduzido consiste em uma nova redução do sinal, porém, sem perda de nenhuma posição, ou seja, as 100 posições originais são mantidas. Segundo

Tafner (1996) esse método basicamente iguala três sinais consecutivos pelo valor mais alto compreendido entre os três. Sendo que esta operação é realizada ao longo de todas as 100 posições (Figura 3.5). Como após a mediação não há redução de posições o sinal amostrado ficará mais intenso quando for apresentado à rede neural, e isto garantirá que apesar do sinal ser reduzido mais uma vez a sua identidade manter-se-á de forma bastante significativa.

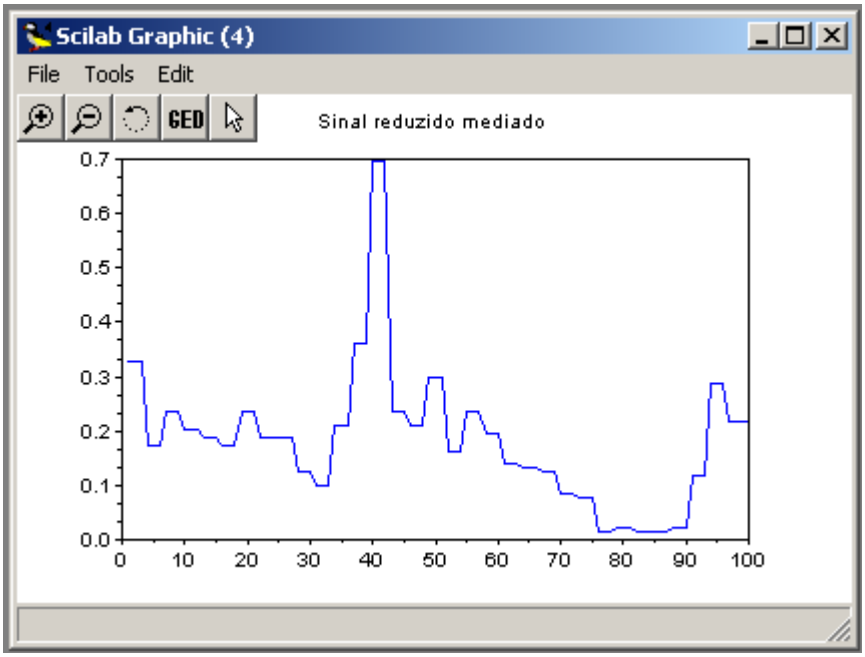


Figura 3.5 – Sinal reduzido mediado

Finalmente, a etapa de normalização do sinal mediado procura anular a diferença da intensidade do sinal como um todo, ou seja, falar duas vezes a mesma palavra, mas com alturas diferentes, produzirá, conseqüentemente, dois sinais com intensidades diferentes, apesar da mesma forma de onda. Esse desalinhamento de sinais pode ser causado por diversos fatores, sendo que o mais comum e fácil de ocorrer é a distância da boca em relação ao microfone (CAMPBELL, 1998). Quanto mais perto a boca estiver do microfone maior será a intensidade do sinal, e quanto mais longe menor será a intensidade.

Uma das maneiras para suprimir essa deficiência é utilizar um método matemático conhecido como normalização, que consiste em ajustar os valores de uma série qualquer em relação a um valor específico. No caso deste trabalho, consiste em igualar todos os valores do sinal mediado à intensidade de 1, eliminando, assim, diferenças de intensidade entre os outros sinais. O valor x_i do sinal mediado x deve ser multiplicado pelo valor 1 e,

depois, dividido pelo valor mais alto encontrado em todo o sinal x, ou seja, necessita apenas ser dividido pelo valor mais alto de x:

(3.2) valor normalizado = valor mediado / maior valor do sinal

Dessa forma, o valor de mais alta intensidade do sinal será igualado a 1, e os demais valores igualados de forma proporcional ao valor de 1. Como pode ser observado pela Figura 3.6, optou-se por normalizar todos valores do sinal mediado à intensidade de 1, pois o Toolbox ANN do Scilab trabalha como padrão com funções sigmoidais, ou seja, funções compreendidas no intervalo de 0 a 1.

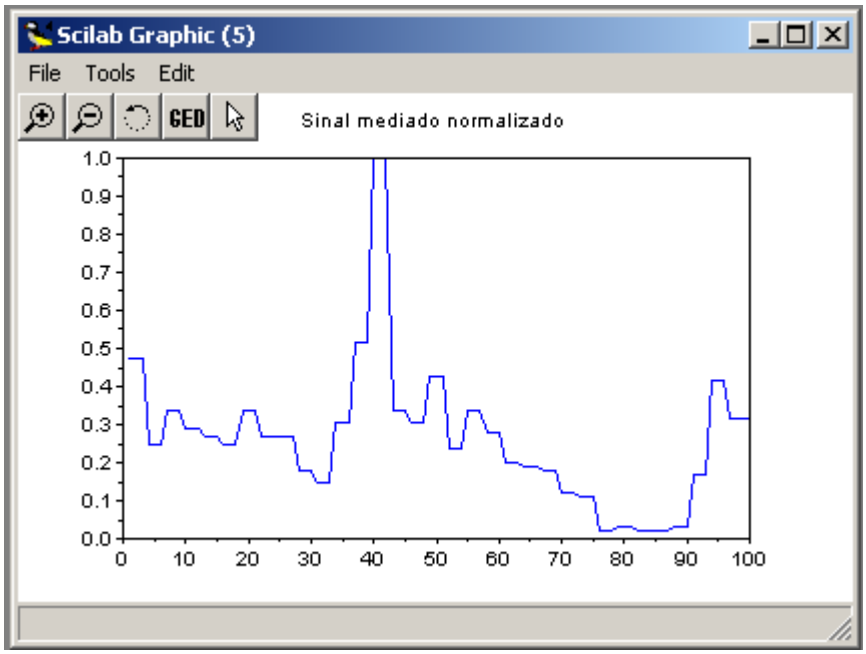


Figura 3.6 – Sinal mediado normalizado

3.3.2. Comparação entre sinais

Nesta seção será ilustrado um conjunto de gráficos que demonstra a diferença na forma de onda, ao longo do tratamento definido anteriormente, de dois sinais sonoros semelhantes (mesma palavra) sendo pronunciados por duas pessoas diferentes. Neste caso, a palavra utilizada foi novamente a palavra “teste”.

A Figura 3.7 ilustra a forma de onda original da palavra “teste” sendo pronunciada pelo um indivívuo α (Figura 3.7 (a)) e pelo indivíduo β (Figura 3.7 (b)).

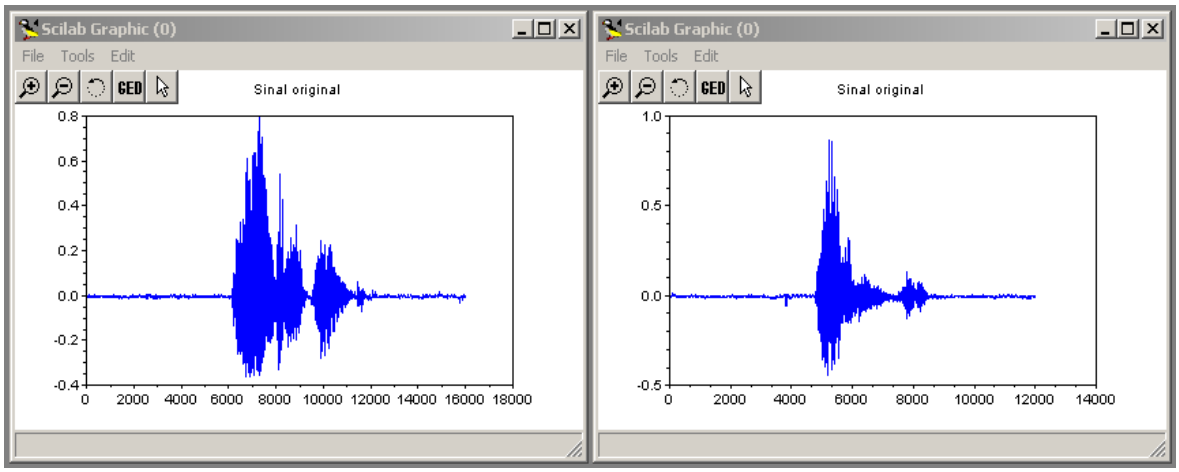


Figura 3.7 – Sinais amostrados originais

(a) Sinal amostrado original do indivíduo α (a esquerda)

(b) Sinal amostrado original do indivíduo β (a direita)

Pode-se observar pela Figura 3.8 abaixo os mesmos sinais sonoros sem as suas respectivas zonas de silêncio (sinais entre -0,1 e 0,1).

