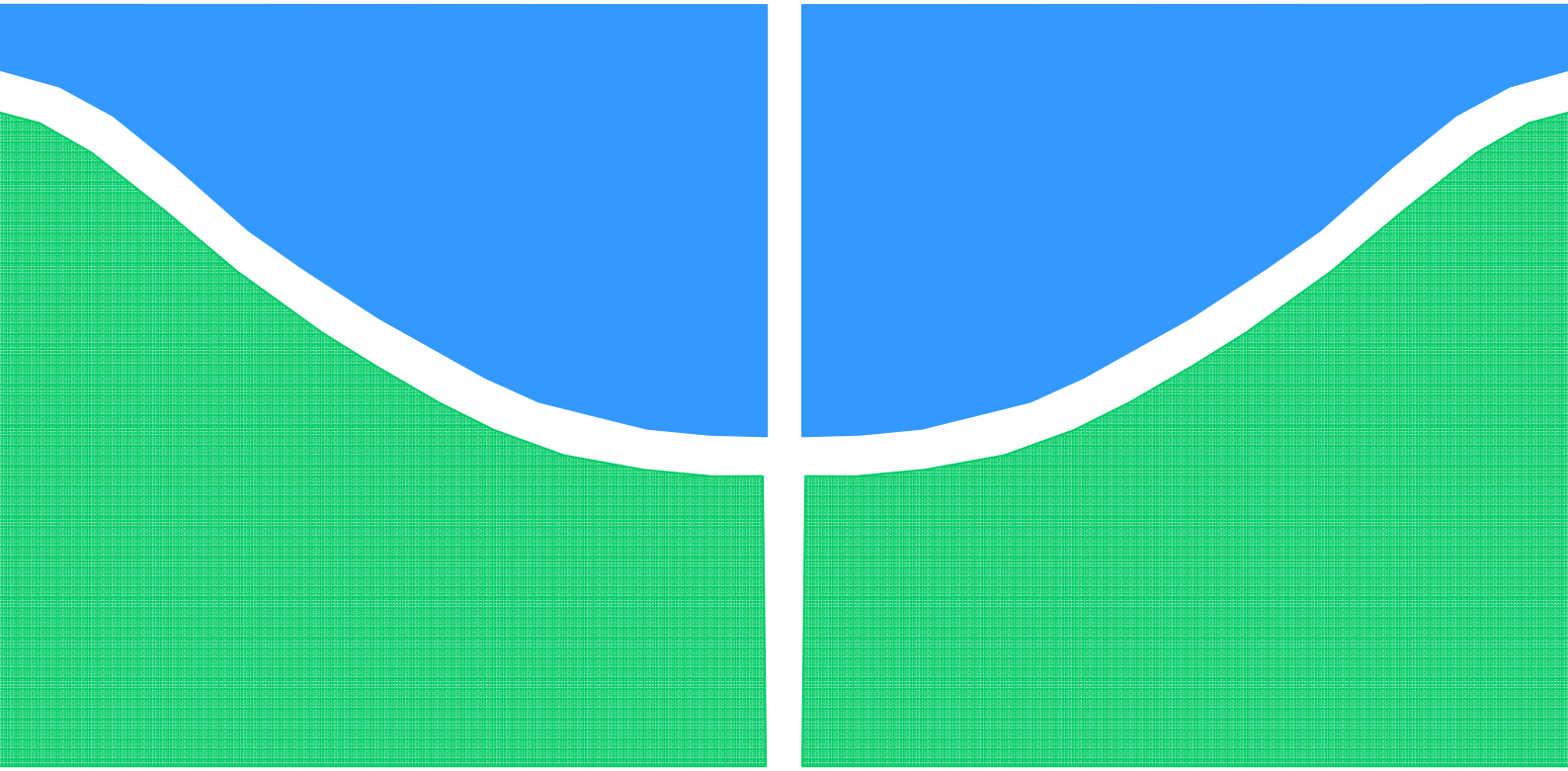

TRABALHO DE CONCLUSÃO DE CURSO

RECONHECIMENTO DE FONEMAS UTILIZANDO REDES NEURAIS PULSADAS

Alexandre Ungaretti Marcondes de Mello
Caio Nunes Nishiyama

Brasília, Setembro de 2010

UNIVERSIDADE DE BRASILIA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia

TRABALHO DE GRADUAÇÃO

RECONHECIMENTO DE FONEMAS UTILIZANDO REDES NEURAIS PULSADAS

Alexandre Ungaretti Marcondes de Mello
Caio Nunes Nishiyama

Relatório submetido como requisito parcial para obtenção
do grau de Engenheiro Eletricista

Banca Examinadora

Prof. Alexandre Ricardo Soares Romariz, Ph.D, UnB/ ENE (Orientador) _____

Prof. Janaína Gonçalves Guimarães, Doutora, UnB/ ENE _____

André Tomaz Gontijo, Mestre, UnB/ ENE _____

FICHA CATALOGRÁFICA

MELLO, ALEXANDRE UNGARETTI MARCONDES DE & NISHIYAMA, CAIO NUNES
Reconhecimento de Fonemas utilizando Redes Neurais Pulsadas,

[Distrito Federal] 2010.

x, 36p., 297 mm (ENE/FT/UnB, Engenheiro Eletricista, 2010). Trabalho de Graduação – Universidade de Brasília. Faculdade de Tecnologia.

1. Redes Neurais
2. Algoritmo Backpropagation
3. Reconhecimento de Fonemas
4. Teorema de Bayes

I. ENE/FT/UnB
II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

MELLO, A. U. M. & NISHIYAMA, C. N. (2010). Reconhecimento de Fonemas utilizando Redes Neurais Pulsadas. Trabalho de Graduação em Engenharia Elétrica, Publicação ENE 01/2010, Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 36p.

CESSÃO DE DIREITOS

AUTORES: Alexandre Ungaretti Marcondes de Mello e Caio Nunes Nishiyama

TÍTULO: Reconhecimento de Fonemas utilizando Redes Neurais Pulsadas.

GRAU: Engenheiro Eletricista ANO: 2010

_____
Alexandre Ungaretti Marcondes de Mello

_____
Caio Nunes Nishiyama

Dedicatórias

Aos meus pais Ligia e Makoto.
Aos meus irmãos Vitor e Leandro.

Caio Nunes Nishiyama

Aos meus pais
A minha família

Alexandre Ungaretti Marcondes de Mello

Agradecimentos

Ao professor Alexandre Ricardo Soares Romariz, por sua orientação e incentivo ao longo do projeto.

Aos meus pais, pelo carinho e apoio de sempre em minhas decisões.

Aos meus irmãos, por sempre estarem ao meu lado.

Ao amigo Caio Nunes Nishiyama, pelo companheirismo e paciência durante o trabalho.

Alexandre Ungaretti Marcondes de Mello

A Deus, Inteligência Suprema, fonte de infinita sabedoria.

Ao Professor Alexandre Ricardo Soares Romariz, por sua orientação, seu incentivo e sua paciência durante todo o desenvolvimento do projeto.

Ao amigo e companheiro de projeto Alexandre Ungaretti Marcondes de Mello, pelo apoio de sempre.

Aos professores, funcionários e colegas do Departamento de Engenharia Elétrica da Universidade de Brasília, por contribuírem de alguma forma para a minha formação.

Caio Nunes Nishiyama

RESUMO

Criamos um sistema para o reconhecimento de palavras baseado em fonemas, uma Rede Neural Pulsada capaz de reconhecer fonemas com pré-processamento reduzido. Os fonemas orais foram utilizados para o treinamento da rede.

A rede neural pulsada foi utilizada devida sua grande capacidade de generalização. Uma de suas características é a capacidade de responder bem a situações inesperadas. Outra vantagem da rede é a imunidade contra o ruído, situação que a rede pulsada também tem boa atuação.

A entrada da rede leva em consideração as parcelas de energia de determinadas faixas de freqüência. Para a segmentação do espectro em bandas de energia um banco de filtros triangulares na escala mel foi usado. O sistema de classificação, baseado no teorema de Bayes da probabilidade condicionada, foi usado após o reconhecimento dos fonemas.

ABSTRACT

Speech recognition is an important subject of study in Computer Science and Engineering. We created a system for words recognition based on phonemes, a Pulsed Neural Network able to recognize phonemes with a little pre-processing. The oral phonemes were used for the network practice.

The pulsed neural network was used because of its ability to generalize and to respond to unexpected situations. Another advantage of the net is its behavior under noise, a situation in which the pulsed neural network has a good performance.

