UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA E DE PRODUÇÃO
CURSO DE ENGENHARIA ELÉTRICA

REDES NEURAIS ARTIFICIAIS APLICADAS AO RECONHECIMENTO DE COMANDOS DE VOZ

ALEXANDRE SANTOS BRANDÃO

VIÇOSA
MINAS GERAIS – BRASIL
JUNHO/2005

REDES NEURAIS ARTIFICIAIS APLICADAS AO RECONHECIMENTO DE COMANDOS DE VOZ

ALEXANDRE SANTOS BRANDÃO

Trabalho de Conclusão de Curso submetido à Universidade Federal de Viçosa para a obtenção dos créditos referentes à disciplina Monografia e Seminário do curso de Engenharia Elétrica.

Aprovada: 29 de junho de 2005.

_____	_____
Prof. David Calhau Jorge (Membro)	Prof. Ricardo dos Santos Ferreira (Membro)
_____	_____
Prof. José Márcio Costa (Coordenador da Disciplina)	Prof. Tarcísio de Assunção Pizziolo (Orientador)

A meus pais Oséas e Cecília e a minha irmã Simone pelo incentivo e confiança depositados em mim.

Agradecimentos

Em primeiro lugar gostaria de agradecer ao professor Roselito de Albuquerque Teixeira pela paciência ao me mostrar os primeiros caminhos a seguir e muitas vezes sanar minhas principais dúvidas. E, principalmente, por ter cedido este projeto no qual me dediquei durante a minha iniciação científica e nos trabalhos restantes para a conclusão deste projeto final de curso.

Ao meu orientador, Tarcísio de Assunção Pizziolo, deixo um agradecimento especial pela escolha, norteamento e por acreditar na minha capacidade para a realização deste trabalho.

Aos colegas de trabalho, Renan Nominato e Matheus Faria, pelo auxílio nas atividades de construção do protótipo e confecção do banco de dados.

Gostaria de agradecer a Isabele Costa, Daniel Cavalieri e Antonio Ribeiro Jr por tantas vezes que me ouviram comentar minhas atividades e às vezes mesmo sem entender prestavam atenção nas minhas palavras.

Aos amigos de classe que fazem parte da minha vida, agradeço pela convivência destes anos que passamos unidos e transpassando todas as barreiras a nós impostas.

Agradeço profundamente aos meus pais, Cecília dos Santos e Oséas Brandão, a minha tia Celina dos Santos e a minha irmã Simone Brandão por me amar, apoiar e incentivar à conclusão deste curso mesmo estando a vários quilômetros de distância.

A Deus por me conceder sabedoria e forças para executar este trabalho e vida para que hoje eu possa agradecer a todas estas pessoas.

RESUMO

REDES NEURAIS ARTIFICIAIS APLICADAS AO RECONHECIMENTO DE COMANDOS DE VOZ

Resumo: *O presente trabalho tem como objetivo a implementação de um sistema de reconhecimento automático de voz (RAV) com vocabulário e número de locutores restritos. Técnicas de processamento digital de sinais e Redes Neurais Artificiais (RNA) são aqui utilizadas. A filtragem do sinal de voz é realizada por software através da aplicação de filtros digitais a fim de minimizar o efeito de borda devido ao truncamento do sinal na etapa de aquisição e os ruídos de fundo inerentes ao sinal. Rotinas para detecção de início e fim de cada Comando de Voz (CVZ) foram implementadas com a finalidade de diminuir o tamanho do vetor o qual contem as amostras do sinal de voz. Na fase de extração de características, utilizam-se os Coeficientes de Predição Linear (LPC). Os coeficientes extraídos são normalizados e utilizados para treinar a RNA. A fim de minimizar o overfitting (superajuste) por parte da RNA, utiliza-se um algoritmo de retropropagação de erro com regularização Bayesiana, que visa a maximização da capacidade de generalização da rede. Esta, após as etapas de treinamento e validação, é capaz de reconhecer CVZ de palavras isoladas para um vocabulário restrito a cinco comandos e um conjunto restrito de três locutores. Executado estas operações sobre um CVZ, pôde-se verificar uma redução no esforço computacional na etapa de extração de características, devido a um menor volume de dados a interpretar, em média 73,6% em relação ao vetor original.*

Palavras-chave: *comandos de voz, processamento digital de sinais, redes neurais artificiais.*

ABSTRACT

The present work has as objective the implementation of an Automatic Voice Recognition (AVR) system with restricted vocabulary and number of speakers. Techniques of Digital Signal Processing (DSP) and Artificial Neural Network (ANN) were used. The voice signal's filtering was realized in software by the application of digital filters. Routines to find the begin and the end of each voice command were implemented to reduce the vector's length whom contains the voice signal samples. Linear Prediction Coefficients (LPC) were utilized to extract the voice's characteristics. These coefficients extracted were standardized and used to train ANN structure. For the purpose to minimize the ANN overfitting, in this work was utilized the back-propagation algorithm with Bayesian regularization. After training and validation, the ANN is capable to recognize voice command of single words in a restricted vocabulary with five commands and a restricted set of three speakers. With these operations, it's possible to minimize 73,6% of the dates and verify a reduction into the computational strive.

Keywords: *Voice Command, Digital Processing Signal, Artificial Neural Network*

Lista de Figuras

Figura 1 - Diagrama de blocos do modelo geral de um sistema RAV *14*
Figura 2 - Diagrama de blocos do filtro digital tipo FIR forma direta *16*
Figura 3 - (a) Comando de Voz (CVZ) no tempo (b) Energia de um CVZ não normalizado *20*
Figura 4 - Neurônio de McCulloch e Pitts *22*
Figura 5 - Classificação das RNA. (a) Rede de Camada única (b) Redes de múltiplas camadas (c) Redes feedback. (a) e (b) são exemplos de redes feed forward. *23*
Figura 6 - Arquitetura das Redes MLP *24*
Figura 7 - Pronúncia de um comando de voz (a) Amostra do padrão "pára" não-saturada (b) Amostra do padrão "atrás" saturada *30*
Figura 8 - Pronúncia do padrão "direita" para visualização de ruídos de fundo na zona de silêncio *30*
Figura 9 - Amplificação de ruído por aplicação indevido de janela de Hamming sobre uma amostra do padrão "atrás" (a) Sinal original adquirido (b) Sinal janelado *31*
Figura 10 - Pronúncia do padrão "esquerda", análise de energia do sinal e centralização do sinal com base no bico de energia *32*
Figura 11 - (a) Sinal de voz filtrado e não normalizado (b) Sinal de voz normalizado tomando como base o maior valor em módulo do sinal (c) Sinal de voz normalizado tomando como referência o máximo valor negativo e positivo do sinal. *35*
Figura 12 - Detecção de início e fim utilizando o método de energia do sinal. E_{i1}, E_{i2}, E_{f1} e E_{f2} são os patamares predefinidos de energia, bem como a largura do salto. *36*
Figura 13 - Análise de energia do Filtro Bilinear e Butterworth *39*
Figura 14 - Análise de Fourier para o sinal original e processado *40*
Figura 15 - Comparação entre a amostra do sinal original e do processado *41*
Figura 16 - Diagrama de Funcionamento do Sistema RAV em Execução no Protótipo *48*

Lista de Tabelas

Tabela 1 – Tipos de janelas generalizadas de co-seno *17*
Tabela 2 - RNA treinada pelo algoritmo BR para 8 LPC *42*
Tabela 3 – RNA treinada pelo algoritmo BR para 10 LPC *42*
Tabela 4 – RNA treinada pelo algoritmo BR para 12 LPC *43*
Tabela 5 – RNA treinada pelo algoritmo LM para 8 LPC *43*
Tabela 6 – RNA treinada pelo algoritmo LM para 10 LPC *44*
Tabela 7 – RNA treinada pelo algoritmo LM para 12 LPC *44*
Tabela 8 – Resumo dos resultados alcançado nos treinamentos dos padrões pelos algoritmos de Levenberg - Marquard e Regularização Bayesiana *45*
Tabela 9 - Codificação das ações de acordo com o CVZ pronunciado *49*
Tabela 10 - Lista de comandos e os resultados de um locutor previamente treinado perante a RNA para a validação do sistema RAV *50*
Tabela 11 - Lista de comandos e os resultados de um locutor desconhecido perante a RNA para a validação do sistema RAV *51*
Tabela 12 – Comparação do comando de voz antes e após processamento digital de sinais *53*

Abreviações

BR *Bayesian Regularization*

CVZ Comando de Voz

DSP *Digital Signal Processing*

FFT *Fast Fourier Transform*

FIR *Finite Impulse Response*

IIR *Infinity Impulse Response*

LPC *Linear Prediction Coefficient*

LM Levenberg - Marquardt

MLP *Multi Layer Perceptron*

MSE *Mean Square Error*

RAV Reconhecimento Automático de Voz

RL Reconhecimento de Locutor

RNA Redes Neurais Artificiais

Sumário

Lista de Figuras ____ *7*

Lista de Tabelas ____ *8*

Abreviações ____ *9*

1 Introdução ____ *11*

2 Revisão Bibliográfica ____ *12*

2.1 Sistemas de Reconhecimento de Voz ____ **12**

2.1.1 Introdução ____ 12

2.1.2 Base de Dados ____ 13

2.1.3 Reconhecimento do Sinal de Voz Baseado na Comparação de Padrões ____ 13

2.2 Processamento do Sinal de Fala ____ **15**

2.2.1 Introdução ____ 15

2.2.2 Filtros Digitais ____ 15

2.2.3 Análise de Energia ____ 19

2.3 Redes Neurais Artificiais ____ **20**

2.3.1 Introdução ____ 20

2.3.2 Definição ____ 21

2.3.3 Classificação das RNA ____ 22

2.3.4 Redes de Múltiplas Camadas ____ 24

3 Objetivos ____ *28*

3.1 Objetivo Geral ____ **28**

3.2 Objetivos Específicos ____ **28**

4 Materiais e Métodos ____ *29*

4.1 Base de Dados ____ **29**

4.2 Processamento do Sinal de Fala ____ **30**

4.2.1 Detecção de Início e Fim ____ 30

4.3 Redes Neurais ____ **37**

5 Resultados e Discussões ____ *39*

6 Conclusões ____ *46*

7 Referências Bibliográficas ____ *47*

Apêndice A – Exemplo de Aplicação ____ *48*

A.1 Desenvolvimento ____ **48**

A.2 Resultados e Discussões ____ **49**

A.3 Conclusão ____ **52**

Apêndice B – Padrões de Voz Antes e Após Processamento Digital de Sinais ____ *53*

1 Introdução

A fala para as pessoas dotadas desta habilidade é uma tarefa trivial que desenvolvemos continuamente no decorrer da vida. De forma similar entendemos ou tentamos entender o que as outras pessoas estejam pronunciando. Todavia, para máquinas e computadores, o processo de reconhecimento de uma locução é uma tarefa bastante complexa, sendo necessário um treinamento prévio e um processamento do sinal de voz.

Segundo Moreira, 1998 **[1]**, o modelo de comunicação utilizando a fala apresenta três etapas: a produção, a transmissão e a recepção da fala. Na produção do sinal de fala, o locutor codifica a informação que pretende transmitir em símbolos de uma estrutura lingüística e, posteriormente, materializa esses símbolos em unidades acústicas. Tendo no próprio aparelho auditivo uma realimentação para avaliar a qualidade do sinal pronunciado. Na transmissão ocorrem as maiores interferências sobre o sinal falado, pois pode haver sobreposição do sinal por ruídos proveniente de outras pessoas que por eventualidade estejam emitindo outros sinais sonoros irrelevantes, como também ruídos de fundo gerados por equipamentos e máquinas. Finalmente, realizando um processo inverso ao de produção, os ouvintes tentam extrair as informações contidas no sinal pronunciado, na etapa recepção da fala.

Para uma integração perfeita homem-máquina seria necessário uma simulação da realidade, pois parte do processo de comunicação entre os seres humanos é constituída por meios explícitos (por exemplo, sorrisos, gestos, olhares, dentre outros), os quais são desconsiderados durante a aquisição de um sinal de voz.

Um sistema de reconhecimento de voz pode objetivar reconhecer o locutor ou o que foi dito por este locutor **[2]**. O Reconhecimento de Locutor (RL) deve verificar se o locutor que pronunciou uma cadeia de caracteres é realmente quem ele diz ser, ou ainda, verificar dentre um conjunto de oradores, qual deles efetuou uma declaração. O Reconhecimento Automático de Voz (RAV) tem como base a identificação de padrões previamente treinados por algum algoritmo de reconhecimento que porventura foi pronunciado por um locutor.

2 Revisão Bibliográfica

2.1 Sistemas de Reconhecimento de Voz

2.1.1 Introdução

Um sistema de reconhecimento de sinais de fala pode ser classificado de acordo com o tamanho do vocabulário, o grau de independência do locutor e o tipo de pronúncia **[3]**.

Considerando o tamanho do vocabulário, este pode ser pequeno (até uma centena de palavras distintas), médio (com algumas centenas de palavras), grande (com milhares de palavras). O sistema de reconhecimento torna-se mais complexo, à medida que o vocabulário aumenta, pois treinar milhares de padrões de referência, por exemplo, tornaria o processo bastante custoso, logo, inviável.