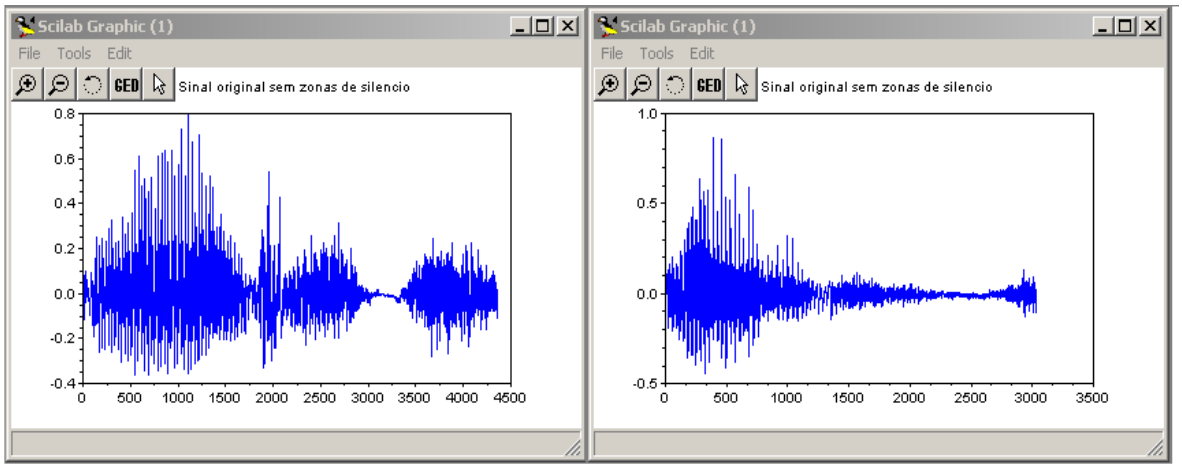


Figura 3.8 – Sinais originais sem zonas de silêncio

(a) Sinal original sem zonas de silêncio do indivíduo α (a esquerda)

(b) Sinal original sem zonas de silêncio do indivíduo β (a direita)

Na Figura 3.9 tem-se as amostras representadas na suas formas sem sinais negativos.

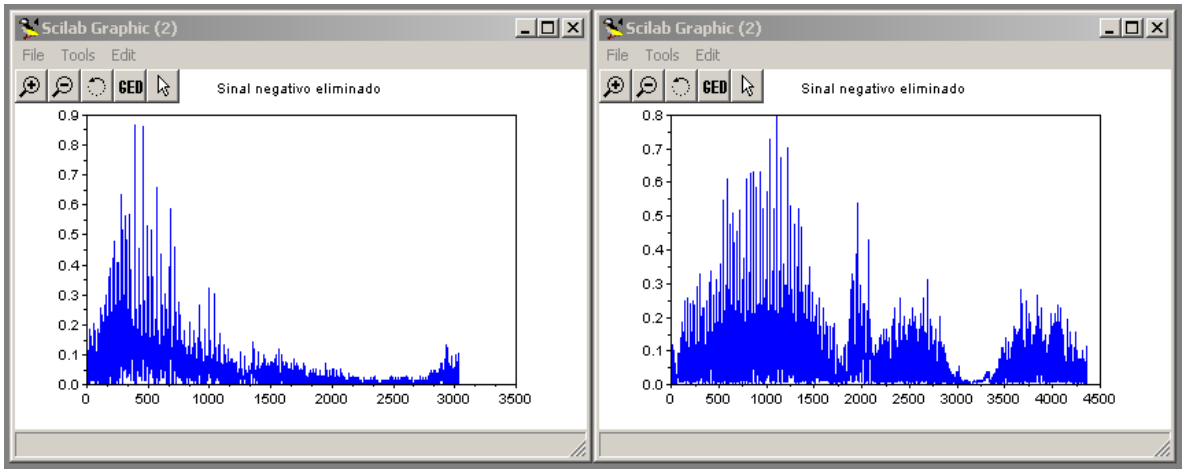


Figura 3.9 – Sinais negativos eliminados

(a) Sinal negativo eliminado do indivíduo α (a esquerda)

(b) Sinal negativo eliminado do indivíduo β (a direita)

A Figura 3.10 ilustra a fase de redução do sinal amostrado, a partir desta etapa os sinais tratados serão plotados em um mesmo gráfico para facilitar a comparação entre os mesmos. Em azul tem-se o indivíduo α, e em vermelho o indivíduo β.

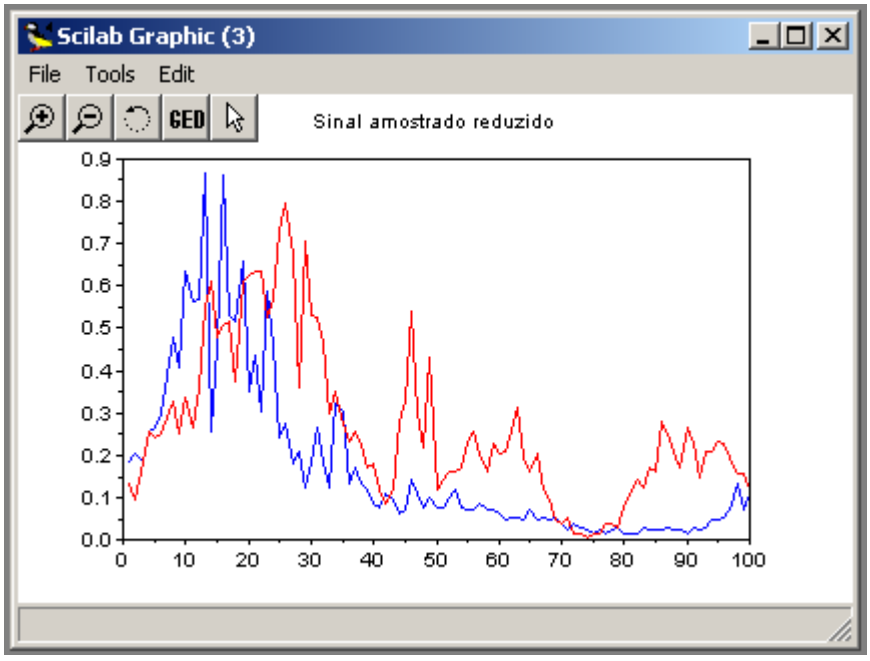


Figura 3.10 – Sinais amostrados reduzidos

As formas de onda dos sinais durante a etapa de mediação das amostras podem ser comparadas pela Figura 3.11. Novamente, em azul tem-se o indivíduo α, e em vermelho o indivíduo β.