The network input is the energy spectrum of the signal in specific frequency bands. For the segmentation of the spectrum in energy bands, a bank of triangular filters was used. The classification system, based on the Bayes Probality Theorema, was used after the phonemes recognition.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

Símbolos Latinos

e	*Tensão de entrada*	*[V]*
f	*Frequência*	*[Hz]*
T	*Tempo*	*[s]*
V	*Tensão*	*[V]*

Símbolos Gregos

δ	Erro	
Δ	Variação entre duas grandezas similares	
ξ	Coeficiente de aprendizagem	
∂	Derivada parcial	
τ	Constante de tempo	[1/s]

Grupos Adimensionais

W	*Peso sináptico*

Subscritos

i	*Número da camada*
j	*Número do neurônio*

Siglas

DC	Correte direta
GMM	Modelos de Mistura Gaussiana
HMM	Modelos Ocultos de Markov
MATLAB	Software interativo voltado para o cálculo numérico
Mel	Unidade de frequência
MFCC	Mel Frequency Cepstral Coefficients
RNA	Redes Neurais Artificiais

1 Introdução

1.1 Aspectos gerais

O processo de entendimento de palavras por humanos e até mesmo alguns animais é de fácil verificação, mas quando a tentativa de reconhecer uma palavra é dada ao computador, torna-se mais complexo. A dificuldade encontrada pelos computadores no reconhecimento de voz deve-se a sua forma de processar a informação. Computadores analisam informações de forma precisa e rápida, entretanto não conseguem lidar com a generalização, isto é, qualquer entrada desconhecida ou não esperada provoca erro. O cérebro processa a informação de forma mais robusta. Tarefas como reconhecer a voz de uma pessoa ou ler uma carta escrita requer conhecimento de padrões anteriormente adquiridos e não somente da informação obtida.

A fala é a comunicação mais utilizada pelos homens e sua compreensão por máquinas e computadores pode auxiliar em diversas áreas de atuação humana. Há multiplicidade de situações em que o reconhecimento de palavras pode ser empregado. O uso de sistemas de reconhecimento de voz já vem sendo empregado, por exemplo, em celulares, programas de transcrição de texto e centrais de atendimento ao consumidor. Problemas encontrados por deficientes físicos, como acesso a máquinas situadas em lugares remotos ou em ambientes inóspitos poderão ser superados com programas de reconhecimento de voz.

Várias técnicas são utilizadas para reconhecimento de palavras atualmente. São exemplos as técnicas Modelos Ocultos de Markov (HMMs), Modelos de Mistura Gaussiana (GMMs) e Redes Neurais Artificiais (RNA). O processo de reconhecimento já está bem avançado, porém nenhuma dessas técnicas tem 100% de sucesso. Para cada uma dessas técnicas existem limitações e vantagens em relação à outra. Os Modelos Ocultos de Markov são menos suscetíveis a variação do locutor, porém apresentam perda de eficiência significativa em ambientes ruidosos[1]. Já os Modelos de Mistura Gaussiana são mais usados para reconhecimento de locutores. As Redes Neurais Artificiais tem como vantagens a capacidade de ajustar novas informações e boas respostas a dados ruidosos.

O grande desafio é obter um método eficiente e robusto ao mesmo tempo. O grande problema no processo de reconhecimento é a qualidade do som extraído, e alguns fatores comprometem essa qualidade[1][2]. Podemos citar alguns:

- Ruído externo que se mistura ao sinal original;
- As características particulares de cada locutor;
- A qualidade do transdutor.

Optamos por trabalhar com redes neurais artificiais para o reconhecimento das palavras. O presente projeto tem como objetivo a identificação dos fonemas com reduzido pré-processamento, ou seja, deseja-se que a rede consiga extrair os padrões sem que haja excesso de pré-processamento, o que deixa o sistema com maior retardo. O foco é nos fonemas, que são as menores unidades sonoras que possuem significado, e posteriormente reconhecimento da palavra.

2 Fundamentos Teóricos

2.1 Fonologia

Fonologia é o ramo que estuda os sons de uma língua. O fonema é a menor unidade sonora da língua. Com a modificação de apenas um fonema podemos mudar o sentido de uma frase toda. Na língua portuguesa temos 34 fonemas que constituem o alfabeto fonético. Podemos separar o alfabeto fonético em três grupos distintos:

- Vogais: 13
- Semivogais: 2
- Consoantes: 19

As características dos fonemas podem ser obtidas através da forma de sua produção no aparelho fonador. As vogais são fonemas que não encontram obstáculos na sua pronúncia, ou seja, o ar passa livremente pela boca. As semivogais são fonemas com som de vogal, mas se apóiam nesta para a construção da sílaba.[3] As consoantes têm em algum momento da sua pronúncia o ar interrompido e não conseguem construir núcleo silábico. A Figura (2.1) mostra os principais integrantes do aparelho de produção da fala humana.

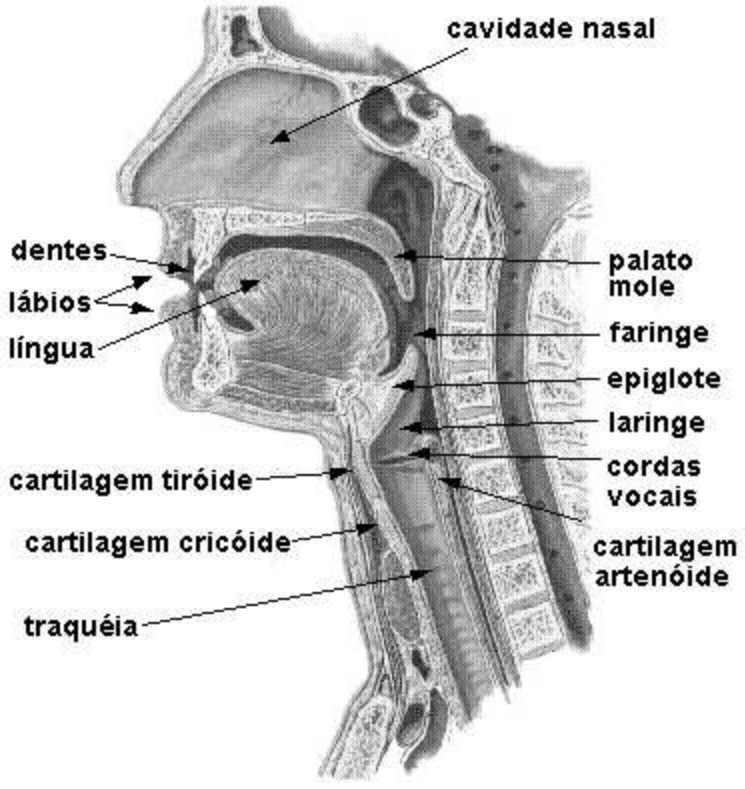

Figura 2.1 - Aparelho fonador.[4]

As vogais podem ainda ser classificadas em relação a passagem do ar pela boca, a zona de articulação, a intensidade e ao timbre. Quanto ao timbre as vogais podem ser abertas, quando se abre o máximo da boca na fala, ou fechadas, quando se abre o mínimo da boca na fala. Já em relação à zona de articulação elas se classificam em posteriores, quando pronunciadas com a língua posicionada no fundo da boca, entre o dorso da língua e o véu palatino, em anteriores, quando pronunciadas com a língua posicionada na frente da boca entre o dorso da língua e o palato duro ou em centrais, quando são pronunciadas com a língua posicionada no centro da boca. Quando se trata de intensidade as vogais podem ser tônicas, subtônicas ou átonas. Em se tratando do modo de articulação as vogais podem ser orais ou nasais.

Na língua portuguesa as vogais são a base das silabas. Elas são as unidades fonéticas com maior energia e que melhor conseguimos identificar na pronuncia das palavras[3]. Por essas características serão utilizadas vogais no processo de reconhecimento das palavras.

2.2 Audição humana

O sistema auditivo humano é dividido em três zonas. A primeira é chamada de ouvido externo que é composto pela orelha e o canal auditivo, essa zona tem a função de transportar o som até o tímpano. A segunda é o ouvido médio que nada mais é do que um acoplamento mecânico entre as membranas do tímpano e a da janela oval da cóclea. Nesse acoplamento estão localizados os três ossos, o martelo, a bigorna e o estribo, responsáveis pela transmissão da energia. Já no ouvido interno temos a cóclea. A cóclea funciona como um transdutor, converte a energia mecânica vinda da segunda zona em impulsos elétricos. Esses sinais são transmitidos até chegarem ao cérebro.

Estudos em psicoacústica revelaram que a percepção humana de frequência de tons ou de sinais de voz não segue uma escala linear. Surgiu então a escala mel que melhor traduz o comportamento da percepção do cérebro. O Mel é uma escala não-linear que simula a percepção do ouvido humano. Em sistemas de reconhecimento de voz técnicas baseadas na escala mel para extração de parâmetro são as mais usadas[5].

2.3 Motivação biológica

O cérebro humano possui mais de 86 bilhões de neurônios. Os neurônios são células base do sistema nervoso. A capacidade de sentir emoções, formular pensamentos e executar as funções sensoriais é atribuída a interconexão dessas células básicas.

As Redes Neurais Artificiais são sistemas de processamento numérico inspiradas no sistema nervoso biológico, sendo constituídas por vários processadores simples amplamente conectados entre si. Difere do computador tradicional possui processamento paralelo e distribuído, ao contrário de um processador central.