Quanto ao grau de dependência do locutor, temos os sistemas dependente e independente de locutor. No primeiro caso, o sistema está restrito a um conjunto específico de locutores previamente treinados. No segundo, encontramos um sistema generalizado capaz de reconhecer Comandos de Voz (CVZ) de uma infinidade de locutores, em contrapartida é um sistema mais difícil de ser construído.

A pronúncia dos CVZ, por sua vez, pode ser executada de forma contínua ou pausada. Existem sistemas de reconhecimento que utilizam palavras isoladas (palavras pronunciadas com pausas entre si), palavras concatenadas (seqüências de palavras pré-estabelecidas faladas de forma contínua) e fala contínua (frases e orações pronunciadas de forma contínua). No último sistema encontramos o maior grau de complexidade, pois os limites de início e fim de um padrão de referência são mais difíceis de encontrar, devido aos efeitos de coarticulação entre as palavras.

2.1.2 Base de Dados

Conforme mencionada anteriormente, a fala é uma das maneiras mais naturais de comunicação entre os seres humanos, todavia há uma série de fatores, que influenciam na identificação de um padrão, como por exemplo: sotaque, dialeto, tamanho do trato vocal, velocidade da pronúncia, dentre outros. Como também, o estado físico-psicológico e cultura do narrador.

Para a confecção de uma base de dados é necessário estabelecer algumas unidades básicas finitas que tenham como características: consistência e treinabilidade. Consistência, a fim de que uma subunidade fonética apresente características similares independentemente do instante em que ocorreu a aquisição; e treinabilidade, para que o sistema conste de um número considerável de amostras visando um modelo mais robusto **[4]**. Unidades maiores, tais como frase ou palavras, são consistentes, todavia são depreciadas quando tratamos da treinabilidade. Por outro lado, o inverso ocorre nas unidades menores, tais como fones, que são treináveis, porém inconsistentes **[5]**.

Para tornar o modelo o mais robusto possível são necessárias aquisições de um número grande de amostras de modo a facilitar a distinção entre os padrões utilizados para treinamento do sistema de reconhecimento.

2.1.3 Reconhecimento do Sinal de Voz Baseado na Comparação de Padrões

A estrutura de processamento para reconhecimento de um sinal de voz deste trabalho pode ser seguida segundo diagrama de blocos da Figura 1 **[6]**.

Processamento do Sinal

Nesta etapa, o sinal analógico de voz é representado sob a forma digital e processado de forma a eliminar ou atenuar ruídos no canal, variações na amplitude ou, ainda, estresses do locutor (estado emocional).

Para reconhecedores de palavras isoladas, como é o caso deste trabalho, torna-se necessário determinar o início e o fim de uma amostra de voz a fim de separar ruídos de fundo do sinal. Para tornar o sinal mais puro possível, para posteriormente extrair as características, filtros digitais são aplicados com objetivo de excluir freqüências irrelevantes.

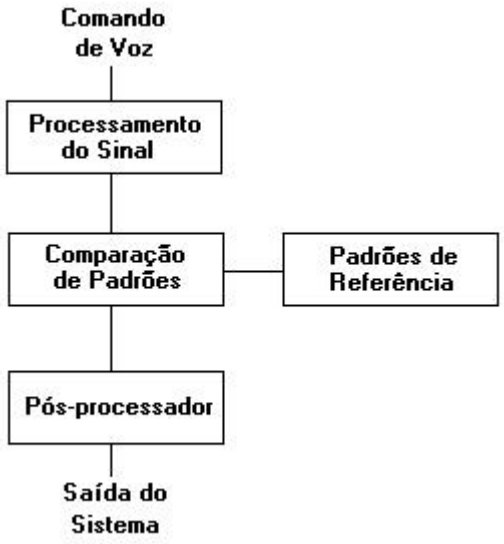


Figura 1 - Diagrama de blocos do modelo geral de um sistema RAV

Padrões de Referência

Os padrões são estabelecidos no momento em que se definem quais as palavras serão reconhecidas, entretanto os padrões de referência são escolhidos aleatoriamente entre todas as amostras adquiridas. Estes padrões são utilizados para treinar o sistema e as demais amostras para fazer o teste e a validação do mesmo **[19]**. Quanto maior for a diversidade existente entre os padrões de referência de uma mesma unidade, maior será a robustez do sistema, pois a quantidade de informação abstraída também será maior durante a etapa de treinamento.

Comparação de Padrões

Após a etapa de treinamento, outro conjunto de amostras distintas dos padrões de referência é inserido no sistema para que se possa analisar a eficiência do reconhecedor. As amostras desconhecidas são comparadas com os padrões de referência

e o reconhecedor apresenta um resultado possibilitando a verificação ou não da generalização do sistema de reconhecimento.

Pós-processador

Designado como a última etapa do sistema RAV, o pós-processador tem como objetivo fazer os ajustes finais de amplitude da saída da estrutura neural avaliada e conceber a afirmação de que o comando de voz foi identificado sem que haja dupla interpretação das demais estruturas envolvidas, ou seja, uma amostra de CVZ seja identificada unicamente pela estrutura que a representa.

2.2 Processamento do Sinal de Fala

2.2.1 Introdução

Para um sistema de RAV é viável converter um sinal de áudio em um conjunto estreito de parâmetros que contenham as informações relevantes do CVZ a fim de facilitar a etapa de treinamento do sistema e minimizar o número de operações matemáticas a serem executadas.

Nesta etapa ocorre a conversão do sinal de modo que o computador possa interpretá-lo. Esta conversão é realizada no momento da aquisição do sinal durante a criação da base de dados, onde ocorre a conversão analógico-digital. Nesta etapa também ocorre a detecção de início e fim, aplicação de filtros e extração dos Coeficientes de Predição Linear (LPC – *Linear Prediction Coefficients*). Este texto é uma adaptação de **[17] [18] [11]**.

2.2.2 Filtros Digitais

Filtros digitais são operadores lineares empregados sobre dados digitalizados (ou amostrados) que permitem a passagem ou o corte de certas freqüências conforme as características do filtro. Estes operadores lineares podem ser descritos pela equação (1).

$$y(n) = \sum_{q=0}^{M} b_q x(n-q) - \sum_{p=1}^{N} a_p y(n-p) \qquad \text{eq. (1)}$$

Onde *x(n)* são os dados amostrados de entrada para o filtro e *y(n)* é a saída resultante do mesmo. Os coeficientes b_q e a_p são os coeficientes de entrada e saída do filtro, respectivamente.

Existem duas grandes classes de filtros digitais. Na primeira, encontram-se os filtros recorrentes, ou filtros IIR (*Infinite Impulse Response*), onde os coeficientes b_q e a_p são valores reais. Na segunda, encontram-se os filtros de resposta finita (FIR – *Finite Impulse Response*), onde os coeficientes a_p apresentam valor nulo para todo valor de *p*. Este último tem como característica a extinção do impulso para um número finito de amostra, além de apresentar uma resposta em fase linear; tornando, portanto viável sua aplicação em tratamento de sinais de voz.

O diagrama de blocos da Figura 2 apresenta as operações realizadas para calcular cada elemento do vetor de saída *y(n)* por um filtro digital FIR.

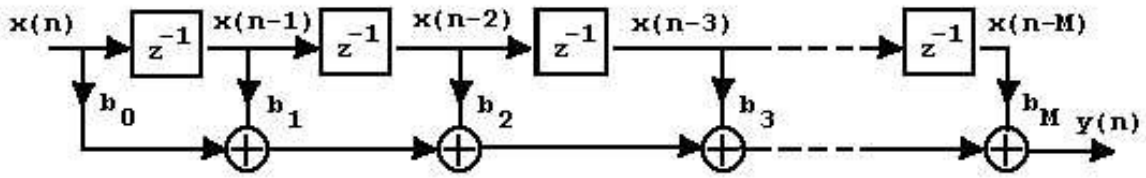


Figura 2 - Diagrama de blocos do filtro digital tipo FIR forma direta

As janelas são filtros digitais que visam reduzir descontinuidades causadas devido ao truncamento do sinal no tempo (denominado janela retangular) tratando suavemente o sinal em suas extremidades e destacando as informações contidas na região central da amostra. Algumas janelas denominadas janelas generalizadas de co-seno apresentam a equação (2) **[10]**:

$$y(n) = A - B\cos\left(\frac{2\pi.n}{N}\right) + C\cos\left(\frac{4\pi.n}{N}\right) \qquad \text{eq. (2)}$$

Onde *N* indica o tamanho da janela, que é equivalente ao tamanho da amostra. A variação dos parâmetros *A*, *B* e *C* possibilita a obtenção de diversas janelas, conforme descrito na Tabela 1 a seguir.

Tabela 1 – Tipos de janelas generalizadas de co-seno

Nome da Janela	Parâmetros		
	A	*B*	*C*
Hanning	0.50	0.50	0.00
Hamming	0.54	0.46	0.00
Blackman	0.42	0.5	0.08

Um outro tipo de filtro digital pode ser conseguido através da transformação bilinear, que converte uma função no domínio da freqüência em uma função discreta equivalente. A transformação bilinear mapeia o plano s no plano z através da seguinte função de transferência:

$$H(z) = H(s)\Big|_{s=2f_s\frac{z-1}{z+1}} \qquad \text{eq. (3)}$$

Os filtros digitais são necessários para suprimir de um sinal de voz informações irrelevantes. Segundo Adami **[2],** um sinal de voz para efeito de percepção da fala apresenta freqüências entre 100Hz e 5kHz. Estas freqüências, denominadas formantes, são suportadas por uma outra denominada freqüência fundamental ou *pitch*, que é uma oscilação quase periódica em torno de 80 a 200 Hz.

A partir das informações de freqüência, pode-se filtrar o sinal de voz, a fim de que este apresente em seu espectro de freqüência, simplesmente as freqüências de interesse. Como por exemplo, um filtro passa faixa de segunda ordem analógico idealizado de acordo com a função de transferência dada pela equação (4).

$$H(s) = \frac{Ks}{s^2 + as + b} \qquad \text{eq. (4)}$$

Onde o módulo máximo da função de transferência dado por:

$$\left|H(j\omega)\right|_{\max} = \frac{K}{a} \qquad \text{eq. (5)}$$

Na freqüência de corte ω_c tem-se que:

$$\left|H(j\omega_c)\right| = \frac{1}{\sqrt{2}}\left|H(j\omega)\right|_{\max} \qquad \text{eq. (6)}$$

Desta forma, as freqüências de corte podem ser calculadas por:

$$\omega_{c1} = \frac{-a + \sqrt{a^2 + 4b}}{2} \text{ e } \omega_{c2} = \frac{a + \sqrt{a^2 + 4b}}{2} \qquad \text{eq. (7)}$$

Tendo em mãos as freqüências de corte, basta encontrar os parâmetros que modelem o filtro desejado.

Os filtros Butterworth de ordem *N* são também filtros digitais, cuja resposta em é módulo é dada pela equação (8), e se enquadram no critério de qualidade de resposta em módulo maximamente plana quando ω tende a ser igual a zero.

$$\left|H(j\omega)\right|^2 = \frac{1}{1 + \left(\frac{\varpi}{\omega_c}\right)^{2N}} \qquad \text{eq. (8)}$$

Onde ω_c é a freqüência de corte deste filtro passa baixa. Considerando que os filtros Butterworth definidos desta forma são isentos de zeros, os seus pólos podem ser obtidos de acordo com a equação (9), a seguir:

$$s_k = \omega_c e^{j\pi(2k+N-1)/2N} \text{ , para k = 0, 1, ..., 2N-1.} \qquad \text{eq. (9)}$$

Os pólos determinados se encontram à direita do plano complexo, indicando a estabilidade do filtro, cuja função de transferência pode ser dada por:

$$H(s) = \frac{H_o}{\prod_{k=1}^{N}(s - p_k)} \qquad \text{eq. (10)}$$

Onde:

$$H_o = \prod_{k=1}^{N}(-p_k) \qquad \text{eq. (11)}$$

E p_k são os pólos encontrados pela equação (9). Para tornar a função de transferência do filtro Butterworth passa baixa em uma passa faixa, faz-se a seguinte transformação de variáveis:

$$s' = \frac{s^2 + \omega_o^2}{Bs} \qquad \text{eq. (12)}$$

Onde ω_o é a freqüência central do filtro definida por:

$$\varpi_0 = \sqrt{\varpi_{c_1}\varpi_{c_2}} \qquad \text{eq. (13)}$$

E B é a largura da faixa de passagem dada por:

$$B = \varpi_{c_2} - \varpi_{c_1} \qquad \text{eq. (14)}$$

2.2.3 Análise de Energia

Toda pronúncia de um sinal de voz possui uma curva de energia característica, que pode ser interpretada como o quadrado do valor ponto a ponto do sinal amostrado. A medida da energia ou potência de um CVZ é dada por:

$$E = \sum_{n=1}^{N} x(n)^2 \qquad \text{eq. (15)}$$

A partir da curva de energia é possível localizar os limites de uma palavra isolada, como é o caso de um CVZ, que apresenta picos de energia durante a evolução do tempo e são nestes picos que se encontram as informações a serem processadas. A Figura 3a mostra um sinal de voz no tempo após aquisição e a Figura 3b, a energia deste sinal não normalizado.