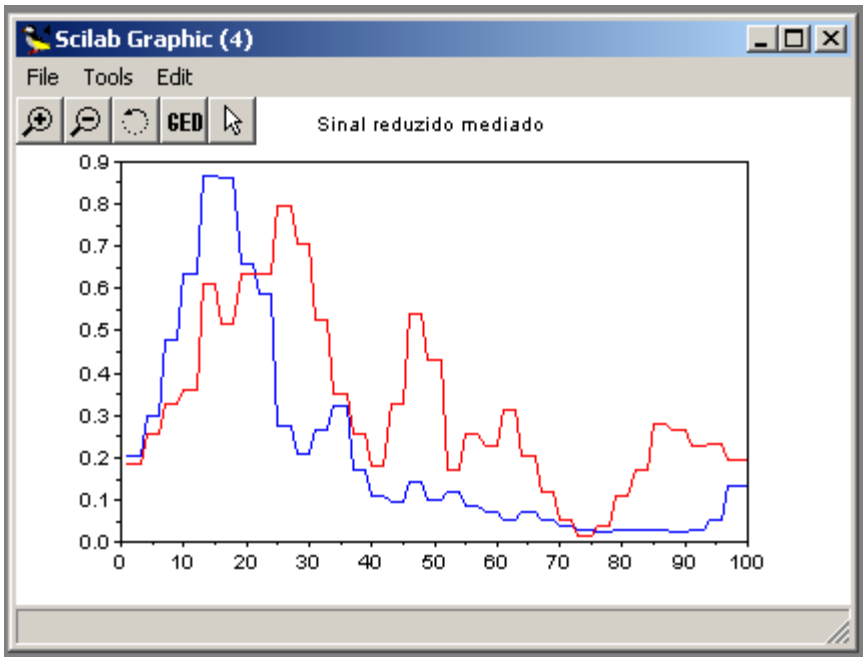


Figura 3.11 – Sinais reduzidos mediados

Finalmente, na Figura 3.12 temos a comparação dos sinais na última fase do tratamento: a normalização do sinal mediado. Em azul tem-se o indivíduo α, e em vermelho o indivíduo β.

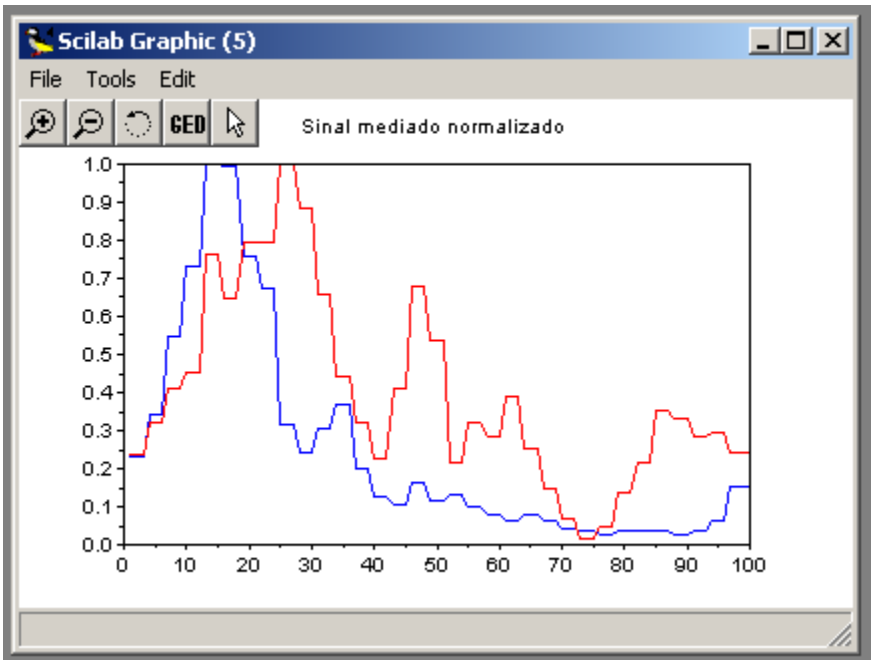


Figura 3.12 – Sinais mediados normalizados

Através destes gráficos se torna evidente as diferenças nas formas de onda de dois sinais sonoros semelhantes (mesma palavra) sendo pronunciados por dois indivíduos

diferentes. Além disso, estes gráficos demonstram, principalmente, que estas diferenças se mantêm ao longo do tratamento definido neste trabalho.

No próximo capítulo é apresentada a rede neural obtida para reconhecer o padrão de voz. O treinamento é realizado utilizando dados reais de fala, conforme exemplificado pela Figura 3.12. Os resultados do desempenho da rede também são apresentados, inclusive com casos de teste em que se utilizou mais de uma rede neural.

4. RESULTADOS E DISCUSSÃO

Neste capítulo será abordado os resultados obtidos com a execução do presente trabalho. Como mencionado anteriormente, para a implementação da rede neural foi utilizado a ferramenta Scilab versão 4.1.1 (Toolbox ANN). De acordo com a definição de Furui (1997), citada no Capítulo 2, este trabalho pode ser classificado como um sistema de verificação de identidade através da fala, ou seja, é um sistema responsável pelo processo de aceitar ou rejeitar a reivindicação de identidade de uma determinada pessoa.

Para compor a base de dados de vozes foram capturadas as pronúncias da palavra "teste" de 16 pessoas diferentes, sendo que 7 (sete) eram mulheres e 9 (nove) eram homens. Logo, o método de verificação desenvolvido é dito como dependente do texto, pois requer que o indivíduo pronuncie uma a palavra chave "teste" tanto para o treinamento quanto para os testes da rede neural.

As amostras capturadas foram divididas em dois grupos: amostras do tipo A e amostras do tipo B. As amostras do tipo A são aquelas que representam os indivíduos que se intenciona aceitar no processo de verificação, enquanto que as amostras do tipo B são aquelas que se planeja rejeitar no processo de verificação. No grupo das amostras do tipo A foram coletadas as pronúncias de um único indivíduo, e no grupo das amostras do tipo B foram capturadas as pronúncias das demais 15 (quinze) pessoas.

Ao todo a base de dados é composta por 132 amostras, sendo 66 pertencentes ao tipo A e 66 pertencentes ao tipo B. Portanto, cada indivíduo pertencente ao grupo de amostras do tipo B pronunciou a palavra "teste" aproximadamente 4,4 vezes, e o único indivíduo que compõem o grupo de amostras tipo A pronunciou a mesma palavra 66 vezes.

A partir das 132 amostras definiu-se que aproximadamente 60% destas (80 amostras, sendo 40 do tipo A e 40 do tipo B) formariam o conjunto de treinamento da rede neural; 20% (26 amostras, sendo 13 do tipo A e 13 do tipo B) integrariam o conjunto de validação; e os restantes 20% (26 amostras, sendo 13 do tipo A e 13 do tipo B) seriam destinados a compor o conjunto de amostras de teste do sistema. Veja a Tabela 4.1 abaixo para obter um melhor detalhamento da divisão das amostas.

Tabela 4.1 – Divisão das amostras (Abordagem 1)

Grupo	Dados Originais	Treinamento	Validação	Teste
Amostras do tipo A	66	40	13	13
Amostras do tipo B	66	40	13	13
Total	132	80	26	26

Antes de se iniciar a discussão sobre a implementação proposta, e os resultados obtidos com a mesma, torna-se necessário definir dois importantes conceitos referentes aos tipos de erros que o sistema pode apresentar: os erros denominados “falsos negativos” e os erros nomeados como “falsos positivos”.

Falsos negativos são os erros que acontecem quando o sistema verifica negativamente um indivíduo, ou em outras palavras, quando o sistema rejeita a verificação de identidade de um indivíduo que teoricamente deveria aceitar.

Por outro lado, falsos positivos são os erros causados quando o sistema verifica positivamente um determinado indivíduo, ou seja, o sistema aceita a verificação de identidade de um indivíduo que teoricamente deveria rejeitar. É importante salientar que, em caso de erro, falsos negativos são mais desejáveis do que falsos positivos, visto que é mais adequado rejeitar uma identidade verdadeira do que aceitar uma identidade falsa.