A análise do neurônio biológico é fundamental para conhecermos algumas características que serão utilizadas em nosso modelo computacional de neurônio. O neurônio, assim como qualquer célula biológica, possui uma fina membrana celular para executar suas funções normais, e além disso possui determinadas propriedades que são essenciais para o funcionamento da célula nervosa. A partir de seu corpo celular, que é o centro de seus processos metabólicos, projetam-se os dendritos e o axônio, mostrados na Fig. (2.2).

Os dendritos são responsáveis pela recepção dos impulsos nervosos e transmissão para o corpo celular, onde a informação é processada e novos impulsos são gerados. O ponto de contato entre dendritos e axônios de diferentes neurônios é o local da sinapse. A rede é então formada pelas ligações sinápticas entre neurônios, e por elas ocorre o fluxo de informação. O efeito da sinapse é variável e pode ser adaptado, esse fator faz com que as ligações entre os neurônios sejam diferentes entre si. Acredita-se que o aprendizado está justamente na variação das ligações sinápticas[6].

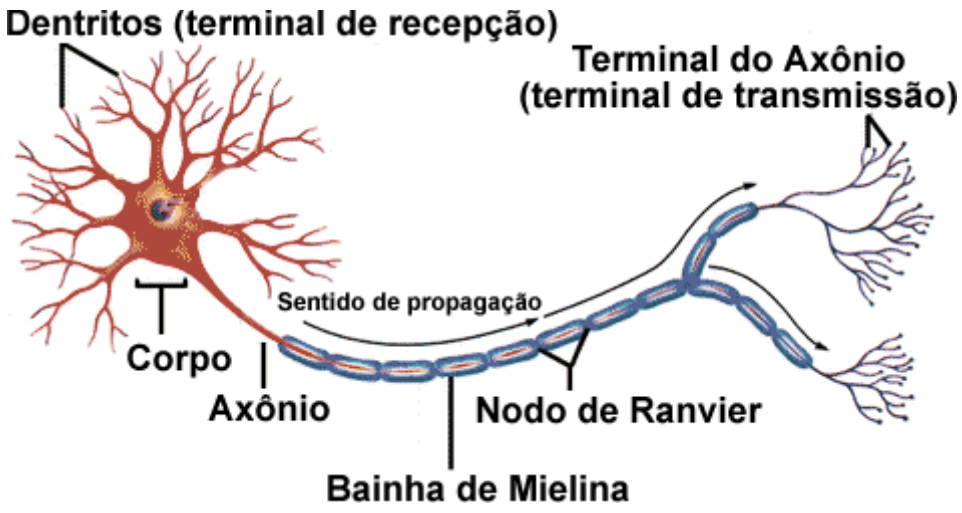

Figura 2.2 - Neurônio biológico com corpo celular, dendritos e axônio.[7]

A transmissão do impulso é feita levando em consideração todos os impulsos recebidos, comparando-se os sinais recebidos. Se o percentual em um intervalo curto de tempo for suficientemente alto, a célula "dispara", transmitindo o pulso.

O pulso ocorre porque existe uma diferença de potencial elétrico na membrana externa do neurônio. Ela é controlada pelas concentrações de sódio e potássio interna e externa à célula nervosa. Uma perturbação no potencial de equilíbrio do neurônio propaga-se pela célula, dos dendritos em direção ao axônio, passando pelo corpo celular, como na transmissão de um sinal elétrico em uma linha de transmissão. Após este fenômeno, a célula precisa de tempo pra voltar à sua condição inicial. Quando ocorre um trem de pulsos elétricos obtidos dos estímulos nos neurônios, a célula nervosa vai somando estes estímulos e os comparando com um limiar. Quando este é atingido, o neurônio emite um pulso que se propaga. O mesmo processo ocorre nos neurônios seguintes.

3 Metodologia

3.1 Visão geral

Foi desenvolvido um programa para o reconhecimento de fonemas através de redes neurais artificiais. Para isso foi utilizado a ferramenta MATLAB. Para um grupo pequeno de palavras a abordagem foi focada nos fonemas á, é, ê, í, ó, ô e û. Escolhemos as vogais por terem maior energia em relação às consoantes, isso facilita a identificação e o reconhecimento[3]. Para o reconhecimento dos fonemas usamos o método de redes neurais pulsadas.

Foi desenvolvido um modelo de neurônio baseado em um integrador, que dispara quando sua excitação alcança um valor limite.

A rede é construída a partir da concatenação de vários neurônios, dispostos aleatoriamente de forma a prover complexidade ao sistema. Os sinais processados na rede foram conectados a um grupo de neurônios na saída com intensidades distintas, aleatoriamente distribuídas, consideradas como pesos para cada sinal de acordo com o neurônio em que se conecta.

Com os dados disponíveis, inicia-se o ajuste desses pesos para que cada neurônio de saída possa ser capaz de identificar um diferente fonema. Este treinamento ocorre até que a identificação dos fonemas seja considerada satisfatória.

Com os pesos ajustados, podemos aplicar o sinal na entrada da rede, o sinal será identificado de acordo com suas características e a saída da rede informará a vogal reconhecida. O Teorema de Bayes foi usado para obter a palavra com maior probabilidade de acerto. O algoritmo analisa os fonemas reconhecidos e fornece a possível palavra, ou seja, a palavra com maior probabilidade de acerto.

3.2 Extração do sinal de voz

O primeiro passo no processo de reconhecimento de palavras é a extração do sinal de voz. Esse procedimento envolve a transdução, processo que transforma o sinal acústico em sinal elétrico. Para a captação do sinal utilizamos um microfone como transdutor.

Como o sinal acústico é analógico, temos que fazer a conversão analógico-digital para o cálculo no MATLAB. O próprio MATLAB foi o software utilizado para conversão. A faixa audível humana é de 20 a 20.000 Hz, entretanto a maior parte da informação esta contida nos primeiros 4.000 Hz [8]. Usamos o critério de Nyquist[9] para reduzir o custo computacional e garantir a inteligibilidade do sinal. A taxa de amostragem utilizada foi de 8000 Hz. Um único canal de áudio foi utilizado e cada amostra tem 16 bits de resolução.

Nesse sinal realizamos algumas modificações a fim de eliminar componentes do sinal que não fazem parte do sinal de voz original. O nível DC, causado por interferências externas, é uma dessas características eliminadas, devido sua alta amplitude. O nível DC modifica o espectro do som puro dos fonemas, dando ênfase nas baixas freqüências, situação em que informações importantes perdem seu valor. Para demonstrar o resultado deste pré-processamento, retiramos o valor DC da vogal “á”, como mostrado nas Fig. (3.1) e (3.2).

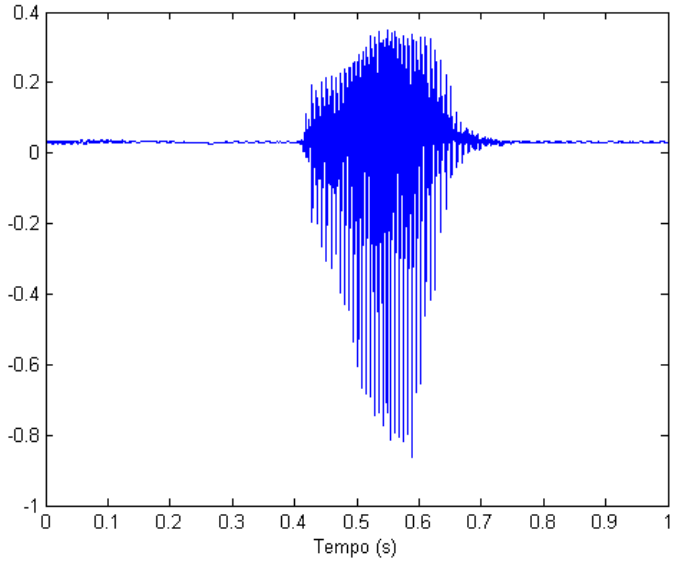

Figura 3.1 – Vogal “a” com nível DC.

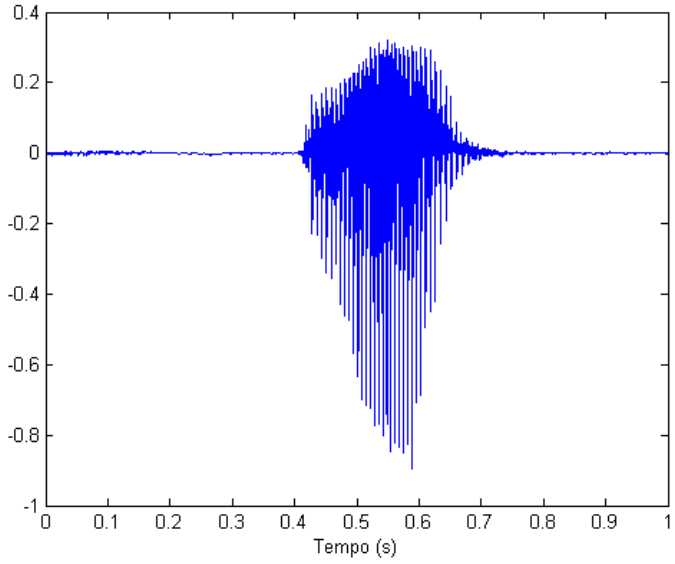

Figura 3.2 - Vogal “a” sem nível DC.