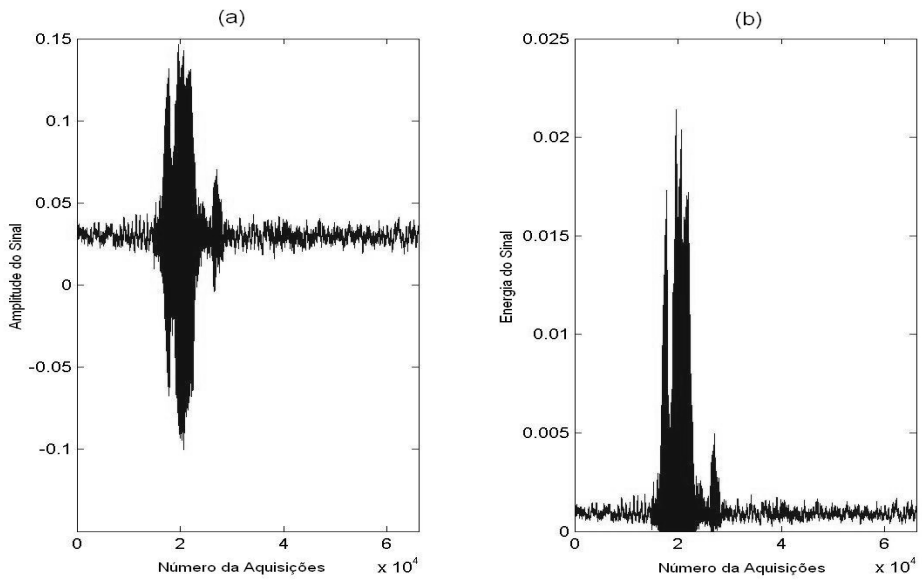


Figura 3 - (a) Comando de Voz (CVZ) no tempo (b) Energia de um CVZ não normalizado

2.3 Redes Neurais Artificiais

2.3.1 Introdução

As Redes Neurais Artificiais (RNA) são estruturas matemáticas capazes de aprender, memorizar e generalizar determinadas situações e problemas a elas apresentadas. Texto adaptado de **[20] [21] [22]**.

As RNA são sistemas paralelos compostos por unidades elementares, denominadas neurônios, que calculam determinadas funções matemáticas geralmente não-lineares, cujo funcionamento é inspirado no próprio cérebro humano.

As soluções por meio das RNA podem se equivaler ou mesmo superar as soluções apresentadas pela programação tradicional. O procedimento pelo qual uma RNA encontra as soluções passa por um processo de aprendizado, onde uma série de amostras de entrada e saída é apresentada às suas unidades elementares que por si só encontram as características necessárias para representar a informação fornecida, e posteriormente, definir o sistema resultante.

As RNA têm como capacidade aprender com os exemplos que lhe são apresentados e generalizar a informação aprendida, sendo possível, portanto, a classificação de amostras de dados desconhecidos, mas que se assemelhe com a informação contida na etapa de treinamento. As RNA são capazes de extrair características que não estejam explicitamente apresentadas sob a forma de exemplos (ou amostras de entrada).

2.3.2 Definição

O cérebro humano é constituído por cerca 10^{11} neurônios, que recebem e enviam informações para milhares de outros a eles conectados. O cérebro é destinado a cuidar em nosso corpo no que se trata de emoção, raciocínio e funções motoras. As RNA, por sua vez, têm como ambição simular este mundo de atividades realizadas pelo cérebro, implementando o seu comportamento básico e sua dinâmica.

O modelo matemático de um neurônio artificial foi proposto por Warren McCulloch, psiquiatra e neuroanatomista, e Walter Pitts, matemático, em 1943. O modelo em si era uma simplificação do neurônio biológico até então conhecido na época. Para representar os dendritos, o modelo constou de n terminais de entrada de informações $x_1, x_2, ..., x_n$ e simplesmente um terminal de saída y, para representar o axônio. Cada entrada apresenta um coeficiente ponderador que visa à simulação das sinapses, sendo que estes coeficientes são valores reais. De forma análoga ao neurônio

biológico, a sinapse só ocorre quando a soma ponderada dos sinais de entrada ultrapassa um limiar pré-definido, realizando, portanto uma atividade semelhante a do corpo. No modelo proposto, o limiar foi definido de forma Booleana, dispara ou não dispara, resultante de uma função de ativação, conforme pode ser visto na Figura 4.

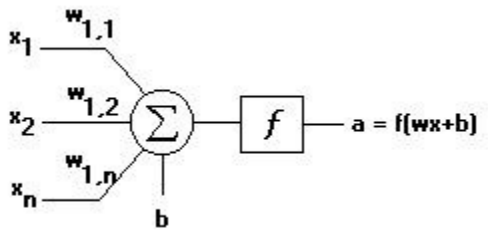


Figura 4 - Neurônio de McCulloch e Pitts

A saída *y* do neurônio de McCulloch e Pitts pode ser equacionada por:

$$y = f(\sum_{i=1}^{n} x_i w_i + b) \qquad \text{eq. (16)}$$

Onde *n* é o número de entradas do neurônio, w_i é o peso associado à entrada x_i e *f* é a função de ativação utilizada.

Todavia, uma restrição existente no modelo criado é que as redes desenvolvidas só conseguem implementar funções linearmente separáveis, ou seja, aquelas que se podem separar os padrões por meio de uma reta.

2.3.3 Classificação das RNA

As RNA podem ser classificadas:

a) Quanto a sua estrutura:
   - Redes de camada única: existe simplesmente um nó entre o vetor de dados de entrada e o vetor de saída (Figura 5a);
   - Redes de múltiplas camadas: existem mais de uma camada de neurônios entre o vetor de entrada de dados e o vetor de saída (Figuras 5b e 5c).

b) Quanto as suas conexões:

- Redes feed forward: são estruturas neurais, onde a saída de um neurônio de uma dada camada realiza conexões com os neurônios das camadas seguintes e não com os das camadas antecessoras (Figuras 5a e 5b).
- Redes feedback: são aquelas cuja saída de uma camada, também pode atuar tanto na entrada dos neurônios das camadas antecessoras como das posteriores (Figura 5c).

A Figura 5 ilustra as estruturas e conexões das RNA.

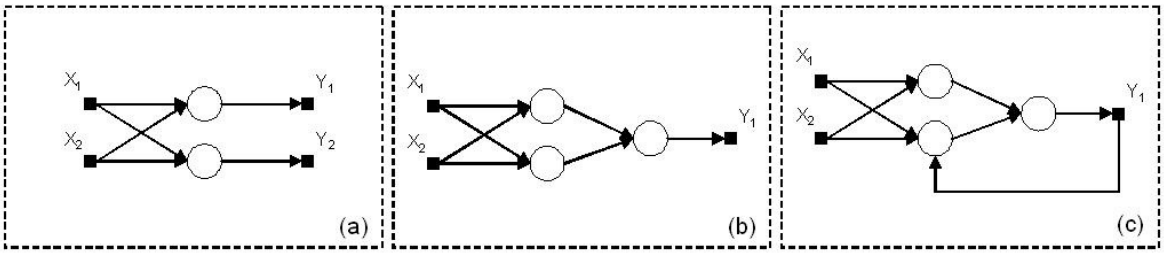


Figura 5 - Classificação das RNA. (a) Rede de Camada única (b) Redes de múltiplas camadas (c) Redes feedback. (a) e (b) são exemplos de redes feed forward.

c) Quanto ao treinamento:

- Supervisionado: caracteriza-se pela existência de um professor, ou supervisor, que monitora a resposta da rede e compara esta com a resposta desejada, e a partir do erro existente entre estas respostas faz-se o ajuste dos pesos sinápticos até que um erro mínimo estabelecido seja alcançado ou o número de iterações pré-estabelecido seja superado.
- Não supervisionado: caracteriza-se pela não existência de saídas desejadas, sendo, portanto, o conjunto de treinamento é estabelecido somente pelo vetor de entrada. A atualização dos pesos sinápticos é obtida com base nos próprios valores de entrada. Este tipo de treinamento é aplicado a problemas de categorização de dados.
- Reforço: é um treinamento que mescla as características do treinamento supervisionado e do não supervisionado. O conjunto de treinamento é

formado simplesmente por dados de entrada, todavia há um crítico, similar ao supervisor, que reforça ou penaliza a saída da rede.

2.3.4 Redes de Múltiplas Camadas

Conforme dito anteriormente, as RNA com uma camada de neurônio são capazes de resolver problemas linearmente separáveis, contudo apesar de resolver uma gama vasta de problemas, existe, por outro lado, uma outra vasta coleção de problemas não linearmente separáveis. Este problema foi proposto por Minsky e Pappert na década de 70, quando, em suas publicações, depreciaram a habilidade das RNA de encontrar soluções para simples problemas, como por exemplo, a modelagem do "Ou Exclusivo" da lógica digital. A solução encontrada para contornar este problema e como conseqüência retomar as pesquisas sobre RNA, até então desacreditas por Minsky e Pappert, foram as estruturas neurais de múltiplas camadas, também conhecida como redes MLP (Multi Layer Perceptron), na década de 80.

As redes MLP apresentam a arquitetura mostrada na Figura 6, onde se encontram a camada de entrada, as camadas intermediárias (ou ocultas) e a camada de saída. O número de variáveis da camada de entrada depende diretamente do número de características agrupadas no vetor das amostras. O número de neurônios das camadas intermediárias depende da complexidade do problema. E a camada de saída contém o número de neurônios necessário para executar a codificação das amostras de entrada.

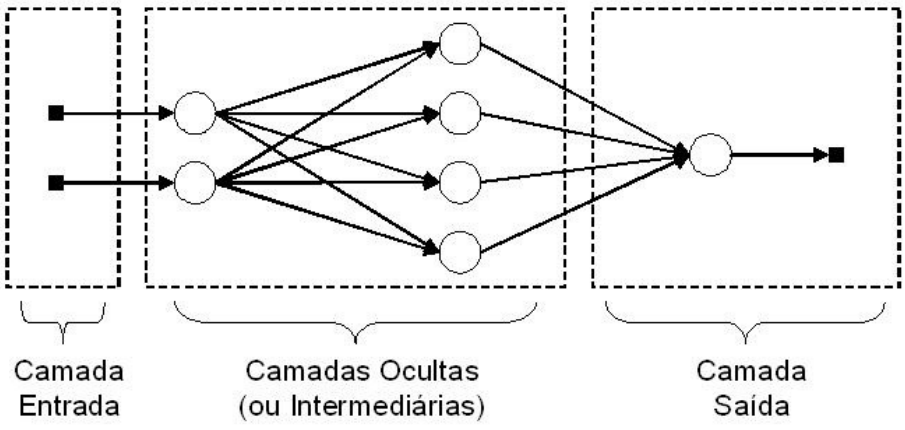


Figura 6 - Arquitetura das Redes MLP

O número de neurônios das camadas intermediárias é determinado de forma empírica, atentando para o caso de *overfitting* (ou superajuste), que é o caso onde existe uma grande quantidade de neurônios e a estrutura ao invés de generalizar as informações, acabar por memorizar os padrões apresentados, não sendo capaz de classificar padrões semelhantes. Outro efeito do superajuste é que a RNA além de armazenar as características relevantes extraídas das amostras, esta guarda em seus pesos informações de ruídos que a princípio não revelam interesse. Por outro lado, caso o número de neurônios seja inferior ao desejado, pode ocorrer um *underfitting* e a RNA não convergir para uma resposta devido a uma sobrecarga de informações a serem armazenadas em poucos pesos.

O treinamento das redes MLP é normalmente realizado pelo algoritmo de retro-propagação do erro (ou *back-propagation*), um algoritmo supervisionado que realiza o ajuste dos pesos, a partir do erro existente entre os pares de amostra de dados de entrada e saída da RNA. Este algoritmo apresenta duas fases denominadas *forward* e *backward*. A fase *forward* é utilizada para que seja encontrada uma saída a partir dos valores de entrada de um dado padrão. A fase *backward* compara esta saída com a saída desejada e retorna atualizando os valores dos pesos das conexões dos neurônios da estrutura.

Basicamente, o processo de treinamento da rede MLP tem como objetivo, durante o treinamento, a minimização da função de erro quadrático médio (mse – *mean square error*), definida pela equação (17). O treinamento da RNA deve ocorre até que se complete um número predeterminado de iterações para atualização dos pesos ou quando o erro quadrático médio encontrar-se abaixo de um valor pré-estabelecido.

$$mse = \frac{1}{k}\sum_{p}\sum_{i=1}^{k}\left(d_i^p - y_i^p\right)^2 \qquad \text{eq. (17)}$$

Onde p é o número de padrões apresentados à estrutura neural, k é o número de unidades de saída, d_i é a i-ésima saída desejada e y_i é a i-ésima saída gerada pela rede.