Para o treinamento da rede neural foi utilizado o algoritmo *backpropagation* com *momentum*. Inicialmente a rede desenvolvida apresentava a seguinte arquitetura de neurônios: [100,100,250,35,1]. A camada de entrada possui esta configuração pois durante o tratamento do sinal sonoro, na etapa de redução do sinal amostrado, definiu-se que o sinal seria reduzido para 100 posições. A camada de saída possui apenas um neurônio devido ao fato de que a saída do sistema possui apenas dois estados possíveis: verdadeiro ou falso, ou seja, o sistema apenas aceita ou rejeita a verificação de identidade de determinado indivíduo.

A taxa de aprendizagem foi configurada com o valor de 0.001, enquanto que a taxa de *momentum* foi ajustada para o valor numérico 0.9. A rede neural foi treinada durante 4000 épocas, alcançando uma taxa de acerto máxima de 69,23% para o conjunto de 26 amostras de teste. Considerando-se, separadamente, apenas as 13 amostras do grupo A, cabíveis de causarem falsos negativos, a melhor taxa de acerto alcançada foi de 76,92%. Enquanto que a taxa de acerto das amostras do grupo B, cabíveis de provocarem falsos positivos, atingiu o valor de 61,54%.

Como descrito anteriormente, em caso de erro do sistema, falsos negativos são mais

adequados do que falsos positivos. Sendo assim, com o objetivo de reduzir a quantidade de falsos positivos, e conseqüentemente aumentar a taxa de acerto dos testes referentes às amostras do tipo B, determinou-se que seria realizado uma diminuição do número de amostras do tipo A durante o treinamento da rede neural, já que devido ao tamanho limitado da base de vozes não seria possível aumentar a quantidade de amostras do tipo B.

Com a redução do número de amostras do tipo A no treinamento do sistema, a rede se torna mais especializada em reconhecer os indivíduos do grupo B, visto que há um desbalanceamento do número de amostras. Por essa razão, a taxa de acerto das amostras do tipo B cresce naturalmente. A nova divisão das amostras pode ser visualizada pela Tabela 4.2.

Tabela 4.2 – Divisão das amostras (Abordagem 2)

Grupo	Dados Originais	Treinamento	Validação	Teste
Amostras do tipo A	66	25	13	13
Amostras do tipo B	66	40	13	13
Total	132	65	26	26

Para validar esta constatação, realizou-se uma bateria de testes com a finalidade de se comparar os resultados obtidos com esta nova abordagem. A partir dos testes realizados, capturou-se os 4 (quatro) melhores resultados alcançados (veja Tabela 4.3) pelas duas abordagens (com a divisão original das amostas e com a nova divisão proposta). Os conjuntos de validação e teste sempre foram os mesmos para todos os testes. Importante notar que a diferença fundamental nos testes realizados é o processo de treinamento individual de cada rede, visto que todas possuem a mesma configuração e o mesmo número de épocas.

Tabela 4.3 – Comparação da taxa de acerto com a redução do conjunto de treinamento

Conj. Treinamento Reduzido			Conj. Treinamento Completo		
Total	Parcial 1 - FN	Parcial 2 - FP	Total	Parcial 1 - FN	Parcial 2- FP
65.38	53.84	76.92	65.38	69.23	61.53
80.76	76.92	84.61	69.23	69.23	69.23
69.23	61.53	76.92	65.38	61.53	69.23
76.92	76.92	76.92	65.38	76.92	53.84

De acordo com a Tabela 4.3 nota-se uma razoável diferença entre os resultados obtidos através do conjunto de treinamento completo em relação aos resultados alcançados com o conjunto de treinamento reduzido. Utilizando-se a nova divisão das amostras a taxa máxima de acerto alcançada foi de 80,76% para o conjunto total de amostras de teste. Analisando-se, separadamente, apenas as 13 amostras do grupo A, cabíveis de causarem falsos negativos (FN), a melhor taxa de acerto, visualizada na segunda coluna da tabela, é de 76,92%. Enquanto que a taxa de acerto das amostras do grupo B, cabíveis de provocarem falsos positivos (FP), atingiu o valor máximo de 84.61%.

Obtendo-se, de modo geral, a média aritmética dos quatro resultados atingidos, levando-se em consideração as duas abordagens, chega-se aos seguintes resultados: para os testes realizados com o conjunto de treinamento completo (Abordagem 1) a taxa média de acerto total do sistema foi de 66,34%, a primeira taxa média de acerto parcial (referente as amostras do tipo A, ou seja, aquelas que podem causar falsos negativos) foi de 69,22%, e a segunda taxa média de acerto parcial (referente as amostras do tipo B, ou seja, aquelas que podem causar falsos positivos) foi de 63,45%. Para os testes realizados com o conjunto de treinamento reduzido (Abordagem 2) a taxa média de acerto total do sistema foi de 73,07%, a primeira taxa média de acerto parcial foi de 67,30%, e a segunda taxa média de acerto parcial foi de 78,84%.

Note que com esta nova abordagem houve uma melhora significativa no desempenho da rede neural sob a ótica da taxa média de acerto das amostras de teste do grupo B (aumentou de 63,45% para 78,84%), que conseqüentemente, ocasionou a diminuição dos falsos positivos. Embora, como seria previsto, também tenha ocasionado uma diminuição na taxa média de acerto das amostras de teste do grupo A (reduzindo de 69,22% para 67,30%), provocando o aumento do número de falsos negativos.

A partir da nova divisão das amostras resolveu-se explorar o desempenho da rede através de épocas diferentes de treinamento. Como pode ser observado pela Tabela 4.4 foram realizados testes com 1000, 3000, 4000, 5000, 6000 e 12000 épocas. Todos os testes

efetuados foram executados com a mesma configuração de rede e com os mesmos conjuntos de treinamento, validação e teste.

Tabela 4.4 – Taxa de acerto por época

Épocas	Total	Parcial 1 - Falsos Negativos	Parcial 2 - Falsos Positivos
1000	69.23	69.23	69.23
3000	73.07	76.92	69.23
4000	84.61	84.61	84.61
5000	80.76	76.92	84.61
6000	80.76	76.92	84.61
12000	73.07	76.92	69.23

Nota-se que o melhor resultado alcançado, durante o treinamento de 4000 épocas, foi de 84,61% para a taxa de acerto total. Considerando-se apenas as amostras cabíveis de causarem falsos negativos obteve-se melhores resultados a partir de 3000 épocas, sendo que o valor máximo da taxa de acerto alcançado para estas amostras também foi de 84,61% no treinamento de 4000 épocas. Agora, tomando-se como teste as amostras cabíveis de provocarem falsos positivos, os melhores resultados foram alcançados com 4000, 5000 e 6000 épocas, atingindo mais uma vez uma taxa de acerto máxima de 84,61%.

Deve-se destacar que com exceção dos treinamentos realizados com 3000 e 12000 épocas, em todos testes efetuados a taxa de acerto das amostras admissíveis de causarem falsos positivos foi maior ou igual a taxa de acerto das amostras admissíveis de provocarem falsos negativos, ou em outras palavras, a quantidade de falsos positivos foi inferior à quantidade de falsos negativos.

Para aumentar o desempenho e a confiabilidade do sistema proposto uma nova implementação foi desenvolvida. Ao invés de se utilizar apenas uma única rede para a verificação de um determinado indivíduo, foi elaborado uma solução que agrupava os resultados de 3 (três) redes distintas em um único resultado. Como elucidado nos capítulos anteriores o Scilab trabalha, como padrão, com funções sigmoidais (funções compreendidas entre 0 e 1). Sendo assim, o resultado de um determinado caso de teste aplicado à rede se resume a um valor numérico contínuo compreendido no intervalo de 0 a 1. Para se obter a resposta final da rede deve-se converter (arredondar para o inteiro mais próximo) este valor contínuo para os valores discretos 0 ou 1 (ou seja, rejeitar ou aceitar).

Para agrupar as soluções propostas pelas 3 redes em uma única resposta foi realizada a média aritmética das 3 implementações e apenas posteriormente foi efetuado o

arredondamento desta média para a resposta final do sistema. Todas as três redes neurais foram configuradas com taxa de *momentum* igual a 0.9 e taxa de aprendizagem igual a 0.001, além disso todas foram treinadas com os mesmos conjuntos de treinamento e teste.