Outro ajuste feito é a normalização, que traz todos os pontos do sinal para uma escala única em amplitude, com valor máximo 1, assim a amplitude máxima dos sinais se torna a mesma, não existindo diferenciação entre o mesmo sinal falado próximo ou longe do transdutor, como ilustra as Fig. (3.3) e (3.4).

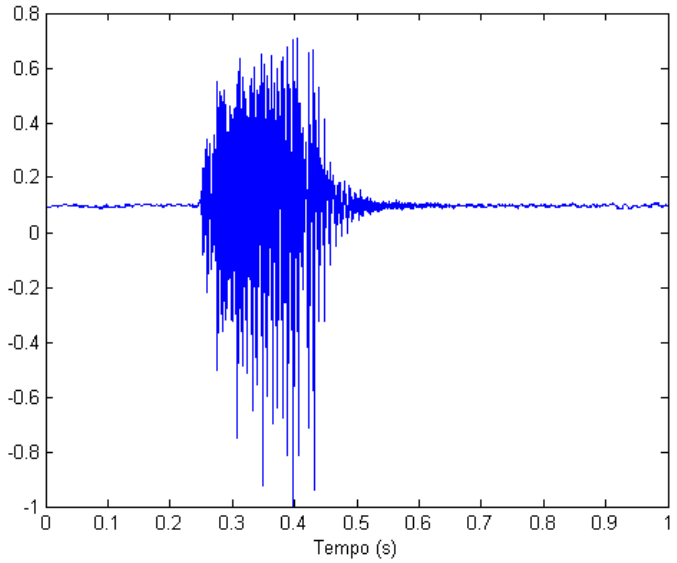

Figura 3.3 - Sinal normalizado.

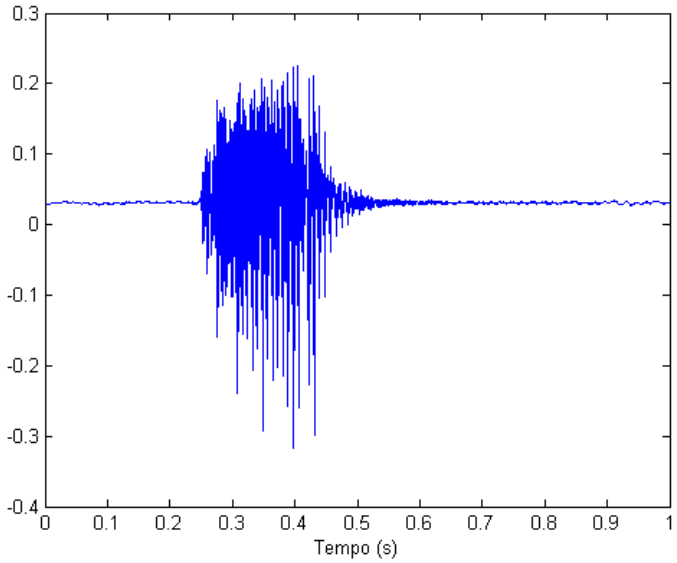

Figura 3.4 - Sinal sem normalização.

A Figura (3.5) mostra um diagrama de blocos com todo o pré-processamento que o sinal recebe até passar pelo banco de filtros.

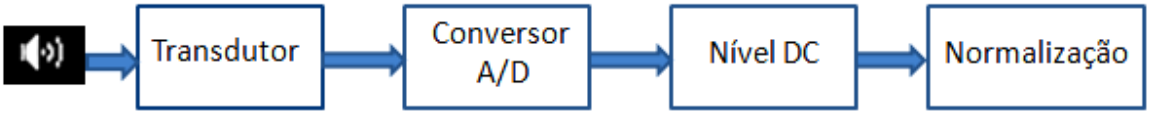

Figura 3.5 - Processamento do sinal.

3.3 Banco de dados

Em sistemas de reconhecimento de voz utilizando redes neurais artificiais temos que ter um banco de dados que possibilite o treinamento dessa rede. A rede reconhece padrões e poderia ser treinada para identificar todas as palavras, bastando ter um banco de dados com todas as palavras desejadas, porém quando trabalhamos com uma grande quantidade de palavras a obtenção dessas e o treinamento das mesmas se torna infausto. Na metodologia de reconhecimento através de fonemas o objeto de reconhecimento da rede se torna os fonemas. Assim, restringimos nosso banco de dados aos fonemas mostrados na Tab. (3.1), o que ajuda tanto em sua formação quanto nos treinamentos das redes.

Sabe-se que os sistemas de reconhecimento são suscetíveis à variação do locutor, ou seja, o sistema reconhece alguns padrões que são específicos do locutor o que deixa o sistema inviável quanto à utilização do público em geral. A obtenção de dados de diferentes locutores e em diferentes recintos leva a um sistema mais robusto. Temos que adequar a rede à sua devida utilização, ou seja, se o sistema será utilizado por todos, o banco de dados deverá ter a maior quantidade de variações de locutores possível.

No presente trabalho temos o propósito de identificar dez palavras utilizando o mesmo locutor. Assim, o banco de dados foi obtido utilizando-se apenas um locutor. Os fonemas utilizados no treinamento da Rede Neural Artificial (RNA) foram obtidos em ambiente de escritório. O mesmo método utilizado na extração do sinal de voz foi empregado para obtenção do banco de dados. Como a técnica é baseada no reconhecimento da palavra usando apenas vogais, o banco de dados para treinamento consiste nas vogais mostradas na Tab. (3.1).

Tabela 3.1 - Fonemas utilizados no sistema de reconhecimento.

Á	É	Ê	I	Ó	Ô	U

O banco de dados possui dez amostras de cada vogal acima, que vão ser utilizadas no treinamento da RNA.

3.4 Banco de filtros

O banco de filtros separa a energia de diferentes faixas de freqüência. A informação que diferencia os fonemas deve ser separada daquela que não nos ajuda a reconhecê-los. Os dados que entrarão na rede devem ter o mínimo de informação possível para identificar os fonemas.

No processamento computacional, devemos ter uma grande preocupação com a obtenção dessas características, que são de grande relevância na qualidade do reconhecimento. São conhecidas diversas formas para extração desses parâmetros, tais como a Codificação por Predição Linear[10] na qual assumimos que a amostra em questão é uma combinação linear de amostras anteriores.

Para obtenção das características que levam a identificação de cada fonema, vamos usar a técnica baseada em MFCC (Mel Frequency Cepstral Coefficients). MFCC é a técnica mais utilizada para reconhecimento de voz. Esta trabalha de modo não linear como a percepção pelo cérebro humano da voz. O Mel é um mapeamento não-linear das freqüências que compõem um sinal. Para a conversão foi adotado como referência a freqüência de 1000 *Hz*, com potência de 40 dB acima do limiar mínimo do ouvido humano, equivalem a 1000 mel[5]. Dessa forma os outros valores foram obtidos experimentalmente, resultando no critério de conversão segundo a Eq. (3.1):

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

Em que f representa a frequência em hertz. Com o sinal convertido na escala mel é utilizado um banco de filtros de formato triangular com largura de banda igual a 300 mels espaçados de 150 mels. Como utilizamos uma taxa de amostragem de 8000 *Hz*, a construção de um banco de 13 filtros é suficiente para cobrir todo o espectro analisado. A saída de cada filtro é uma entrada na rede neural artificial, diferentemente da técnica MFCC, que com o resultado dos filtros ainda calcula os coeficientes mel-cepstrais[5], como mostra a Fig. (3.6).

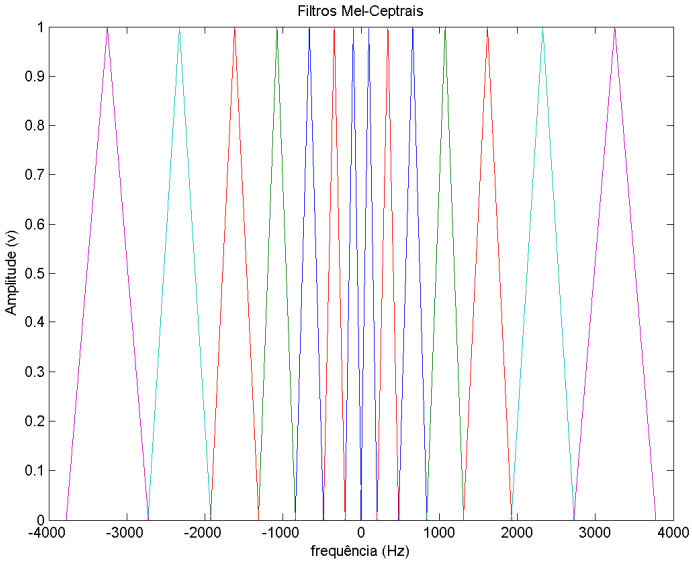

Figura 3.6 - Banco de filtros.