Uma grande dificuldade de encontrar uma solução viável diz respeito aos mínimos locais, que apresentam características semelhantes às do mínimo absoluto,

todavia resultam em resposta equivocadas e geralmente incorretas. Existem algumas formas de aumentar a generalização das RNA evitando a priori a incidência dos mínimos locais, dentre estas estão a regularização e a parada precoce **[13]**.

A regularização envolve diretamente uma modificação na função de desempenho, que é a função de erro quadrático médio. A modificação é realizada ao acrescentar um termo que estará diretamente ligado à soma dos quadrados dos pesos da rede neural, definido por:

$$msw = \frac{1}{k}\sum_{p}\sum_{i=1}^{k}\left(w_i^p\right)^2 \qquad \text{eq. (18)}$$

E o erro quadrático médio da regularização é dado por:

$$msereg = \gamma * mse + (1-\gamma) * msw \qquad \text{eq. (19)}$$

Onde γ é a razão de desempenho. Este tipo de função de desempenho força uma redução nos valores dos pesos e conseqüentemente força uma redução do superajuste. Contudo um outro problema encontrado é a determinação do valor ótimo para o coeficiente γ. Tendo em vista que γ deve estar compreendido na faixa $0 \leq \gamma \leq 1$, se houver uma elevação do valor da razão de desempenho acima do necessário, poderá ocorrer um superajuste; por outro lado, se a razão for muito pequena, a rede não estará ajustando adequadamente os dados de treinamento. Para facilitar a busca por este ponto ótimo, foi desenvolvido por David Mackay um algoritmo de regularização automática dos parâmetros, denominado regularização Bayesiana, que computa, após cada iteração, o valor de γ a ser utilizado.

O método de parada precoce (ou *early stopping*) é um método de melhoria da generalização das RNA, que oferece modificação durante a etapa de treinamento. Esta técnica fraciona as amostras em três grupos distintos. O primeiro grupo são as amostras de treinamento, que são utilizadas para ajustar os pesos. O segundo grupo são as amostras de validação. O erro de validação é monitorado durante a etapa de treinamento, caso este cresça o treinamento pára, independente se o erro obtido na etapa

de treinamento esteja decrescendo. Normalmente, no início do treinamento, tanto o erro de treinamento quanto o erro de validação decrescem no decorre das iterações, entretanto, à medida que começa a ocorre um *overfitting* o erro de validação começa a crescer. O terceiro grupo é constituído das amostras de teste, as quais deverão se apresentadas à RNA treinada.

3 Objetivos

3.1 Objetivo Geral

Este trabalho tem como objetivo identificar comandos de voz (CVZ) de palavras isoladas com dependência do locutor que as pronuncia utilizando técnicas de Processamento Digital de Sinais (DSP – *Digital Signal Processing*), para aquisição, filtragem e extração de características do sinal de voz e Redes Neurais Artificiais (RNA) para realizar a tarefa de reconhecimento de padrões.

3.2 Objetivos Específicos

- Estudar o processo de formação da fala;
- Identificar as características do sinal de voz a serem utilizadas para facilitar o processo de reconhecimento do mesmo;
- Estudar Redes Neurais Artificiais (RNA) e, posteriormente, aprofundar na área dos problemas envolvendo reconhecimento de sinais de voz;
- Estudar métodos de reconhecimento por comparação de padrões; e,
- Verificar a possibilidade de implementação de um sistema de Reconhecimento Automático de Voz (RAV) em tempo real.

4 Materiais e Métodos

4.1 Base de Dados

Visando um aprendizado inicial sobre RAV, os padrões escolhidos para constituir a base de dados são as posições: frente, atrás, direita, esquerda e pára. O objetivo futuro é manipular uma máquina específica que possa ter seu controle ativado por voz.

Para a confecção da base de dados, foram selecionados três locutores do sexo masculino estudantes de graduação. Cada locutor pronunciou os padrões supracitados na ordem descrita por vinte vezes, para não ocorrer depreciação do sistema devido a treinabilidade do padrão.

Como característica, o sinal de voz apresenta sua natureza contínua, logo foi necessária uma conversão analógico-digital, para que o mesmo pudesse ser armazenado no disco rígido. A aquisição dos dados foi realizada num ambiente de laboratório relativamente silencioso a fim de facilitar a extração das características, através de um microfone direcional e uma placa de som SoundBlaster AWE 64. As amostras foram adquiridas a uma taxa de amostragem de 22.050Hz com resolução de 16 bits e armazenadas no formato wave. Logo, segundo o Teorema da Amostragem de Nyquist, as amostras possuem freqüências de até 11kHz.

Depois de realizada a aquisição, cada amostra de cada padrão teve sua forma de onda analisada em um programa gráfico e, utilizando fones de ouvido, verificou-se a qualidade do sinal. Excluíram-se as amostras onde ocorriam à saturação pela placa de som e que ocasionaria ruídos irreversíveis (Figura 7). Para manter a igualdade de número de amostras por padrão, foram realizadas novas aquisições pelos mesmos locutores das amostras saturadas, a fim de substituí-las.

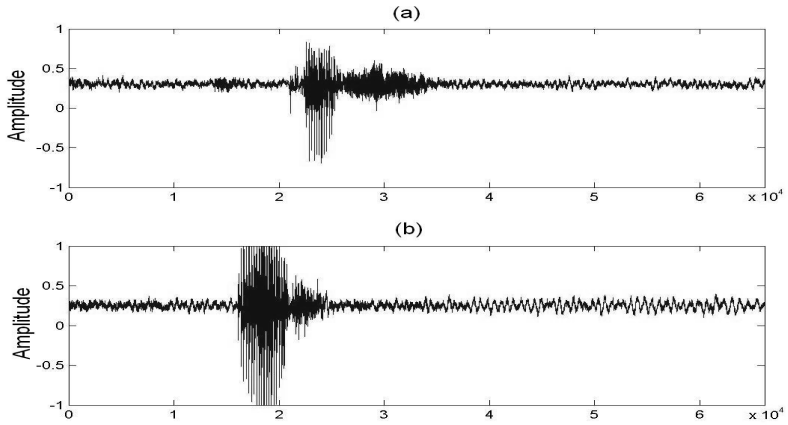


Figura 7 - Pronúncia de um comando de voz (a) Amostra do padrão "pára" não-saturada (b) Amostra do padrão "atrás" saturada

4.2 Processamento do Sinal de Fala

4.2.1 Detecção de Início e Fim

A Figura 8 apresenta uma grande parte do CVZ que é constituído simplesmente de ruídos de fundo na zona de silêncio, sendo, portanto irrelevante para o sistema de RAV. A detecção de início e fim do CVZ visa eliminar esta zona e resultar simplesmente na região onde realmente se encontra a pronúncia do padrão.

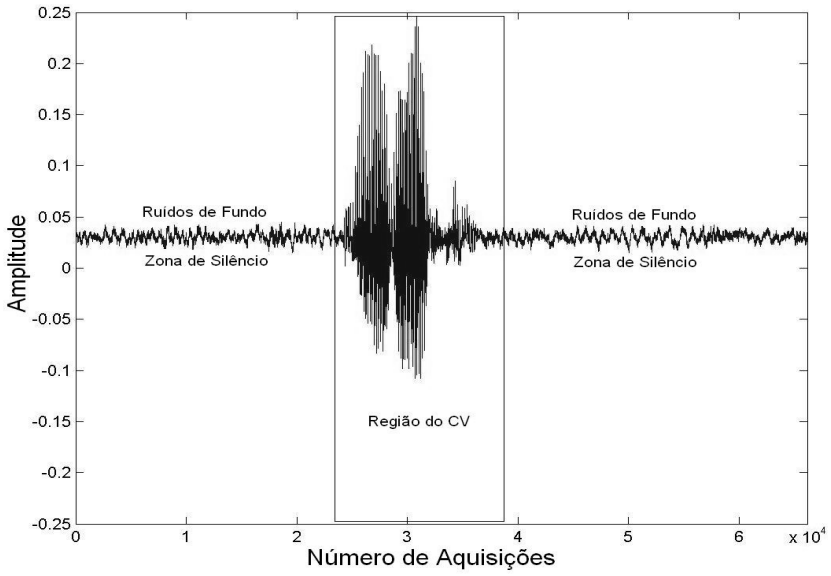


Figura 8 - Pronúncia do padrão "direita" para visualização de ruídos de fundo na zona de silêncio

Centralização do Comando de Voz

Nem todas as amostras adquiridas apresentam o pico de maior amplitude em módulo nas proximidades da região central conforme Figura 8 e este fato pode ser visualizado na Figura 9, onde o CVZ não está centralizado.

Uma amostra de padrão não centralizada não apresenta uma depreciação quanto à qualidade do sinal, contudo, caso seja necessário aplicar uma janela de Hamming para atenuar os efeitos do truncamento durante a aquisição, por exemplo; isto ocasionaria uma amplificação dos ruídos (Figura 9).

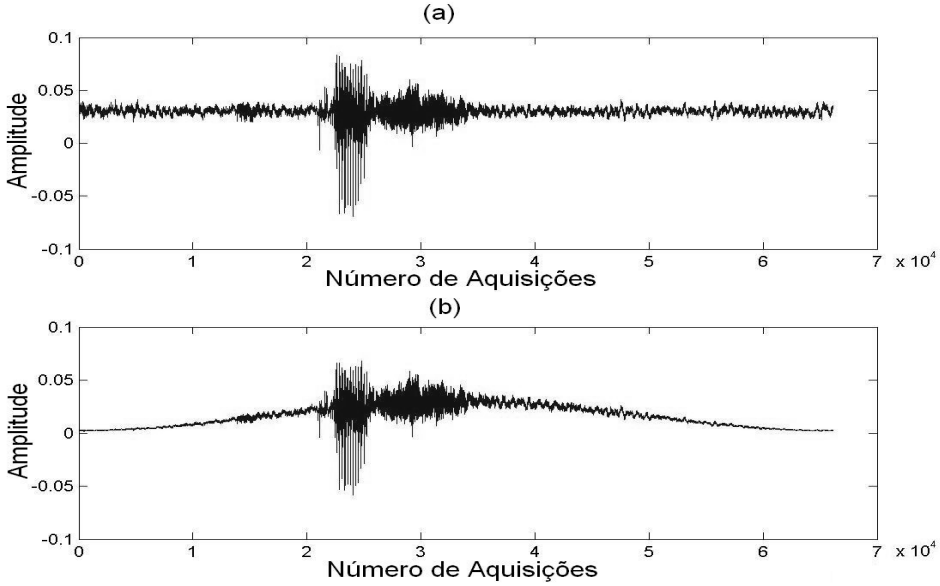


Figura 9 - Amplificação de ruído por aplicação indevido de janela de Hamming sobre uma amostra do padrão "atrás" (a) Sinal original adquirido (b) Sinal janelado

Para evitar problemas futuros como este, durante a aplicação de janelas, faz-se a centralização do CVZ tomando como base o valor absoluto de maior valor energético. Após encontrar este valor, verifica-se à distância entre o início da amostra e o ponto de pico e a distância entre o pico e o fim. A menor distância é adotada e espelhada sobre o maior de modo que o pico de energia esteja no centro da amostra (Figura 10) e conseqüentemente há uma redução no tamanho da zona de silêncio.

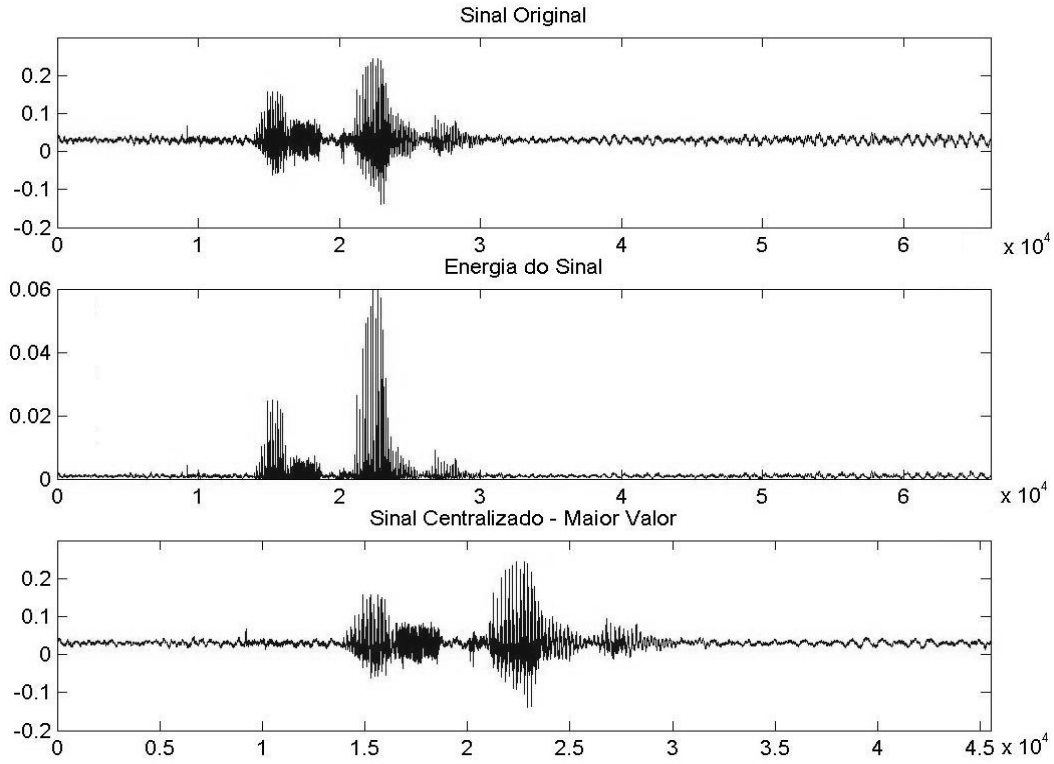


Figura 10 - Pronúncia do padrão "esquerda", análise de energia do sinal e centralização do sinal com base no bico de energia.