Após os testes realizados o melhor desempenho obtido foi alcançado após um treinamento de 6000 épocas. Atingiu-se uma taxa de acerto máxima de 80,76% para o conjunto de 26 amostras de teste. Considerando-se, novamente, apenas as 13 amostras do grupo A, cabíveis de causarem falsos negativos, a melhor taxa de acerto alcançada foi de 76,92%. Enquanto que a taxa de acerto das amostras do grupo B, cabíveis de provocarem falsos positivos, atingiu o valor de 84,61%.

O gráfico do erro quadrático obtido durante o treinamento das três redes pode ser visualizado pela Figura 4.1. Pelo gráfico nota-se que a partir de 2000 épocas o erro quadrático de treinamento se manteve constante para todas as três redes neurais.

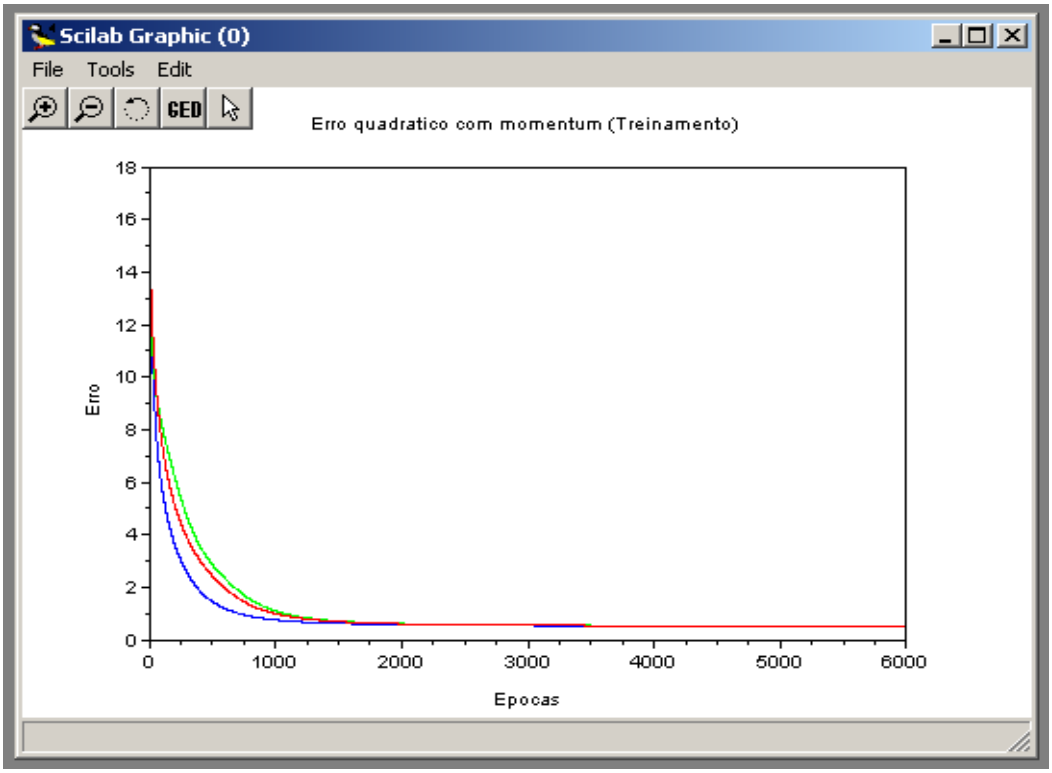


Figura 4.1 – Erro quadrático de treinamento das três redes

Contudo, analisando-se o gráfico do erro quadrático de validação das três redes (Figura 4.2) nota-se que o gráfico é continuamente crescente no intervalo de 6000 épocas. Importante salientar que em todos os testes executados com esta arquitetura todas as curvas de erro quadrático de validação se apresentaram de forma crescente logo a partir das primeiras épocas de treinamento.

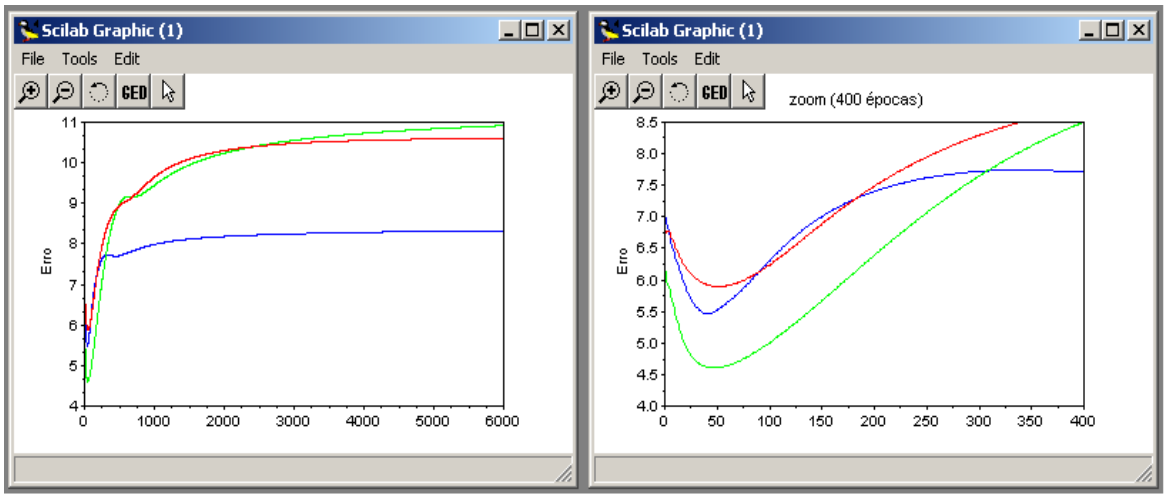


Figura 4.2 – Erro quadrático de validação das três redes

(a) Curva completa

(b) Curva parcial durante as 400 primeiras épocas – Zoom (400 épocas)

Tem-se também que, mesmo os testes executados com uma quantidade inferior de épocas, o que conseqüentemente provocaria um erro de validação menor, apresentaram resultados inferiores aos do exemplo acima. Porém, nos testes executados com um número superior a 6000 épocas o desempenho da rede foi drasticamente reduzido, inviabilizando a utilização do sistema.

Para validar a declaração de que os testes executados em redes com um período de treinamento menor, e simultaneamente com um erro de validação também menor, apresentaram resultados inferiores aos testes realizados em redes treinadas por 4000 épocas, foi desenvolvido um sistema composto por um comitê[1] de nove redes neurais. Para testar o sistema foram selecionadas 26 amostras da base de vozes, sendo 13 amostras do tipo A e 13 do tipo B. Cada rede do comitê foi treinada com um conjunto de 80 amostras, selecionadas aleatoriamente do restante das amostras na base de dados. Como critério de parada para o processo de treinamento, utilizou-se um conjunto, também aleatório, de 26 amostras de validação. Dessa maneira, o treinamento é interrompido assim que a curva do erro quadrático de validação começa a se tornar crescente. Foi definido a utilização das mesmas taxas de aprendizagem e *momemtum*, além da mesma arquitetura de neurônios, dos outros sistemas citados anteriormente neste trabalho. A Figura 4.3 apresenta o gráfico do erro quadrático de treinamento deste sistema, onde a curva do erro quadrático de cada

1 Um comitê de redes neurais é um arranjo de redes independentes, trabalhando em paralelo, mas com a finalidade de classificação única e consensual. De modo geral, a resposta final é o resultado de algum tipo de procedimento de votação envolvendo os classificadores membros do sistema. (SANTOS ET AL., 2007)

rede é representada por uma cor diferente.