3.5 Modelo de neurônio

O modelo de neurônio desenvolvido foi baseado no funcionamento de um circuito RC série [11], ilustrado na Fig. (3.7), e, como tal, seu funcionamento é semelhante a um integrador de acordo com a Eq. (3.2). Sua saída, porém, é sempre nula, exceto quando a integração atinge seu limiar de saturação. Neste caso, o neurônio emite um pulso de saída. Seu funcionamento completo está ilustrado na Fig. (3.8).

$$\frac{\partial V(t)}{\partial t} = \frac{e(t) - V(t)}{\tau} \tag{3.2}$$

Em que e(t) representa o sinal de entrada, V(t) a saída e τ representa a constante de tempo da integração. Este valor foi definido por meio de experimentos, de tal forma que sua magnitude foi adaptada de acordo com a freqüência do sinal de entrada.

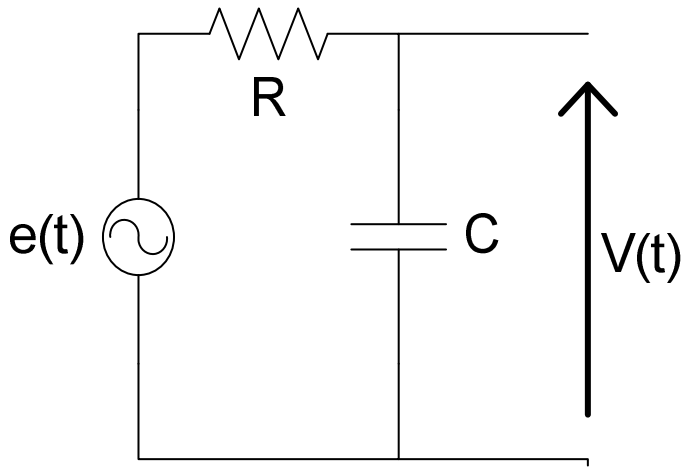

Figura 3.7 - Circuito em que se baseia o funcionamento do neurônio.

Podemos perceber um tempo de espera para o recomeço da integração. Ao finalizar o pulso, o integrador só volta à ativa após o período refratário, esse período limita a frequência de disparos. Existem dois tipos de períodos refratários, o período refratário absoluto e o relativo. No período relativo, o integrador após o pulso se comporta de forma diferente ao integrador pré-pulso. Ele, durante um intervalo de tempo, se comporta de maneira menos suscetível ao sinal. Já no período absoluto o integrador não atua no período refratário. Com isso um sinal só dispara mais de um pulso se houver energia suficiente para o integrador alcançar o limiar, esperar o período refratário e ainda existir a energia requerida para outro disparo.

Determinamos a largura do pulso de maneira experimental. Devemos tomar cuidado para que a largura do pulso não fique demasiadamente grande, o que causaria uma sobre excitação nos neurônios seguintes.

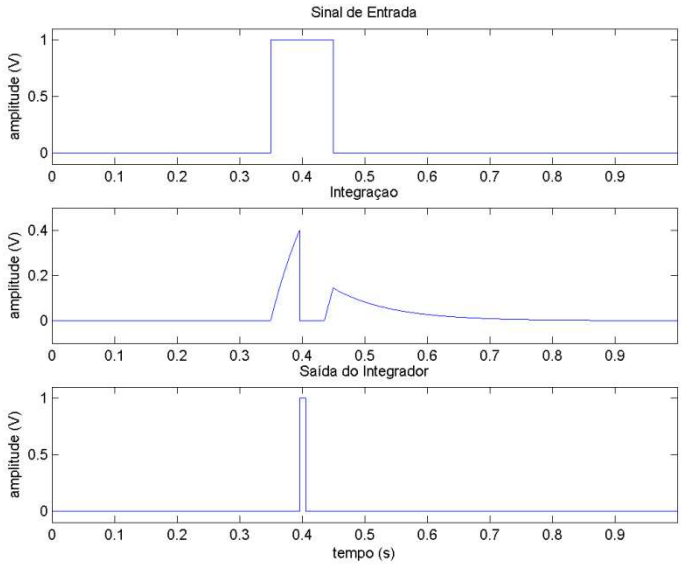

Figura 3.8 - Funcionamento do neurônio

3.6 A rede neural

A complexidade do sistema não está no processamento do sinal que o modelo de neurônio desenvolvido executa, e sim nas ligações existentes entre vários neurônios. Foram desenvolvidos dois sistemas utilizando este modelo, sendo o primeiro um sistema dinâmico interconectado aleatoriamente para representar o histórico do sinal de entrada, que é a rede neural. O segundo sistema foi desenvolvido para captar os sinais de saída, sendo que estes serão os responsáveis por tomar decisões sobre o fonema que foi pronunciado.

A rede neural recebe como entradas os sinais processados pelo banco de filtros, além de receber os valores de polarização. O trem de pulsos gerado na saída de cada neurônio serve de entrada para o neurônio seguinte, e sua saída decorre do processamento do somatório de todas as suas entradas. As ligações que ocorrem na rede neural podem ser observadas na Fig. (3.9).

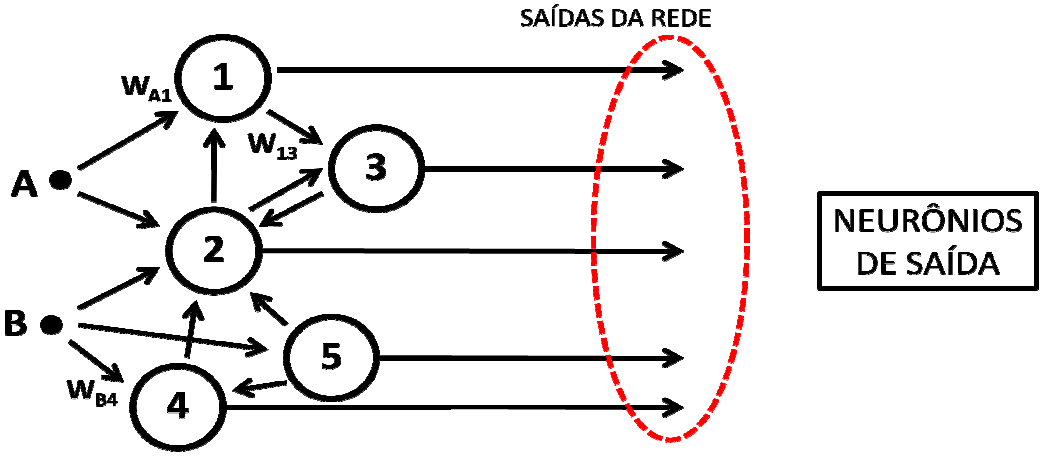

Figura 3.9 - Conexões da rede neural.

Cada sinal de entrada em um neurônio é ponderado por um peso W_{ij}, que representa a intensidade da ligação entre dois neurônios ou entre o neurônio e o meio externo, sendo que o índice 'i' representa a camada em que o neurônio se encontra e o índice 'j' representa o número do neurônio.

Para a construção da rede, os neurônios foram conectados de tal forma que todos se comunicam com intensidades diferentes. Estas intensidades foram selecionadas de forma aleatória. Além disso, cada neurônio está preparado para receber as entradas externas, sendo que cada entrada está associada a um determinado grupo de neurônios. Esta associação também foi determinada de forma aleatória, pela função *rand* do MATLAB. Os sinais gerados na rede neural serão processados pelos neurônios de saída.

3.7 Neurônios de saída

Os neurônios de saída são responsáveis por interpretar a informação que foi gerada na rede neural. Seu funcionamento é semelhante ao do neurônio utilizado na rede, porém a constante de tempo τ é significativamente maior, tornando a integração mais lenta. Neste tipo de neurônio não há condição de limiar para que a saída pulse. A saída deste será a integração do sinal de entrada.

Foram utilizados sete neurônios na construção da saída. Cada neurônio será responsável por identificar um fonema, através de um pulso no momento em que o fonema for pronunciado, enquanto as outras saídas permanecem sem reação.

3.8 Treinamento

Para que os neurônios de saída reconheçam o fonema pronunciado, é necessário que a intensidade das ligações entre os neurônios da rede e os de cada saída sejam diferentes. Esta diferença torna cada neurônio de saída especialista em reconhecer um determinado fonema.

No modelo do neurônio de saída, os pesos W_{ij} que cada sinal em sua entrada recebe foram selecionados aleatoriamente, de forma que a saída não respondesse a nenhum fonema. Estes pesos foram então ajustados, por meio de um treinamento, para que respondessem melhor a determinado sinal. O treinamento foi desenvolvido de forma supervisionada, isto é, a rede recebeu amostras do sinal de entrada e a respectiva saída que se desejava obter.