Filtragem do Sinal de Voz

A determinação dos limites de um CVZ utiliza basicamente a técnica da análise de energia, e em alguns casos a técnica da taxa de cruzando por zero do sinal no tempo **[7]**. Neste trabalho é implementada a primeira técnica. Contudo antes de separar a zona de silêncio do sinal relevante de voz é necessário trabalhar sobre o sinal adquirido e tentar minimizar os ruídos inerentes do processo de aquisição, filtrar as freqüências indevidas e normalizar a amplitude do sinal **[8]**.

Primeiramente, aplica-se um filtro digital do tipo FIR (*Finite Impulse Response*), denominado filtro de pré-ênfase, que apresenta a seguinte transformada Z:

$$f(z) = 1 - \alpha . z^{-1} \qquad \text{eq. (20)}$$

Implementado *f(z)* através da diferenciação a seguir:

$$y(n) = x(n) - \alpha.x(n-1) \qquad \text{eq. (21)}$$

Onde *x(n)* é o sinal amostrado e o parâmetro α pode variar entre 0,9 e 1,0 para sinais sonoros, e ter valores próximos de zero, para sinais surdos **[12]**. Neste trabalho adotou-se o valor de 0,95, pois sinais de voz apresentam características sonoras. Esta diferenciação age como um filtro passa alta, com o objetivo de compensar a atenuação de 6dB/oitava nas altas freqüências, devido à radiação da fala nos lábios de (+6dB/oitava) **[9]**. O filtro de pré-ênfase atua como um atenuador da tensão de *off-set* inserida no sinal de voz no momento da aquisição.

Com a finalidade de atenuar parte do ruído branco inerente no processo de aquisição, foi utilizado um filtro média móvel de terceira ordem, composto simplesmente zeros, conseqüentemente, estáveis, dado pela equação (25):

$$y(n) = \frac{1}{3}\sum_{k=0}^{2} x(n-k) \qquad \text{eq. (22)}$$

Conforme foi descrito anteriormente, as freqüências formantes dos sinais de voz correspondem à faixa de 100Hz a 5kHz. Com vista nestas freqüências, foram implementados filtros para desprezar as freqüências que se encontrassem fora desta região.

De acordo com a equação (4), um filtro analógico passa faixa pode ser implementado simplesmente ajustando os parâmetros *a* e *b*, de acordo com a equação (7). No entanto a função de transferência de um filtro passa faixa analógico no plano s é dada por:

$$H(s) = \frac{4900s}{s^2 + 4900s + 500000} \qquad \text{eq. (23)}$$

Onde os parâmetros encontrados para *a* e *b* são:

$a = 4900$
$b = 500000$

Adotou-se $K = a$ com o objetivo de manter o módulo máximo igual à unidade. Tendo em vista que uma amostra de CVZ é discreta, torna-se necessário converter o filtro descrito acima para o domínio discreto, para isto utiliza-se à transformação bilinear (equação (3)) que resultará na seguinte função de transferência:

$$H(z) = \frac{0.3278 - 0.3278z^{-2}}{1 - 1.3311z^{-1} + 0.3445z^{-2}} \qquad \text{eq. (24)}$$

Implementou-se dois filtros digitais destinados a executar funções similares ao filtro IIR resultante da transformação linear. O primeiro é um do tipo FIR passa baixa de 20ª ordem com freqüência de corte ω_c de 5kHz, ou $5/11 rad/s$ na freqüência de Nyquist. O segundo é um Butterworth passa faixa de 5ª ordem com freqüências de corte definidas pela faixa das freqüências formantes.

Após a etapa de filtragem fez-se a normalização do CVZ visando enquadrá-lo entre os limites unitários $|x(n)| \geq 1$. Um método de normalização da amostra seria encontrar o maior valor em módulo do sinal e, posteriormente, efetuar a divisão de todos elementos do vetor por este valor. A equação (25) descreve esta operação:

$$y(n) = \frac{x(n)}{\max\|x(n)\|} \qquad \text{eq. (25)}$$

Outro método utilizado foi encontrar o maior valor positivo e negativo e levá-los a unidade atentando para o sinal correspondente (equação (26)). Porém esta última apresenta a desvantagem de deslocar o zero caso os máximos não apresentem simetria de valor com a origem (distintos em módulo), o que não ocorre no primeiro caso.

$$y(n) = 2\frac{x(n) - \min(x(n))}{\max(x(n)) - \min(x(n))} - 1 \qquad \text{eq. (26)}$$

Ambas operações de normalização podem ser observadas na Figura 11.

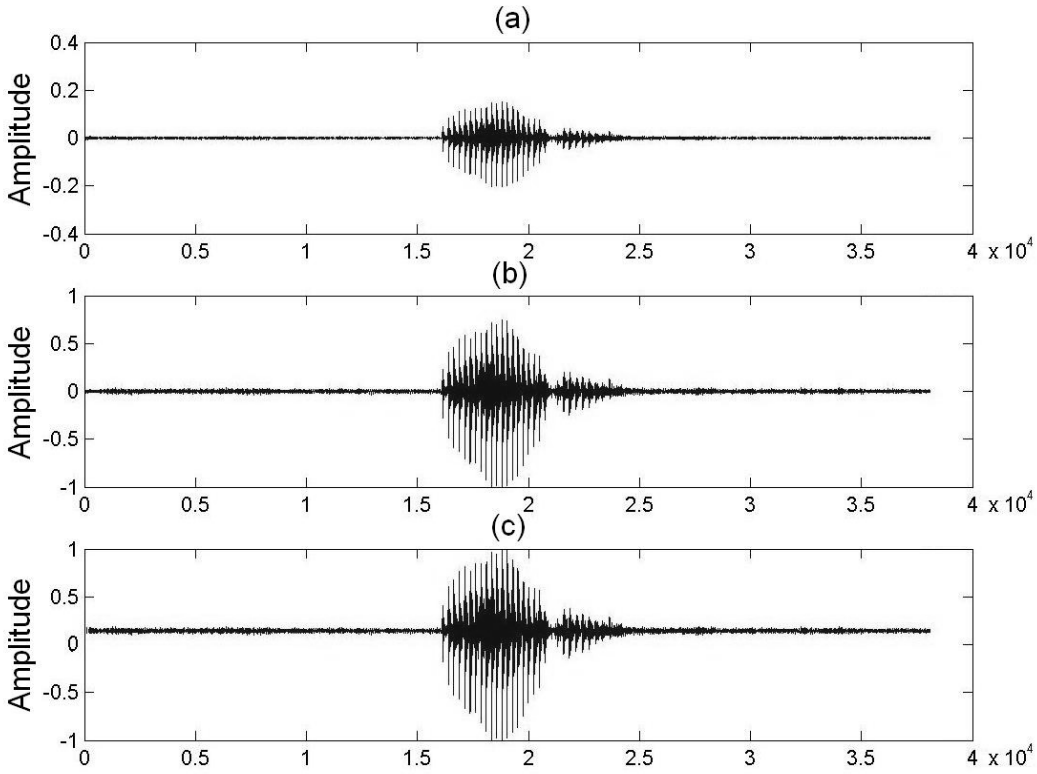


Figura 11 - (a) Sinal de voz filtrado e não normalizado (b) Sinal de voz normalizado tomando como base o maior valor em módulo do sinal (c) Sinal de voz normalizado tomando como referência o máximo valor negativo e positivo do sinal.

Análise de Energia

A análise de energia foi apresentada anteriormente com o objetivo de centralizar o CVZ a partir do pico máximo de energia do sinal e neste momento terá como finalidade encontrar os limites das palavras.

Encontrar os limites precisos de início e fim de um CVZ é um problema difícil de ser resolvido **[16]**, todavia para um sinal normalizado, a detecção dos limites pela análise de energia se torna mais simples, pois possibilita estipular patamares fixos. Para o caso em que o nível de ruído de fundo é de baixa intensidade, a tarefa de encontrar os limites é trivial, pois basta determinar um patamar de energia acima da energia do ruído e comparar durante o comando de voz. Contudo, pode ocorrer que em certos casos imprevisíveis um nível de ruído ultrapasse este patamar, sendo necessário, portanto, determinar outro patamar de maior intensidade com a finalidade de indicar que a energia do sinal está se elevando. Outro artifício a ser utilizado é dar saltos (incrementos

em relação ao primeiro ponto) no vetor que contenha o CVZ no momento em que o algoritmo de busca encontra um valor que ultrapasse o primeiro patamar. O salto tem a finalidade de evitar ruídos pontuais, os quais são desconsideráveis.

O início de uma palavra é definido no primeiro ponto que ultrapassa o primeiro patamar de energia e tem, após o salto, um seu sucessor com uma energia superior ao segundo patamar estabelecido. A busca pelo fim do CVZ ocorre de modo semelhante, porém ocorre do último para o primeiro elemento do vetor. A Figura 12 exemplifica as etapas para a detecção de início e fim de um CVZ.

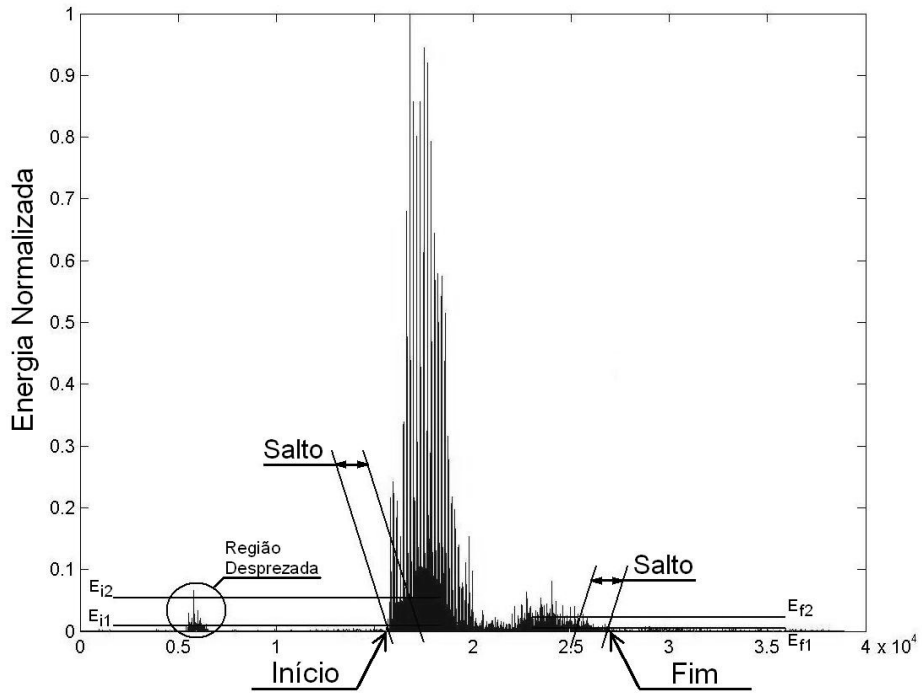


Figura 12 - Detecção de início e fim utilizando o método de energia do sinal. E_{i1}, E_{i2}, E_{f1} e E_{f2} são os patamares predefinidos de energia, bem como a largura do salto.

Neste trabalho foi adotado um salto equivalente a 20ms, pois para um sinal de voz não há tanta alteração na amplitude como ocorre nos casos de ruídos de impacto, conforme acontece na região desprezada na Figura 12. Os patamares de energia E_{i1} e E_{f1} são denominados patamares de silêncio de início e de fim da pronúncia, respectivamente. Os percentuais adotados para estes patamares são 1% para o início e 0,2% para o fim em relação à amplitude unitária da energia do CVZ. Os patamares E_{i2} e E_{f2} são destinados à confirmação de que a energia do comando de voz ainda cresce

mesmo após o salto, tanto do início para o fim da amostra quanto do fim para o início. Os valores adotados foram 2,5% e 0,5% da energia máxima do CVZ.

Após encontrar os limites do CVZ define-se um novo vetor, desprezando o que o algoritmo interpretou como zona de silêncio ou ruído de fundo. Este novo vetor apresenta um menor volume de dados com as mesmas informações de interesse. Após o corte na amostra original, aplica-se uma janela de Hamming para atenuar a inserção de freqüência nas bordas devido ao novo truncamento do sinal e realiza-se uma nova normalização da amostra.