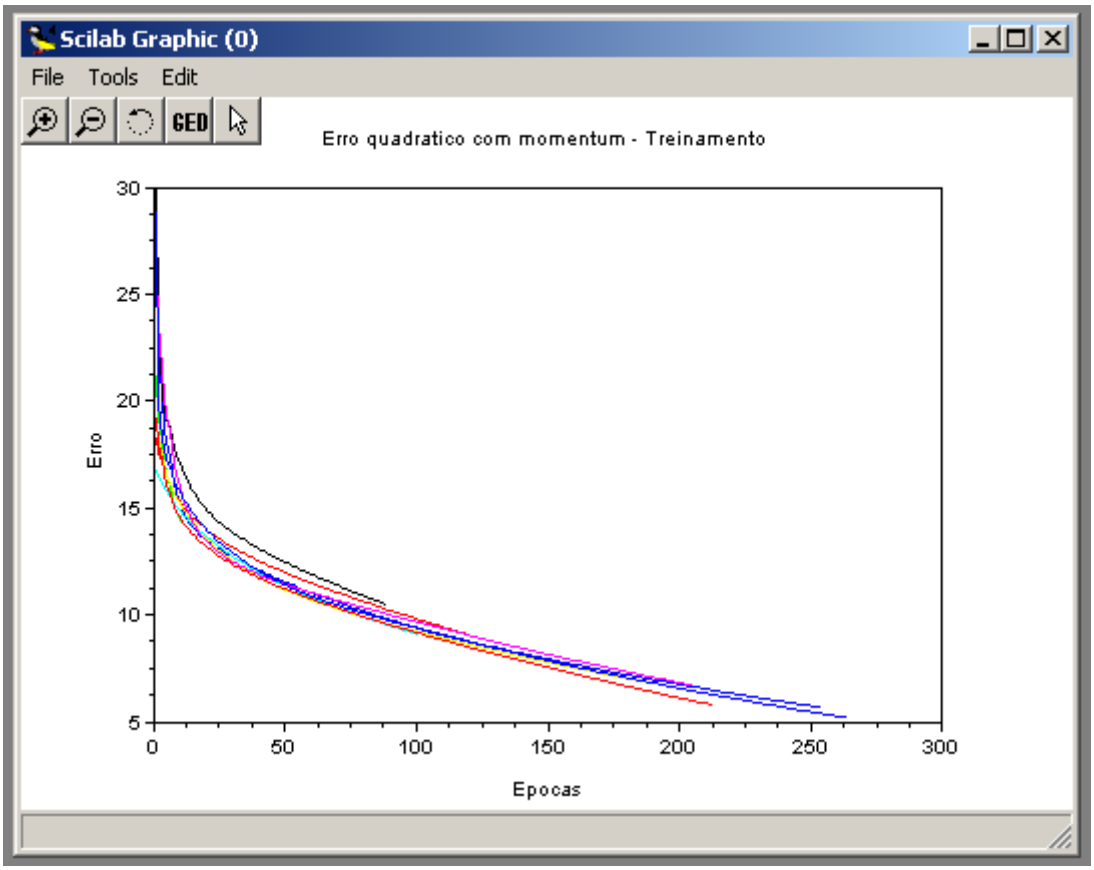


Figura 4.3 – Erro quadrático de treinamento do comitê

Como pode ser observado pelas Figuras 4.3 e 4.4, a rede que conseguiu ser treinada por mais tempo atingiu o valor aproximado de 250 épocas. Nos testes realizados, este sistema alcançou uma taxa de acerto máxima de 61,53% para o conjunto de 26 amostras de teste. Considerando-se, separadamente, apenas as 13 amostras do grupo A, cabíveis de causarem falsos negativos, a melhor taxa de acerto alcançada foi de 69,23%. Enquanto que a taxa de acerto das amostras do grupo B, cabíveis de provocarem falsos positivos, atingiu o valor de 53,84%.

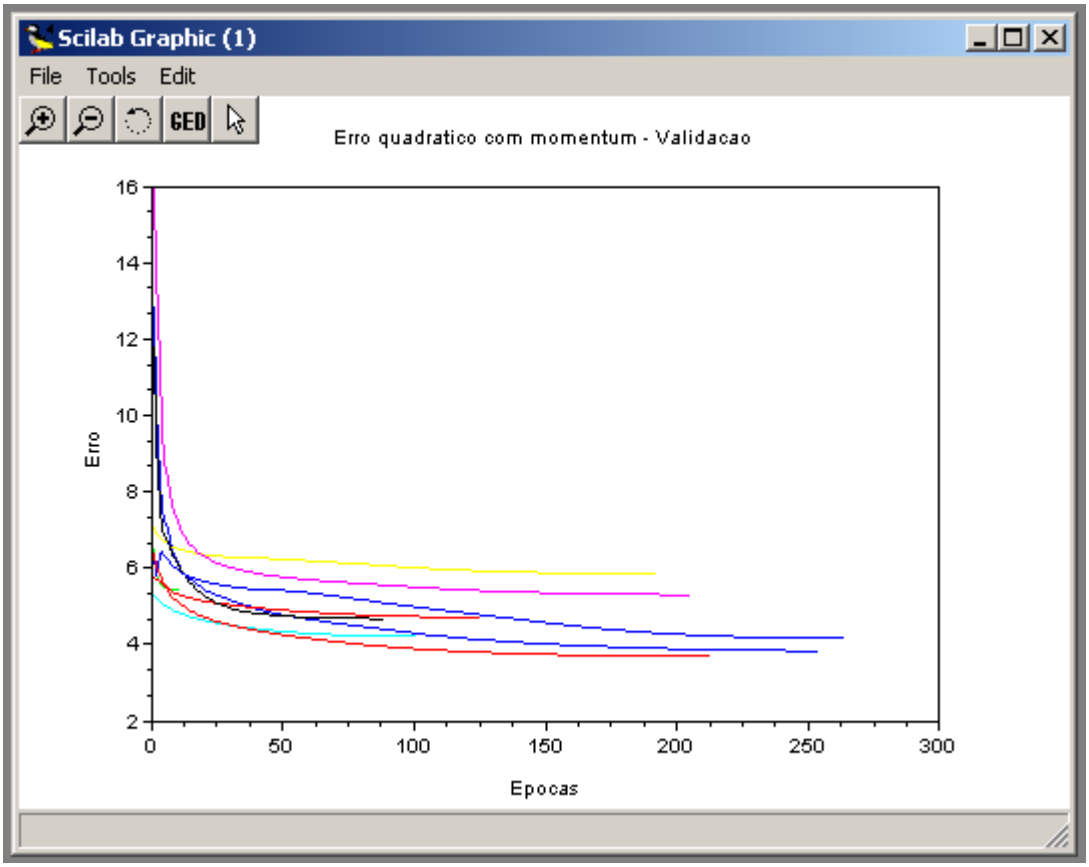


Figura 4.4 – Erro quadrático de validação do comitê

Porém com um treinamento de 4000 épocas, tendo-se a mesma configuração do sistema anterior, os resultados obtidos são: taxa de acerto máxima de 65,38% para o conjunto de 26 amostras de teste. Considerando-se as 13 amostras do grupo A a melhor taxa de acerto alcançada foi de 69,23%. Enquanto que a taxa de acerto das amostras do grupo B atingiu o valor de 61,53%. A Figura 4.5 ilustra o erro quadrático de treinamento durante as 4000 épocas.