A primeira parte consistiu na propagação do sinal pela rede. A saída obtida foi comparada com o sinal desejado, e essa diferença, representada pelo erro δ, foi utilizada para determinar a direção em que os pesos W_{ij} seriam corrigidos. De posse deste erro, o algoritmo de minimização de erro quadrático foi aplicado, corrigindo somente os pesos W_{ij} dos neurônios de saída, sem alterações nas demais interconexões da rede. Pela Figura (3.10), podemos observar um diagrama de blocos do algoritmo de treinamento.

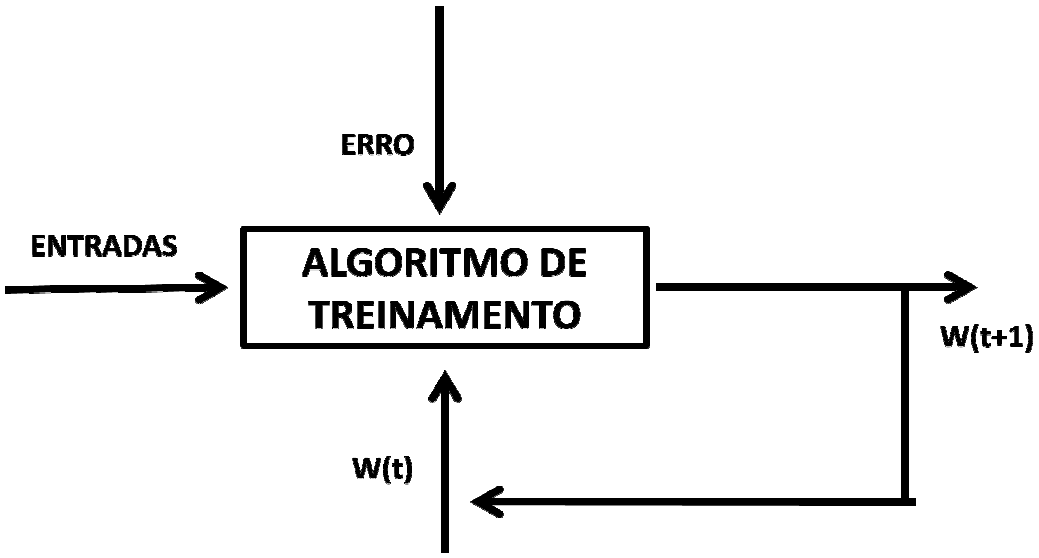

Figura 3.10 - Diagrama de blocos do algoritmo de treinamento da rede.

O valor da correção aplicada aos pesos foi regido pelas Eq. (3.3) e (3.4), em que a entrada $e(t)$ é ponderada pelo erro δ obtido. O passo de correção é governado pelo coeficiente de aprendizado ξ. Para valores elevados de ξ o ajuste dos pesos tende a ser rápido, porém com maior probabilidade de divergência. Por outro lado, valores reduzidos desta constante fazem com que a trajetória calculada seja suave, desta forma o tempo de duração do treinamento aumenta. Temos então uma relação de compromisso entre convergência do erro e tempo de treinamento[12]. Pela Figura (3.11), podemos observar a convergência do erro em relação ao número de épocas. Definiu-se época o conjunto de iterações do algoritmo de treinamento para os quais todas as amostras de sinais de entrada foram apresentadas para a rede.

$$\Delta W = \xi \cdot (\delta \cdot e) \tag{3.3}$$

$$W(t+1) = \mathrm{W(t)} + \Delta W \tag{3.4}$$

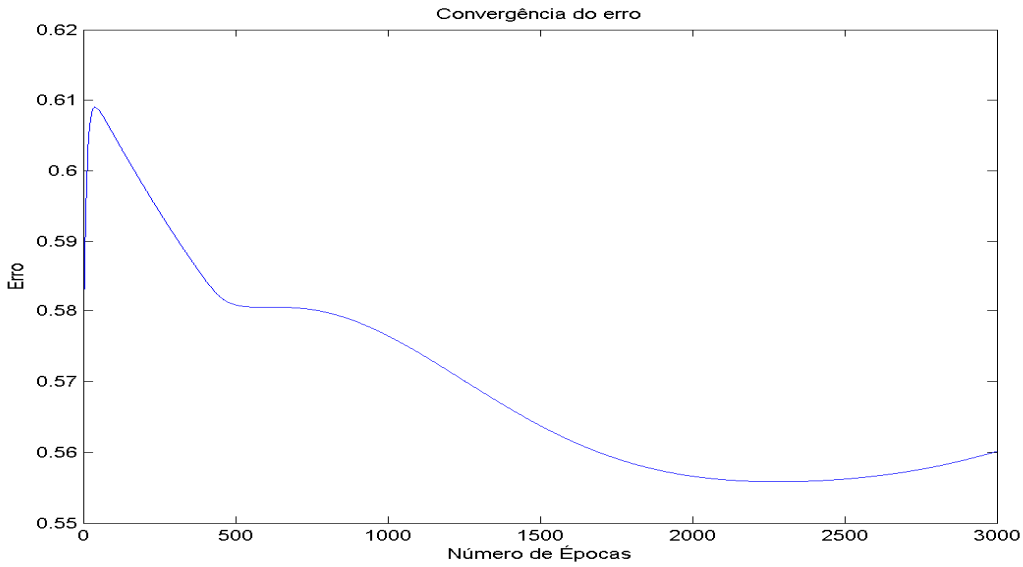

Figura 3.11 - Convergência do erro.

3.9 Sistema de classificação

O sistema de classificação é usado para definir a palavra de saída do sistema. Contudo, antes de aplicarmos o sistema de classificação propriamente dito, devemos analisar a saída da rede. Cada sinal que passa pela rede fornece sete saídas. Cada uma destas corresponde à identificação de um fonema, porém um mesmo sinal pode ser identificado por mais de uma saída. A saída é então comparada com um limiar pré-determinado a fim de só se levar em consideração a saída com certo nível de reconhecimento. O fonema só será reconhecido se a rede permanecer durante certo tempo no patamar escolhido.

O teorema de Bayes da probabilidade condicionada foi usado para a classificação das palavras levando em consideração os fonemas. Esse teorema trata da relação de um fato dado outro, ou seja, a ocorrência de um permite atualizar a estimativa de outro que com esse tenha relação. No escopo do projeto esse teorema condiciona a possível palavra com o fonema reconhecido pela rede. A saída da rede nos fornece o fonema reconhecido pela RNA, esses fonemas servem de entrada do sistema de classificação. Utilizamos Eq. (3.5) para classificação.

$$P(Pa|F) = \frac{P(F|Pa)\cdot P(Pa)}{P(F|Pa)\cdot P(Pa)+P(F|nPa)\cdot\left(1-P(Pa)\right)} \tag{3.5}$$

Em que $\mathrm{P}(\mathrm{Pa} \mid \mathrm{F})$ é a probabilidade da palavra Pa dado o fonema F, $\mathrm{P}(\mathrm{F} \mid \mathrm{Pa})$ é a probabilidade do fonema F dado a palavra Pa, $\mathrm{P}(\mathrm{Pa})$ é a probabilidade da palavra Pa e $\mathrm{P}(\mathrm{F} \mid \mathrm{nPa})$ é a probabilidade do fonema F se a palavra não for Pa. Na Tabela (3.2) temos as probabilidades $\mathrm{P}(\mathrm{F} \mid \mathrm{Pa})$, essa probabilidade nos mostra o erro associado ao projeto de reconhecimento:

Tabela 3.2 - Probabilidade de encontrar o fonema dado a palavra.

Palavra	Á	É	Ê	I	Ó	Ô	U
Zero	0,2	0,8	0,2	0,2	0,2	0,8	0,2
Um	0,2	0,2	0,2	0,2	0,2	0,2	0,8
Dois	0,2	0,2	0,2	0,8	0,2	0,8	0,2
Três	0,2	0,2	0,8	0,2	0,2	0,2	0,2
Quatro	0,8	0,2	0,2	0,2	0,2	0,8	0,8
Cinco	0,2	0,2	0,2	0,8	0,2	0,8	0,2
Seis	0,2	0,2	0,8	0,8	0,2	0,2	0,2
Sete	0,2	0,8	0,8	0,2	0,2	0,2	0,2
Oito	0,2	0,2	0,2	0,8	0,2	0,96	0,2
Nove	0,2	0,2	0,8	0,2	0,8	0,2	0,2

A cada fonema reconhecido, a probabilidade de cada palavra é atualizada e ao final do sinal de voz teremos na saída um vetor que nos dará a palavra com a maior probabilidade de acerto, mostrando assim a palavra reconhecida.

4 Resultados

4.1 Saída do integrador

Os primeiros ensaios do modelo de neurônio desenvolvido foram feitos com sinais simples. Nestes ensaios, adotou-se o valor da constante de tempo τ de $0{,}1\ s^{-1}$ e o limite de saturação foi de $0{,}3\ V$. A duração de um pulso de saída foi de $1\ ms$ e o período refratário de $3\ ms$. Podemos observar pela Fig. (4.1) o comportamento do neurônio para um pulso retangular de $10\ ms$ de duração e $0{,}4\ V$ de amplitude. Nota-se que a excitação não foi suficiente para que o neurônio disparasse. Pela Fig. (4.2), agora com o mesmo pulso retangular, porém de amplitude $0{,}7\ V$, podemos observar o comportamento do neurônio quando sua saída é ativada.