4.3 Redes Neurais

Após a etapa de processamento do CVZ é suposto que o mesmo apresente simplesmente o que é relevante para o reconhecimento. Todavia, mesmo excluindo grande parte da zona de silêncio, a amostra de CVZ ainda contém milhares de elementos no vetor de dados sendo, portanto inviável criar uma estrutura neural com esta infinidade de entradas.

Os Coeficientes de Predição Linear (LPC), por sua vez, são aqui utilizados para extrair as características do CVZ pré-processado e representar os milhares de dados em algumas unidades ou dezenas pré-determinadas. A idéia da predição linear é que uma amostra de voz pode ser representada por uma combinação linear de amostras de voz passadas **[15][2]**. Neste trabalho foram utilizando LPC de 8^{a}, 10^{a} e 12^{a} ordem. Após a determinação dos LPC, estes foram normalizados pela equação (26) com o objetivo de proporcionar uma convergência mais rápida na etapa de treinamento das RNA.

Sabendo quantos parâmetros de entrada serão utilizados, a camada de entrada na RNA já está definida. O número de neurônios da camada oculta foi pré-estabelecido em 5, 10, 15 e 20, para que fosse possível analisar dentre as possíveis combinações a que apresentaria a melhor resposta. A camada de saída apresenta simplesmente um neurônio cuja resposta é positiva para o padrão a ser reconhecido ou negativa para a rejeição dos demais padrões. Para cada padrão foi treinada uma RNA específica com uma taxa de

aprendizagem do algoritmo back-propagation de 0,010. Cada estrutura de redes foi avaliada cinco vezes e salvando aquele que obtivesse menor erro dentre as repetições.

Foram utilizadas funções sigmóides tangente hiperbólico dada pela equação 27 para as funções de ativação dos neurônios da camada intermediária e de saída.

$$fa(n) = \frac{1 - e^{-b.n}}{1 + e^{-b.n}} \qquad \text{eq. (27)}$$

Para treinamento do sistema foram utilizados os algoritmos back-propagation de Levenberg-Marquardt (LM) e de Regularização Bayesiana (BR) para efeito de comparação. O algoritmo LM apresenta como vantagem principal à agilidade na qual converge para uma solução quando o problema de reconhecimento de padrões seja linear. A principal desvantagem deste algoritmo é o intenso consumo de memória computacional. O algoritmo LM é baseado no método de Newton, o qual utiliza o Jacobiano da função erro para ajuste dos pesos durante uma época. O grande esforço computacional se concentra no cálculo da matriz Hessiana (segunda derivada) a partir do Jacobiano **[14]**. O algoritmo BR é uma modificação do algoritmo LM com as vantagens da regularização citadas anteriormente.

As atualizações dos pesos foram realizadas em até 200 ciclos ou então quando o erro quadrático médio (mse) atingisse um valor inferior a 0,012.

As amostras de cada padrão foram separadas aleatoriamente constando com a razão de 2/3 das amostras destinadas ao treinamento da RNA e 1/3 para validação e teste da estrutura treinada. O treinamento foi realizado separadamente para cada padrão. Para as RNA específicas foram apresentados 40 amostras de CVZ considerados verdadeiros perante a estrutura, como também foram apresentados 160 amostras de CVZ falsos.

5 Resultados e Discussões

A definição do tipo de filtro a ser utilizado dentre os citados anteriormente para a determinação do início e fim do CVZ foi escolhida analisando a energia normalizada da resposta destes. A Figura 13 ilustra estas respostas.

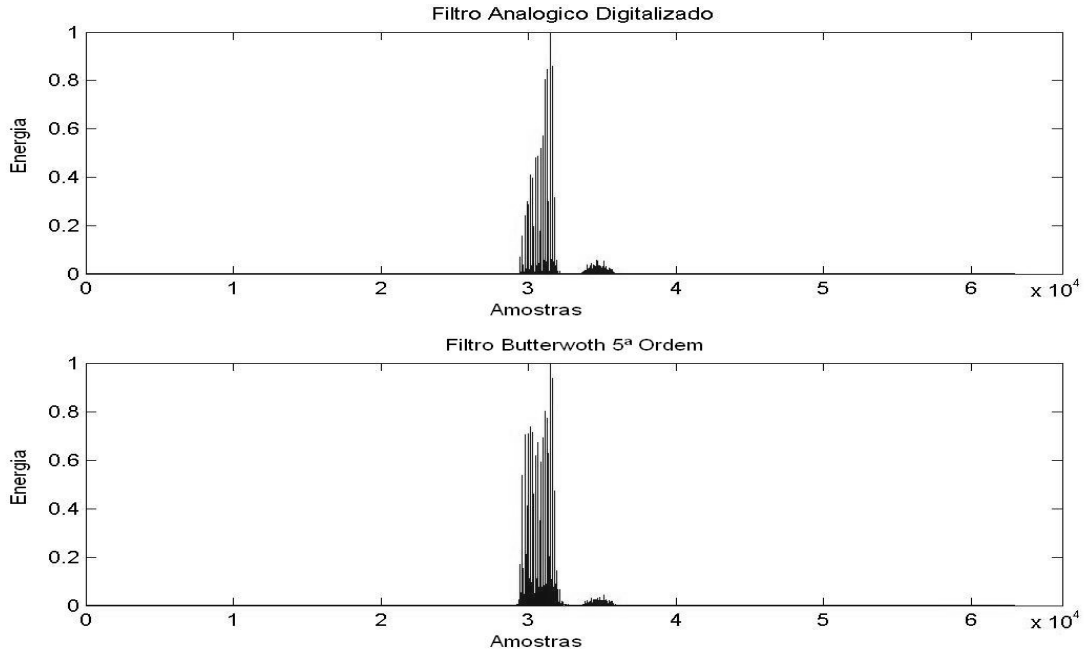


Figura 13 - Análise de energia do Filtro Bilinear e Butterworth

Como é possível visualizar, a energia do filtro analógico digitalizado pela transformação bilinear é mais bem definida na fase final do comando de voz. Visualmente é uma diferença sutil, todavia para efeitos computacionais a determinação do ponto de término do comando se torna uma tarefa mais simples.

Todavia a qualidade sonora apresentada pelo filtro Butterworth é bem superior a do filtro supracitado, logo o que se fez foi encontrar os elementos que determinam os limites na resposta do filtro analógico e procurá-los na resposta do filtro Butterworth, para que o corte da amostra pudesse ser efetuado.

Foram testados também modelos híbridos entre o filtro Butterworth e o da transformação linear, os quais não obtiveram bons resultados quanto à definição clara dos limites das palavras.

Foi feita uma comparação entre o filtro FIR de 20ª ordem e o filtro Butterworth de 5ª ordem, ambos obtiveram uma resposta aceitável, todavia o Butterworth computa a resposta com esforço computacional muito menor que o FIR.

Para a avaliação do processamento do sinal de voz, foi verificado o espectro de freqüência da amostra antes e após a filtragem e a normalização (Figura 14) através da transformada rápida de Fourier (FFT – *Fast Fourier Transform*). É fácil observar que a amostra de voz pós-processada não apresenta amplitudes significativas na região inferior a 125Hz e superior a 5kHz, o que caracteriza a funcionalidade dos filtros digitais aplicados aos CVZ.

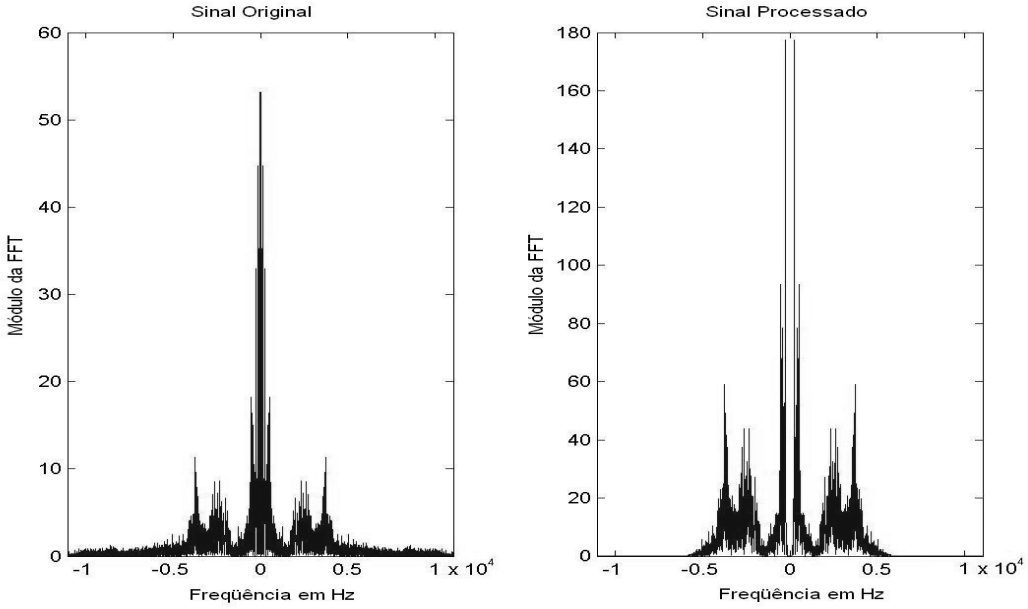


Figura 14 - Análise de Fourier para o sinal original e processado

A Figura 15 destaca a discrepância existente entre o sinal original e o sinal resultante após a aplicação dos filtros diferencial, média móvel, Butterworth e janela de Hamming.

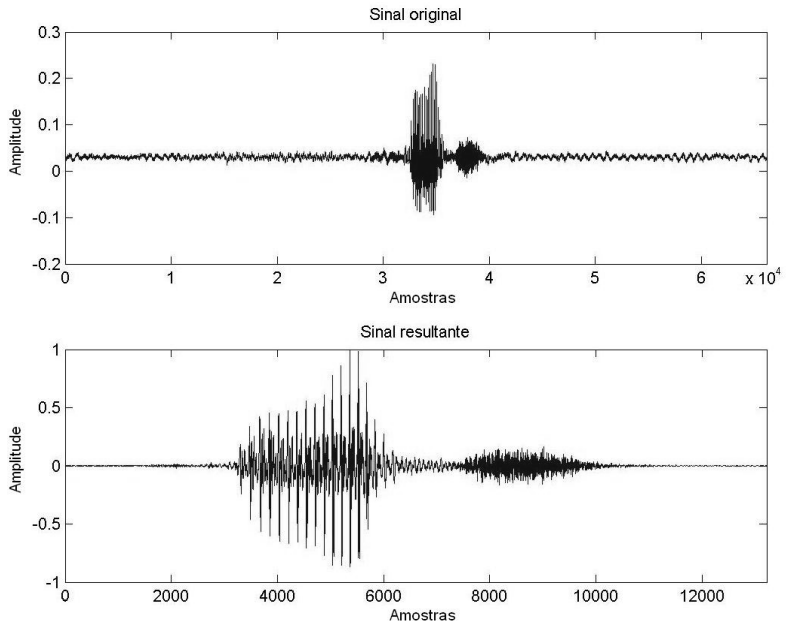


Figura 15 - Comparação entre a amostra do sinal original e do processado

Como é possível verificar na Figura 15, a amostra de CVZ pós-processada apresenta uma amplitude bem definida como também uma atenuação dos ruídos e eliminação da tensão de *off-set*.

Realizando as operações de processamento de sinais e de detecção de início e fim do CVZ foi possível obter uma redução de 73,6% em média no número de dados do vetor que continha a amostra, tendo, por isso, uma redução de esforço computacional.

Quanto ao desempenho das RNA treinadas foram avaliados dois aspectos destinados à qualificação do sistema:

a) Erro: é o percentual da amostras que foram identificadas para um dado padrão, mas que na verdade pertencem à outra classe;
b) Equívoco: é o percentual de amostras de um dado padrão que foi indevidamente rejeitado, sendo, portanto considerado de outra classe.

As Tabelas de 2 a 7 demonstram o percentual de erro e equívoco encontrados tanto na etapa de treinamento quanto na etapa de validação e testes. As linhas indicam a função de treinamento predefinida, o número de LPC extraídos, o número de neurônios

da camada oculta (nnco) e a resposta final da rede. A resposta final (rf) é determinada pela porcentagem de todas as amostras que a rede ou equivocou ou errou.