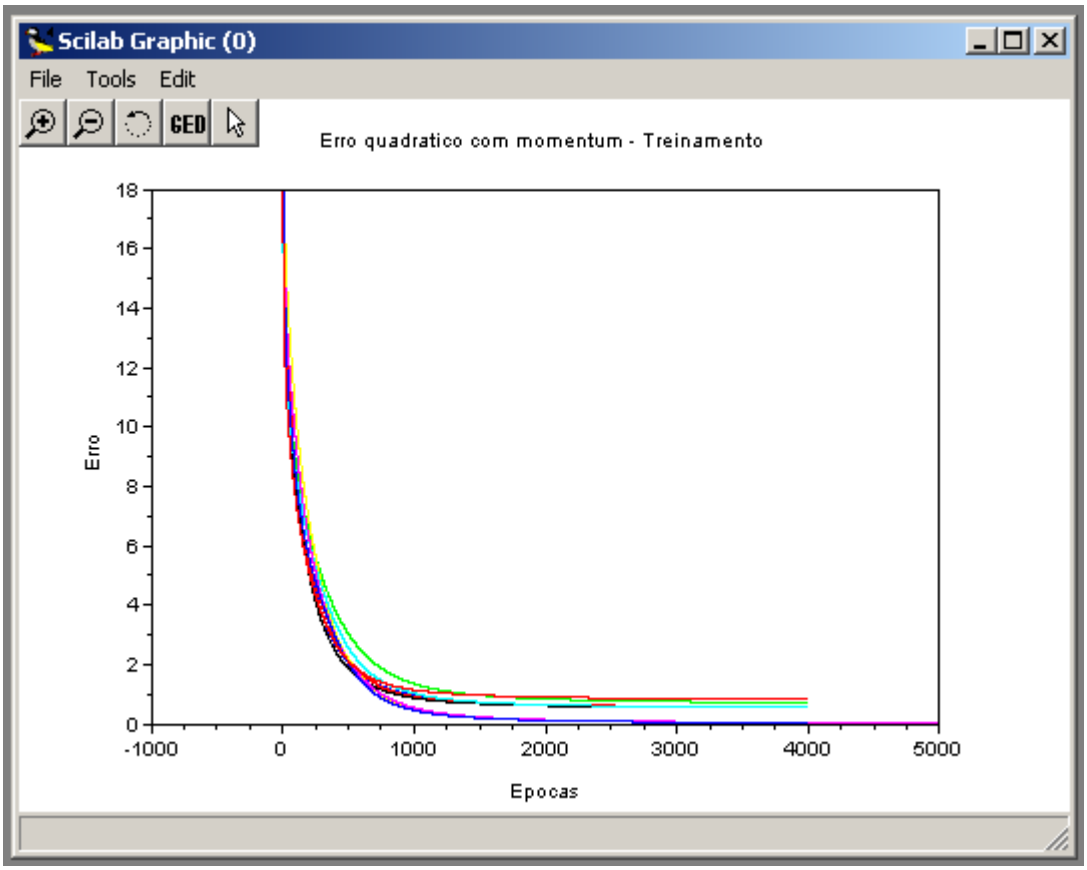


Figura 4.5 – Erro quadrático de treinamento do comitê (4000 épocas)

Na Figura 4.6 tem-se o gráfico do erro quadrático de validação durante o treinamento de 4000 épocas.

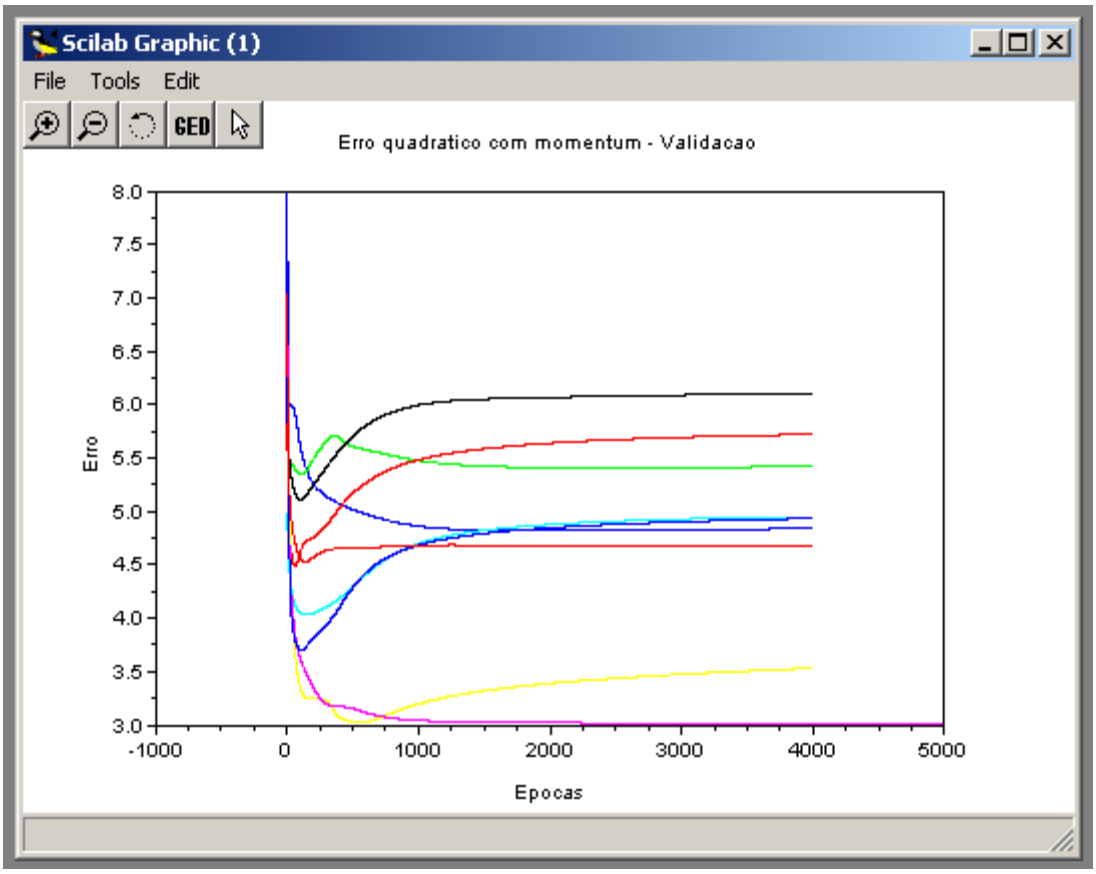


Figura 4.6 – Erro quadrático de validação do comitê (4000 épocas)

Comparando-se os resultados encontrados até aqui com os de outros autores, como por exemplo Wang et al. (2002), nota-se que as taxas de acerto alcançadas neste trabalho demonstram serem resultados expressivos. Os referidos autores realizaram uma comparação de uma MLP (Multilayer Perceptron) com outras técnicas híbridas na identificação de um indivíduo através de um segmento de voz de 8 (oito) segundos. A melhor taxa de identificação alcançada pelos autores utilizando uma MLP foi de 63,81%, enquanto que os resultados obtidos com técnicas híbridas variam de 65,50% a 92,37%.

Logo, os resultados obtidos neste trabalho são satisfatórios considerando-se apenas as redes MLP. Comparando-se com técnicas híbridas os resultados alcançados demonstraram ser um pouco inferiores, porém este era um resultado esperado, visto que técnicas híbridas geralmente apresentam um desempenho melhor se comparado às demais metodologias.

Chen et al. (1996) também realizou trabalhos para a identificação de indivíduos através da fala, comparando redes MLP e redes neurais HME (Hierarchical Mixtures of

Experts[2]). Através de redes MLP, estes autores alcançaram taxas de acerto que variam entre 84,80% a 96,33%, enquanto que com redes HME os autores atingiram resultados que variam de 88,4% a 98,78%.

Uma das causas que podem ser visualizada para esta diferença significativa de resultados, levando-se em consideração as taxas de acerto obtidas neste trabalho, é a considerável desigualdade no número de amostras utilizadas pelos sistemas. Como o desempenho de uma rede é afetado diretamente pelo número de amostras durante o treinamento, os resultados obtidos neste trabalho ainda demonstram ser satisfatórios considerando-se a limitada base de vozes utilizada.

2 HME são redes neurais que possuem múltiplas sub-redes que cooperam entre si, baseadas no princípio de dividir para conquistar, para lidar com um dado problema. (CHEN ET AL., 1996)

5. CONCLUSÃO

O reconhecimento automático de indivíduos através da fala se resume a utilização de computadores para reconhecer uma pessoa através de uma palavra, ou frase, pronunciada. Os sistemas de reconhecimento de pessoas través da voz podem ser usados para identificar uma pessoa em particular ou para verificar a reivindicação de identidade de determinada pessoa. Sendo que neste trabalho foi estudado a aplicação de redes neurais MLP no processo de verificação de identidade.