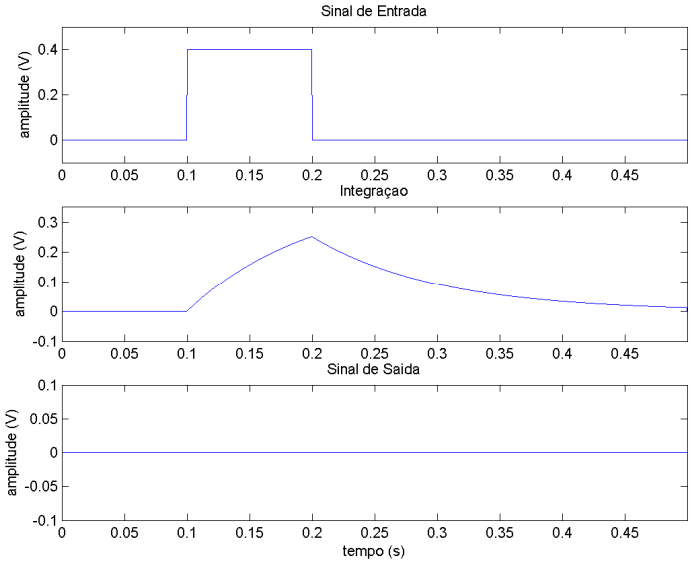

Figura 4.1 - Funcionamento do neurônio para pulso retangular sem ativação da saída

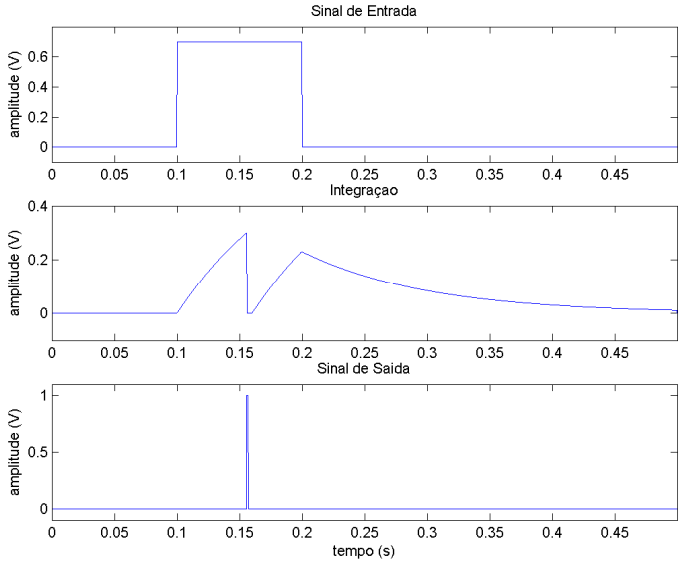

Figura 4.2 - Funcionamento do neurônio para um pulso retangular com ativação da saída

Utilizando os mesmos parâmetros, a rede foi testada para um sinal senoidal de freqüência $4\ Hz$, como ilustrado na Fig. (4.3).

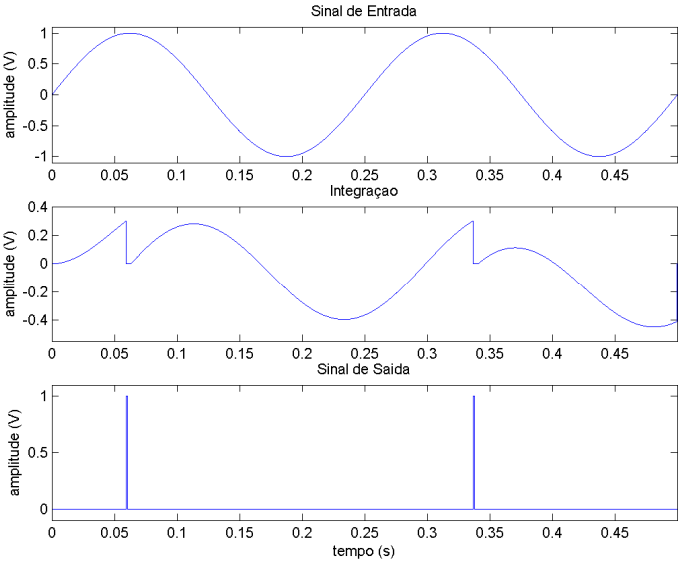

Figura 4.3 - Funcionamento do neurônio para um pulso senoidal.

Para o processamento dos sinais de voz, a constante de tempo τ utilizada foi de $5 \cdot 10^{-4}\ s^{-1}$ e o limite de saturação foi de $0{,}21\ V$. A duração de um pulso de saída é de$1\ ms$ e o período refratário é de $3\ ms$. Estes valores foram escolhidos por meio de experimentos da resposta do integrador para vários sinais de voz.

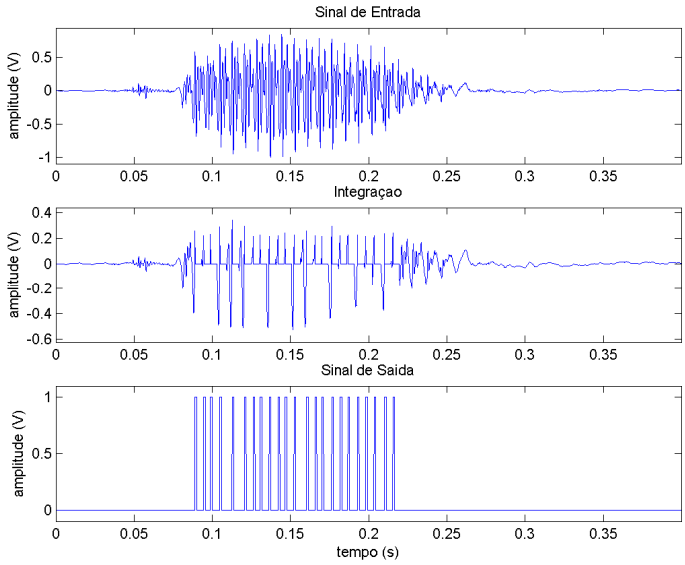

Figura 4.4 - Funcionamento do neurônio para o fonema "ó".

É possível observar pela Fig. (4.4) o comportamento do neurônio para o fonema “ó”. Como os sinais de voz possuem valores de freqüência consideravelmente superiores aos sinais utilizados anteriormente, os parâmetros do modelo de neurônio foram alterados para responder com maior velocidade.

4.2 Saída da rede

Foram feitas simulações para observar o comportamento dos neurônios conectados formando a rede neural. Todos os pesos W_{ij} utilizados foram gerados aleatoriamente, em uma faixa de $-0{,}5$ a $0{,}5$. Podemos observar, pela Fig. (4.5), a resposta de cada neurônio em uma rede de quatro integradores ao sinal retangular que foi utilizado nas simulações anteriores, em que o sinal tracejado de vermelho representa o sinal na entrada da rede, e o trem de pulsos em azul a resposta de cada neurônio.

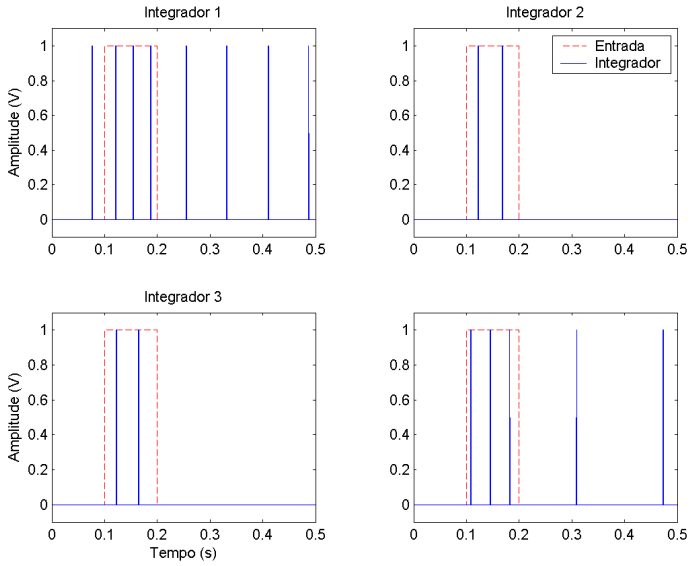

Figura 4.5 - Resposta da rede de quatro integradores a um sinal retangular

Pelas Figuras (4.6) e (4.7), observamos a resposta de uma rede de dez neurônios a um sinal senoidal de 8 Hz.