Tabela 2 - RNA treinada pelo algoritmo BR para 8 LPC

algoritmo	NNCO	posição	treinamento		validação		RF
			erro	equívoco	erro	equívoco	
Regularização Bayesiana 8 coeficientes LPC	5	frente	0,00	0,00	5,00	5,00	1,67
		atrás	0,00	0,00	0,00	6,25	1,67
		direita	7,50	0,00	15,00	2,50	2,67
		esquerda	10,00	0,00	15,00	10,00	5,00
		para	0,00	0,62	0,00	8,75	2,67
	10	frente	0,00	0,00	10,00	5,00	2,00
		atrás	0,00	0,00	0,00	5,00	1,33
		direita	5,00	0,00	15,00	7,50	3,67
		esquerda	0,00	0,00	25,00	11,25	4,67
		para	0,00	0,62	15,00	7,50	3,33
	15	**frente**	**0,00**	**0,00**	**10,00**	**3,75**	**1,67**
		atrás	**0,00**	**0,00**	**5,00**	**1,25**	**0,67**
		direita	2,50	0,00	20,00	8,75	4,00
		esquerda	**0,00**	**0,00**	**15,00**	**5,00**	**2,33**
		para	0,00	0,62	15,00	3,75	2,33
	20	frente	0,00	0,00	15,00	5,00	2,33
		atrás	0,00	1,88	5,00	6,25	3,00
		direita	**7,50**	**0,00**	**10,00**	**3,75**	**2,67**
		esquerda	2,50	0,00	20,00	8,75	4,00
		para	**0,00**	**0,62**	**5,00**	**3,75**	**1,67**

Tabela 3 – RNA treinada pelo algoritmo BR para 10 LPC

algoritmo	NNCO	posição	treinamento		validação		RF
			erro	equívoco	erro	equívoco	
Regularização Bayesiana 10 coeficientes LPC	5	**frente**	**0,00**	**0,62**	**30,00**	**1,25**	**2,67**
		atrás	0,00	0,00	5,00	6,25	2,00
		direita	**2,50**	**0,00**	**20,00**	**3,75**	**2,67**
		esquerda	**2,50**	**0,00**	**25,00**	**5,00**	**3,33**
		para	**0,00**	**0,00**	**5,00**	**2,50**	**1,00**
	10	frente	5,00	1,88	20,00	6,25	4,67
		atrás	**0,00**	**0,00**	**15,00**	**2,50**	**1,67**
		direita	10,00	0,00	35,00	11,25	6,67
		esquerda	7,50	0,00	35,00	7,50	5,33
		para	0,00	0,00	20,00	2,50	2,00
	15	frente	2,50	1,88	15,00	6,25	4,00
		atrás	2,50	0,00	10,00	3,75	2,00
		direita	5,00	0,62	35,00	2,50	4,00
		esquerda	20,00	1,88	20,00	3,75	6,00
		para	5,00	0,00	0,00	3,75	1,67
	20	frente	7,50	1,88	20,00	7,50	5,33
		atrás	5,00	0,00	0,00	7,50	2,67
		direita	5,00	0,00	25,00	5,00	3,67
		esquerda	5,00	0,00	30,00	11,25	5,67
		para	10,00	0,00	35,00	3,75	4,67

Tabela 4 – RNA treinada pelo algoritmo BR para 12 LPC

algoritmo	NNCO	posição	treinamento		validação		RF
			erro	equívoco	erro	equívoco	
Regularização Bayesiana 12 coeficientes LPC	5	**frente**	**0,00**	**0,00**	**20,00**	**3,75**	**2,33**
		atrás	0,00	0,62	5,00	6,25	2,33
		direita	**17,50**	**0,62**	**25,00**	**7,50**	**6,33**
		esquerda	**5,00**	**0,00**	**30,00**	**7,50**	**4,67**
		para	**0,00**	**0,00**	**10,00**	**1,25**	**1,00**
	10	frente	0,00	0,00	20,00	11,25	4,33
		atrás	2,50	0,00	10,00	1,25	1,33
		direita	7,50	1,25	30,00	23,75	10,00
		esquerda	7,50	0,00	30,00	6,25	4,67
		para	0,00	0,62	10,00	3,75	2,00
	15	frente	0,00	0,00	25,00	7,50	3,67
		atrás	0,00	0,00	25,00	2,50	2,33
		direita	5,00	0,00	30,00	22,50	8,67
		esquerda	2,50	0,00	25,00	11,25	5,00
		para	2,50	0,62	5,00	1,25	1,33
	20	frente	0,00	0,00	20,00	7,50	3,33
		atrás	**0,00**	**0,00**	**5,00**	**3,75**	**1,33**
		direita	5,00	0,62	55,00	17,50	9,33
		esquerda	7,50	0,00	40,00	6,25	5,33
		para	0,00	0,62	10,00	5,00	2,33

Tabela 5 – RNA treinada pelo algoritmo LM para 8 LPC

algoritmo	NNCO	posição	treinamento		validação		RF
			erro	equívoco	erro	equívoco	
Levenberg-Marquardt 8 coeficientes LPC	5	**frente**	**0,00**	**0,00**	**5,00**	**3,75**	**1,33**
		atrás	0,00	0,62	10,00	3,75	2,00
		direita	10,00	0,00	5,00	3,75	2,67
		esquerda	**2,50**	**0,00**	**30,00**	**3,75**	**3,33**
		para	0,00	0,62	10,00	2,50	1,67
	10	frente	2,50	0,00	10,00	2,50	1,67
		atrás	2,50	1,25	5,00	11,25	4,33
		direita	7,50	0,00	20,00	6,25	4,00
		esquerda	7,50	0,00	30,00	7,50	5,00
		para	**0,00**	**1,25**	**5,00**	**1,25**	**1,33**
	15	frente	0,00	0,00	20,00	2,50	2,00
		atrás	0,00	0,00	5,00	6,25	2,00
		direita	2,50	0,00	5,00	13,75	4,33
		esquerda	0,00	0,00	30,00	6,25	3,67
		para	0,00	0,00	10,00	5,00	2,00
	20	frente	0,00	0,00	10,00	6,25	2,33
		atrás	**0,00**	**0,00**	**0,00**	**7,50**	**2,00**
		direita	**5,00**	**0,00**	**10,00**	**2,50**	**2,00**
		esquerda	7,50	0,00	30,00	10,00	5,67
		para	0,00	1,25	10,00	2,50	2,00

Tabela 6 – RNA treinada pelo algoritmo LM para 10 LPC

algoritmo	NNCO	posição	treinamento		validação		RF
			erro	equívoco	erro	equívoco	
Levenberg-Marquardt 10 coeficientes LPC	5	frente	0,00	1,25	10,00	10,00	4,00
		atrás	0,00	0,62	15,00	2,50	2,00
		direita	**5,00**	**0,00**	**15,00**	**2,50**	**2,33**
		esquerda	15,00	0,00	30,00	5,00	5,33
		para	0,00	0,62	0,00	6,25	2,00
	10	frente	2,50	1,25	10,00	6,25	3,33
		atrás	2,50	0,62	5,00	2,50	1,67
		direita	2,50	0,00	20,00	3,75	2,67
		esquerda	**0,00**	**0,00**	**15,00**	**10,00**	**3,67**
		para	0,00	0,00	10,00	2,50	1,33
	15	frente	0,00	0,62	30,00	7,50	4,33
		atrás	**2,50**	**0,00**	**0,00**	**3,75**	**1,33**
		direita	2,50	0,00	30,00	13,75	6,00
		esquerda	7,50	0,00	20,00	16,25	6,67
		para	2,50	0,00	15,00	1,25	1,67
	20	**frente**	**0,00**	**0,62**	**15,00**	**3,75**	**2,33**
		atrás	2,50	0,00	25,00	2,50	2,67
		direita	7,50	1,25	25,00	3,75	4,33
		esquerda	5,00	0,62	20,00	12,50	5,67
		para	**2,50**	**0,00**	**10,00**	**0,00**	**1,00**

Tabela 7 – RNA treinada pelo algoritmo LM para 12 LPC

algoritmo	NNCO	posição	Treinamento		validação		RF
			erro	equívoco	erro	equívoco	
Levenberg-Marquardt 12 coeficientes LPC	5	frente	2,50	0,00	20,00	6,25	3,33
		atrás	0,00	0,00	5,00	3,75	1,33
		direita	**7,50**	**0,00**	**5,00**	**1,25**	**1,67**
		esquerda	**7,50**	**0,00**	**15,00**	**3,75**	**3,00**
		para	**0,00**	**0,00**	**10,00**	**0,00**	**0,67**
	10	**frente**	**5,00**	**0,00**	**10,00**	**2,50**	**2,00**
		atrás	0,00	0,00	5,00	5,00	1,67
		direita	7,50	1,88	40,00	11,25	7,67
		esquerda	10,00	0,62	35,00	7,50	6,00
		para	0,00	0,62	5,00	3,75	1,67
	15	frente	0,00	0,00	25,00	3,75	2,67
		atrás	0,00	0,00	5,00	5,00	1,67
		direita	10,00	0,00	40,00	20,00	9,33
		esquerda	0,00	0,00	40,00	10,00	5,33
		para	2,50	0,00	5,00	5,00	2,00
	20	frente	0,00	0,00	15,00	18,75	6,00
		atrás	**0,00**	**0,00**	**0,00**	**1,25**	**0,33**
		direita	0,00	0,00	35,00	10,00	5,00
		esquerda	5,00	0,00	35,00	7,50	5,00
		para	0,00	0,00	30,00	5,00	3,33

As linhas em negrito indicam a estrutura com RF mais baixa para cada padrão em função do algoritmo de treinamento e do número de LPC extraídos, variando simplesmente o NNCO.

As Tabelas mostraram que alguns padrões obtiveram um baixo índice de erro ou equívoco como, por exemplo, pára e atrás.

A Tabela 8 a seguir explicita um resumo dos resultados tomando como base o algoritmo de treinamento das estruturas neurais.

Tabela 8 – Resumo dos resultados alcançado nos treinamentos dos padrões pelos algoritmos de Levenberg - Marquard e Regularização Bayesiana

Algoritmo	Levenberg - Marquard		Regularização Bayesiana	
Padrão	Erro da Rede	Estrutura	Erro da Rede	Estrutura
Frente	1,33%	[8 – 5 – 1]	1,67%	[8 – 15 – 1]
Atrás	0,33%	[12 – 20 – 1]	0,67%	[8 – 15 – 1]
Direita	1,67%	[12 – 5 – 1]	2,67%	[10 – 5 – 1]
Esquerda	3,00%	[12 – 5 – 1]	2,33%	[8 – 15 – 1]
Pára	0,67%	[12 – 5 – 1]	1,00%	[10 – 5 – 1]

Segundo Adami [2], um sistema RAV deve ter um erro da rede inferior a 2,50% e conforme explicitado na Tabela 8, os padrões esquerda e direita resultantes do algoritmo de treinamento de Levenberg - Maquard e Regularização Bayesiana, respectivamente, extrapolam este valor.

Foi possível verificar durante teste que à medida que cresce o número de parâmetros efetivos da RNA não há uma melhora tão significativa, como ocorre nos algoritmos de treinamento LM e BR para 12 LPC e 20 neurônios na camada intermediária. Logo, para números superiores de parâmetros não haverá redução de erro ou equívoco, provavelmente.

Foram testadas RNA com duas camadas intermediárias, as quais não apresentaram respostas satisfatórias, além de serem muito mais complexas matematicamente e difíceis de treinar.

6 Conclusões

Primeiramente, a redução da quantidade de elementos da amostra de CVZ devido ao processamento agilizou o processo de extração de características através dos coeficientes de predição linear (LPC).

Outro fato observado foi que nem sempre o aumento da quantidade de variáveis do sistema de reconhecimento, como por exemplo, o número de neurônios da camada intermediária ou o número de coeficientes de predição extraídos, reduzem os erros inerentes ao processo de classificação de padrões. Há sempre um ponto intermediário onde a resposta mais favorável se encontra, todavia a busca por este é um processo empírico.

Uma melhoria do sistema RAV com redução de erro e equívoco poderia ser conseguida se houvesse uma quantidade maior de amostras e um nível menor de ruídos de fundos. Uma quantidade maior de indivíduos auxiliaria na diversidade do sistema e um nível mais baixo de ruído contibuiria na detecção dos limites do CVZ e na extração das características.

Este trabalho alcançou o objetivo de contribuir para o aperfeiçoamento de sistemas de reconhecimento automático de voz e estimular a implementação real do sistema para atuar em equipamentos ou como uma interface auxiliar para computadores.

Para trabalhos futuros, com os conhecimentos aqui adquiridos, é interessante verificar, por exemplo, se existem outros métodos mais eficientes que coeficientes de predição para extração de características, como por exemplo, coeficientes espectrais. E outras técnicas diferentes de redes neurais artificiais para reconhecimento de padrões, tais como Modelos Ocultos de Markov.