5.1. Considerações Finais

Sistemas de reconhecimento de padrões, especialmente aqueles na área biométrica, possuem um vasto horizonte de aplicações, por exemplo para a identificação criminal e para o controle de acesso. E dentre estas aplicações a verificação de identidade de indivíduos através da voz, discutida nos capítulos anteriores, é uma metodologia que oferece bons resultados.

A rede neural desenvolvida neste trabalho apresentou um bom desempenho. Porém, para a sua utilização em uma aplicação do mundo real deve-se levar em consideração alguns critérios importantes, como por exemplo, a segurança e a confiabilidade. Visto que nenhuma das implementações realizadas atingiram uma taxa de erros nula, deve-se considerar o nível de confiança desejado para um sistema de verificação, antes de se resolver pela utilização de redes neurais como metodologia de reconhecimento.

Duas importantes características neste trabalho podem ser mencionadas: primeiro, o baixo volume de amostras que se dispunha na base de vozes, tanto para realizar os testes do sistema quanto, e principalmente, para efetuar o treinamento adequado da rede neural. Uma vez que o desempenho do sistema proposto depende essencialmente das amostras com que o treinamento da rede neural é realizado. Segundo, a análise mais detalhada da taxa de acerto do sistema, considerando-se não apenas a taxa de acerto global, mas também a taxa de acerto individual das amostras cabíveis de causarem falsos negativos e a taxa de acerto individual das amostras admissíveis de provocarem falsos positivos.

5.2. Trabalhos Futuros

Finalmente, como trabalhos futuros, novas pesquisas podem ser realizadas a partir desta, como por exemplo a verificação e solidificação das taxas de acerto através de novos testes, preferencialmente com uma base de dados maior e até mesmo com novas palavras ou frases.

Além disso, a partir dessa base, pode-se desenvolver novos estudos considerando-se as demais vertentes dos sistemas de reconhecimento de indivíduos através da fala, como a verificação não dependente do texto e os sistemas de identificação de indivíduos, dependentes ou não do texto.

Por fim, pode-se efetuar pesquisas com o objetivo de se construir aplicações reais. Construindo sistemas que sejam capazes de reconhecer tanto as palavras, ou comandos, quanto as pessoas que os pronunciaram.

REFERÊNCIAS BIBLIOGRÁFICAS

ARBIB, M. A. **The handbook of brain theory and neural networks**. A Bradford book: Second Edition, 2003. 1309 páginas.

BARRETO, J. M. **Introdução às Redes Neurais Artificiais**. Florianópolis/SC : Laboratório de Conexionismo e Ciências Cognitivas-UFSC, 2002. 57 páginas.

BEN-YACOUB, S.; ABDELJAOUED, Y.; MAYORAZ, E. Fusion of face and speech data for person identity verification. **IEEE Transactions on Neural Networks**, vol. 10, número 5, 1999.

BISHOP, C. M. **Neural Networks for Pattern Recognition**. USA/New York: Oxford, 1995.

BRUNELLI, R.; FALAVIGNA, D. Person identification using multiple cues. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, vol. 17, p. 955-966. Italy: Istituto per la Ricerca Scientifica e Tecnologica,1995.

CAMPBELL, J. P. **Speaker Recognition**. Fort Meade: Department of Defense, 1998.

CHEN, K.; XIE, D.; CHI, H. A modified HME architecture for text-dependent speaker identification. **IEEE Transactions on Neural Network**.Vol. 7, número 5, p. 1309-1314, 1996.

DREYFUS, G. **Neural Networks – Methodology and Applications**. Springer, 2005. France/Paris : Laboratoire d'Électronique.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification.** Second edition, 1973. 738 páginas.

FERNEDA, E. **Redes neurais e sua aplicação em sistemas de recuperação de informação**. Ciência da Informação, v. 35, p. 25 – 30, 2006.

FREEMAN, J. A.; SKAPURA, D. M. **Neural networks: algorithms, application, and programmin techniques**. Computation and Neural Systems Series, 1991. 414 páginas.

FURUI, S. Speaker Recognition. **Survey of the State of the Art in Human Language Technology**. Web Edition, 1997. 543 páginas.

HU, Y. H.; HWANG, J. N. **Handbook of neural network signal processing**. Electrical engineering and applied signal processing (Series), 2002. 384 páginas.

KASABOV, N. K. **Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering**. England: London, 1998. A Bradford Book: Second Edition, 581 páginas.

KRÖSE, B.; VAN DER SMAGT, P. **An introduction to neural networks**. Amsterdam, 1996. Eighth edition, 135 páginas.

JAIN, L. C.; MARTIN, N. M. **Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications**. CRC Press, 1998. 297 páginas.

NUSSENZVEIG, H. M. **Curso de Física Básica**. Volume 2 . Editora Edgard Blücher Ltda. Terceira edição, 1997. 315 páginas.

PAULA, M. B. **Reconhecimento de palavras faladas utilizando Redes Neurais Artificiais**. Monografia de Graduação. Pelotas/SP: UFPEL, 2000.

RABUÑAL, J. R.; DORADO, J. **Artificial Neural Networks in Real-Life Applications**. Idea Group Publishing, 2006. 395 páginas.

RAO, V. B. **C++ Neural Networks and Fuzzy Logic**. M&T Books, 1995. 577 páginas.

REIS, C. F.; ALBUQUERQUE, M. P.; CASTRO, S. B. **Introdução ao reconhecimento de padrões utilizando redes neurais**. Rio de Janneiro/RJ : Centro Brasileiro de Pesquisas Físicas, 2001.

RESNICK, R.; HALLIDAY, D. **Física**. Volume 2. Livros Técnicos e Científicos Editora: Quarta Edição, 1986. 309 páginas.

SANTOS, R. O. V.; FEITOSA, R. Q.; VELLASCO, M. M. B. R.; TANSCHEIT, R. Sistemas muilti-redes para classificação de imagens multitemporais. **Anais XIII Simpósio Brasileiro de Sensoriamento Remoto**. Florianópolis, 2007, p. 6135-6142.

SEARS, F.; ZEMANSKY, M. W.; YOUNG, H. D. **Física – Mecânica dos Fluidos, Calor e Movimento Ondulatório**. Livros Técnicos e Científicos Editora: Segunda Edição, 1996. 510 páginas.

TAFNER, M. A. **Reconhecimento de palavras faladas isoladas usando redes neurais artificiais.** Florianópolis/SC: UFSC - Dissertação de Mestrado, 1996.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern recognition**. Elsevier Academic Press. Second Edition, 2003. 710 páginas.

TIPLER, P. A. **Física para cientistas e engenheiros – Gravitação, Ondas e Termodinâmica**. Volume 2, Terceira Edição, 1991.

VEELENTURF, L. P. J. **Analysis and Applications of Artificial Neural Networks**. Prentice Hall, 1995. 121 páginas.

VON ZUBEN, F. J. **Uma caricatura funcional de redes neurais artificiais**. Campinas/SP: DCA-FEEC-Unicamp, 2003. 11 páginas.

ZAMBALDE, A. L.; SILVA E PÁDUA, C. I. P. **O documento científico em ciência da computação – suas partes e sua redação: estudo e análise em uma instituição federal de ensino superior (IFES)**. Belo Horizonte/MG: DCC-UFMG, 2005.

WANG, L.; CHEN, K; CHI, H. Capture interspeaker information with a neural network for speaker identification. **IEEE Transactions on Neural Network**.Vol. 13, número 2, p. 436-446, 2002.

YOUNG, H. D.; FREEDMAN R. A. **Física II : Termodinâmica e Ondas**. Pearson Addison Wesley: Décima Edição, 2003. 328 páginas.