7 Referências Bibliográficas

[01] Moreira, F. (1998). Reconhecimento automático de fala contínua. Trabalho de Conclusão de Curso. Engenharia Elétrica AAPS. Portugal – 1998. 11

[02] Adami, A.G. (1997). Sistemas de reconhecimento de locutor utilizando Redes Neurais Artificiais. Tese de Mestrado. Ciência da Computação - Universidade Federal do Rio Grande do Sul – 1997. .... 11, 17, 37

[03] Furui, S. (1989). Digital speech processing, synthesis and recognitions. Marcel Dekker, Inc – 1989. 12

[04] Lee, K. F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continous speech recognition. IEEE transactions on acoustics speech and signal processing. April 1990**.** 13

[05] Ynoguti, C. A. (1999). Reconhecimento de fala contínua usando modelos ocultos de Markov. Tese de Doutorado. UNICAMP, 1999. 13

[06] Cox, R.V., Rabiner, L.R. (1999). Digital signal processing handbook, speech processing.Chapman & Hall – 1999. 13

[07] Rabiner, L. R. (1974). An Algorithm for Locating the Beginning and End of an Utterance Using ADPCM Coded. Speech, L. H. Rosenthal, R. W. Schafer and L. R. Rabiner, *Bell System Tech. Journ.,* Vol. 53, No. 6, pp. 1127-1135, July-August 1974 32

[08] Ribas, J. C., Cunha, F. L., Cliquet Jr, A. (2002). Sistema de Controle por Voz Aplicado à Reabilitação Humana. XVIII Congresso Brasileiro de Engenharia Biomédica, Vol. 1, pp 149-154, São José dos Campos, SP, Brasil – 2002.. 32

[09] Dias, R. S. F. (2000). Normalização de locutor em sistema de reconhecimento de fala. Tese de Mestrado, UNICAMP – 2000. 33

[10] The MathWorks Inc. (2000). Signal Processing Toolbox User's Guide for Use with Matlab – 2000.. 16

[11] Johnson, D. E., Hilburn, J. L., Johnson, J. R. (1994). Fundamentos de Análise de Circuitos Elétricos. Ed LTC - 2000. 15

[12] Sotomayor, C. A. M. (2003). Realce de Voz Aplicado à Verificação Automática de Locutor. Tese de Mestrado, IME – 2003. 33

[13] The Mathworks, Inc. (2000). Neural Network Toolbox, User's Guide for Use with Matlab – 2000.. 26

[14] Cunha, A. R., Racz, A., da Silva, V. F. (2002). Sistema de Reconhecimento de Escrita Baseado em Redes Neurais Artificiais. Projeto Final de Curso da USP – 2002. 38

[15] Tanprasert, C., Wutiwiwatchai, C., Sae-Tang, S. (1999). Text-dependent Speaker Identification Using Neural Network on Distinctive Thai Tone Marks. Internacional joint Conference on Neural Network, July, 1999. 37

[16] Scavone, A. P. R. (1996). Reconhecimento de Palavrras por Modelos Ocultos de Markov. Tese de Mestrado, USP - 1996. 35

[17] Diniz, P. S. R., da Silva, E. A. B., Netto, S. L. (2004). Processamento Digital de Sinais, Projeto e Análise de Sistemas. Ed. Bookman - 2004. 15

[18] Haykin, S., Veen, B. V. (2001). Sinais e Sistemas. Ed Bookman - 2002. 15

[19] Martins, J. A. (1997). Avaliação de Diferentes Técnicas para Reconhecimento de Fala. Tese de Doutorada, UNICAMP - 1997. 14

[20] Braga, A. P., Carvalho, A. C. P. L. F., Ludermir, T. B. (2000). Redes Neurais Artificiais, Teoria e Aplicações. Ed LTC - 2000. 20

[21] Haykin, S. (2001). Redes Neurais, Princípios e Prática. Ed. Bookman - 2001. 20

[22] Página na internet acecssada no dia 27 de fevereiro de 2005: http://www.din.uem.br/ia/neurais/#neural 20

Apêndice A – Exemplo de Aplicação

A.1 Desenvolvimento

Um protótipo foi desenvolvido durante a execução deste trabalho para validar o sistema de Reconhecimento Automático de Voz (RAV). Este tem como objetivo agir no meio externo ao microcomputador segundo os Comandos de Voz (CVZ) previamente treinados.

Os comandos selecionados para a confecção do banco de dados e utilizados para treinar, validar e testar as estruturas neurais foram frente, atrás, direita, esquerda e pára. Estes comandos almejam o controle do dispositivo de posicionamento de um carro guiado por voz. O protótipo do carro apresenta duas rodas que atuam de forma independente.

O princípio de funcionamento do sistema RAV é descrito segundo o diagrama da Figura 16.

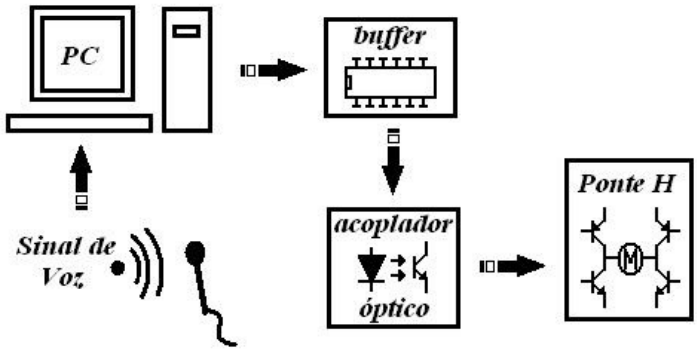


Figura 16 - Diagrama de Funcionamento do Sistema RAV em Execução no Protótipo

Primeiramente é realizada a aquisição do CVZ o qual é processado, filtrado e, posteriormente, classificado segundo as RNA treinadas. A RNA que for excitada pelo CVZ pós-processado será a responsável pelo envio do código referente à ação a ser executada. Foram estabelecidos conforme apresentados na Tabela 1 os códigos referentes às ações.

Tabela 9 - Codificação das ações de acordo com o CVZ pronunciado

Comandos	Motor Direito		Motor Esquerdo		Codificação (4 bits)
	Rotação Horária	Rotação Anti-horária	Rotação Horária	Rotação Anti-horária	
Frente	Ligado	Desligado	Desligado	Ligado	1001
Atrás	Desligado	Ligado	Ligado	Desligado	0110
Direita	Desligado	Desligado	Desligado	Ligado	0001
Esquerda	Ligado	Desligado	Desligado	Desligado	1000
Pára	Desligado	Desligado	Desligado	Desligado	0000

Caso duas ou mais RNA tenham suas saídas excitadas simultaneamente o algoritmo desenvolvido estabelece que a ação a ser executada é referente ao comando "pára" e a mensagem "Comando não identificado" é exibida ao usuário.

O código binário é enviado através da porta paralela do microcomputador (PC) para o meio externo. A fim de eliminar o acoplamento elétrico entre o circuito de comando e o circuito de carga (motor) foram utilizados um *buffer* e acopladores ópticos protegendo, pois, este dispositivo de saída do PC.

E por fim, foi montado o circuito de uma ponte H para cada motor de corrente contínua acoplado às rodas para atuar invertendo a rotação do mesmo e agir de acordo com os CVZ pronunciados por alguns dos locutores previamente treinados.

A.2 Resultados e Discussões

Para validação do sistema de simulação foi estabelecida uma Tabela de comandos para que fosse possível analisar o comportamento do sistema perante locutores previamente treinados pelas RNA e outros locutores desconhecidos por estas. As Tabelas 9 e 10 destacam a situação de reconhecimento ou não de CVZ para um locutor previamente treinado e outro desconhecido, respectivamente. Os comandos desejados estão selecionados com a cor cinza e os comandos pronunciados e classificados pelas RNA são denotados pela cor preta. Para os casos onde ocorra a sobreposição de cores há, portanto a evidência de que o comando foi identificado corretamente pelo sistema RAV. Para o caso em que numa mesma linha existam duas

células preenchidas há neste caso a situação de comando não identificado (equívoco) ou classificação de um comando em uma outra classe (erro).

Tabela 10 - Lista de comandos e os resultados de um locutor previamente treinado perante a RNA para a validação do sistema RAV

comando	frente	atrás	direita	esquerda	pára	não ind
frente						
frente						
frente						
atrás						
atrás						
atrás						
direita						
direita						
direita						
esquerda						
esquerda						
esquerda						
pára						
pára						
pára						
frente						
atrás						
direita						
esquerda						
pára						
pára						
esquerda						
direita						
atrás						
frente						
esquerda						
pára						
direita						
frente						
atrás						
esquerda						
direita						
pára						
atrás						
frente						
esquerda						
atrás						
direita						
pára						
frente						

Tabela 11 - Lista de comandos e os resultados de um locutor desconhecido perante a RNA para a validação do sistema RAV

comando	frente	atrás	direita	esquerda	pára	não ind
frente						
frente						
frente						
atrás						
atrás						
atrás						
direita						
direita						
direita						
esquerda						
esquerda						
esquerda						
pára						
pára						
pára						
frente						
atrás						
direita						
esquerda						
pára						
pára						
esquerda						
direita						
atrás						
frente						
esquerda						
pára						
direita						
frente						
atrás						
esquerda						
direita						
pára						
atrás						
frente						
esquerda						
atrás						
direita						
pára						
frente						

Nas Tabelas 9 e 10 é possível verificar a influência de um treinamento prévio do locutor, recordando que o sistema RAV desenvolvido é dependente de locutor. As discrepâncias existentes nas Tabelas destacam o fato de que locutores previamente treinados ao pronunciar erroneamente um comando, mais de uma RNA é excitada e o

sistema responde que o comando não foi identificado em 89,7% dos casos durante a validação do sistema. Para locutores desconhecidos pelas RNA este número reduz para 51,7%, existindo, por conseguinte, uma tendência maior de um locutor desconhecido pronunciar um CVZ, por exemplo, “frente” e o sistema RAV classificar como “direita”.

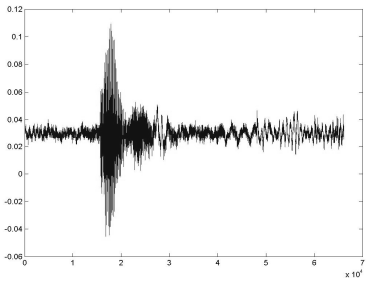
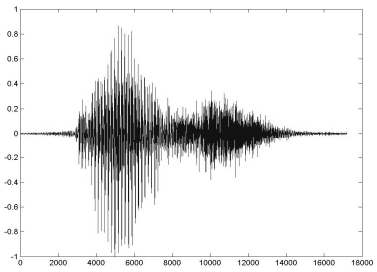
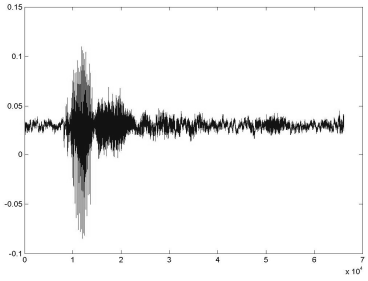
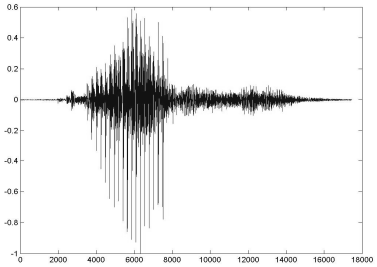
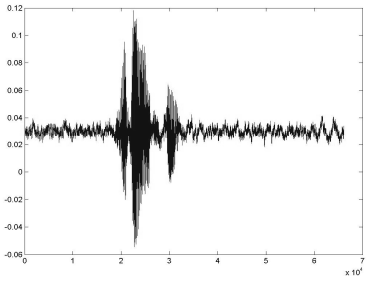
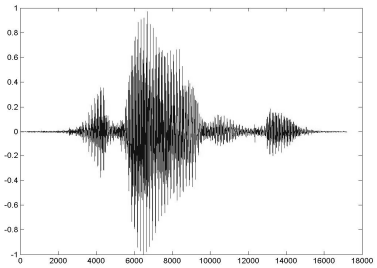
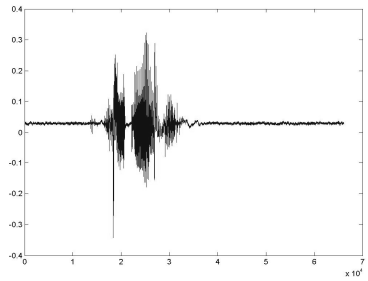
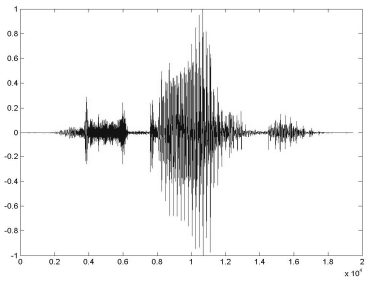
O sistema RAV desenvolvido como qualquer outro sistema tem o atraso de tempo inerente aos cálculos realizados para a determinação da saída classificada, neste trabalho o sistema apresentou um tempo de processamento em média de 4,74 segundos. Esta medida foi realizada a partir do momento em que o locutor inicia a pronuncia do comando de voz até o instante em que o motor responde ao comando dado ou o programa em execução exibe a mensagem de comando não identificado. É importante salientar que a parcela fixa de 3,00 segundos é destinada à aquisição do CVZ e o tempo restante, denominado tempo morto ou tempo de resposta, em média de 1,74 segundos é aquele no qual o sistema entra em processo de cálculo matemático.

A.3 Conclusão

Para um sistema RAV dependente de locutor torna-se evidente a relação entre os locutores previamente treinados e outros desconhecidos durante o processo de reconhecimento dos comandos de voz. Os locutores treinados se destacam até no momento de equívoco, onde mais de uma RNA é excitada e o sistema permanece estável ou repouso exigindo que um novo comando seja pronunciado em 89,7% dos casos. Este fato ocorre também para locutores desconhecidos, todavia em menores proporções, 51,7% dos casos, enfatizando o erro que resulta na operação indevida da saída do sistema, neste caso, a operação dos motores.

Apêndice B – Padrões de Voz Antes e Após Processamento Digital de Sinais

Tabela 12 – Comparação do comando de voz antes e após processamento digital de sinais

Padrão	Antes	Após
Frente		
Atrás		
Direita		
Esquerda		
Pára