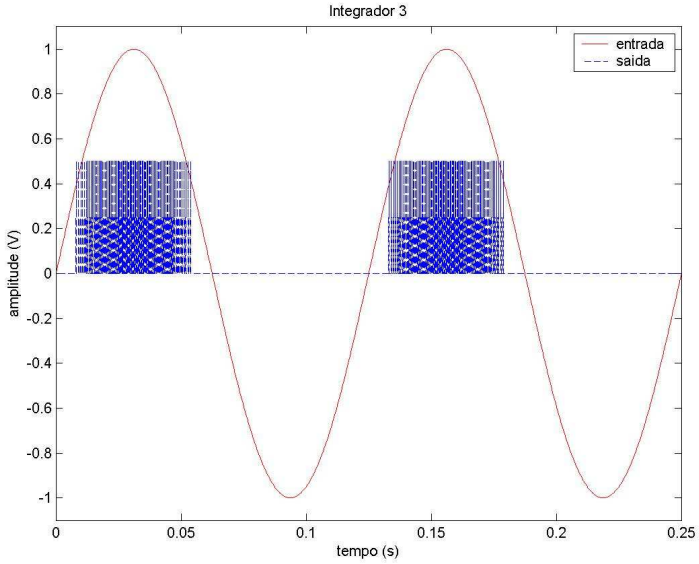

Figura 4.6 - Resposta do integrador três da rede de dez neurônios a um sinal senoidal

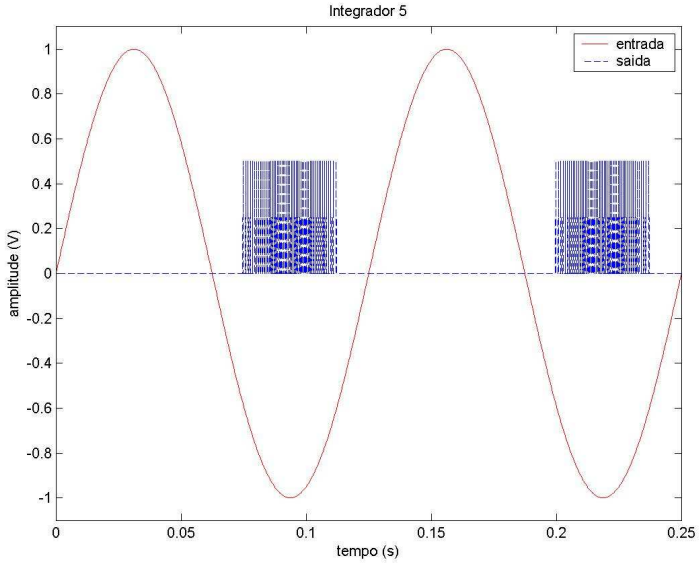

Figura 4.7 - Resposta do integrador cinco da rede de dez neurônios a um sinal senoidal

Podemos observar que houve resposta da rede mesmo quando a amplitude do sinal foi igual ou inferior a zero. Isso ocorreu devido ao sinal de *bias* que se somou à entrada de cada neurônio, ou devido aos valores negativos do peso W_{ij}, para o caso da amplitude negativa da onda senoidal.

A simulação com sinais de voz foi executada utilizando uma rede composta por 20 neurônios. Esta rede recebeu 13 diferentes sinais de entrada, que foram os 13 sinais filtrados por frequências mel-cepstrais. Na Figura (4.8) podemos observar a reação de quatro dos 20 neurônios para o fonema “ô”.

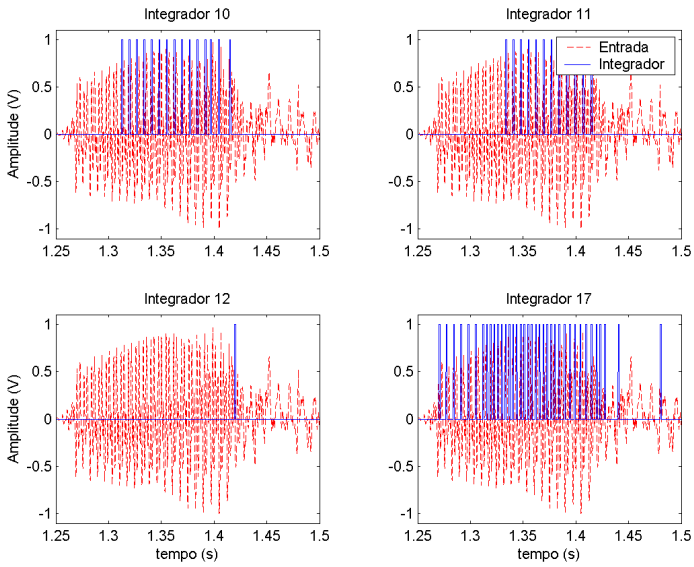

Figura 4.8 - Resposta de quatro dos 20 neurônios ao fonema "ô"

Observa-se a diferença de resposta em cada neurônio, apesar do sinal de entrada ser o mesmo. Pela Figura (4.9) podemos verificar as diferenças de resposta de um mesmo neurônio para diferentes fonemas, como o caso do integrador 11 que não responde ao fonema “ó”, enquanto o integrador 19 não responde ao fonema “i”. Estas diferenças serão importantes para a etapa de treinamento dos neurônios de saída, pois cada um destes sinais receberá uma ponderação de acordo com a sua capacidade de resposta a determinado fonema.

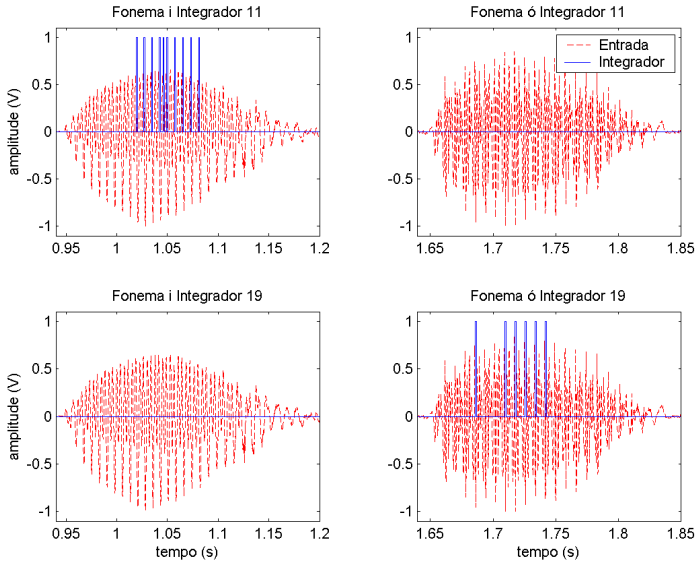

Figura 4.9 - Respostas de dois neurônios para diferentes fonemas.

Pela Figura (4.10), podemos observar o comportamento dos neurônios de saída para um trem de pulsos, que é tipo de sinal que se espera receber na entrada destes neurônios. Nota-se que ele responde tanto a pulsos de amplitude positiva quanto negativa, podendo, desta forma, modelar o sinal de saída para que o seu retorno a zero ocorra de forma mais lenta ou mais brusca.

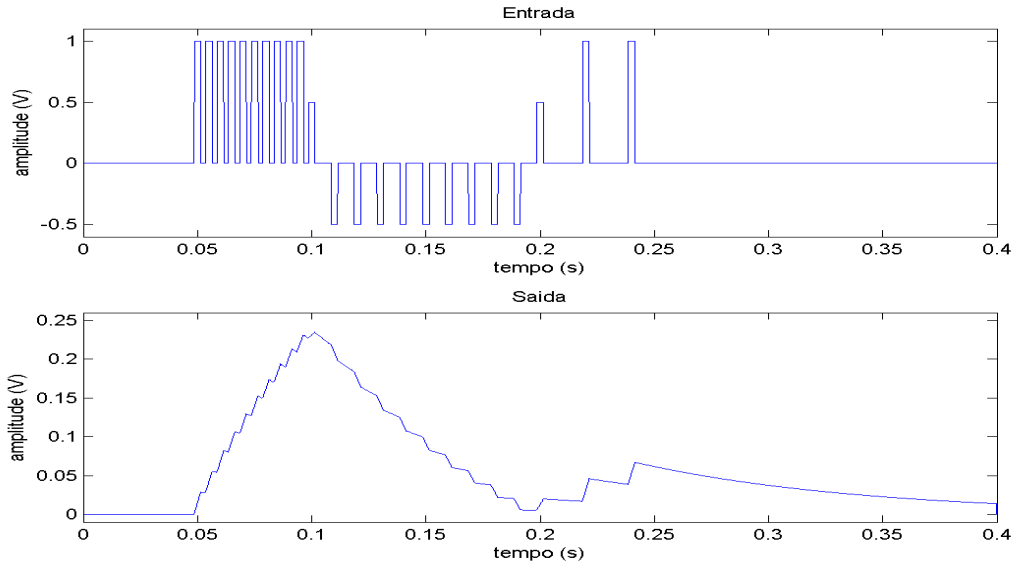

Figura 4.10 - Resposta do neurônio de saída a um trem de pulsos.

4.3 Treinamento

Para a etapa de treinamento foram utilizados sete neurônios de saída. A Figura (4.11) mostra de que forma foram organizadas as conexões entre a saída da rede neural e a entrada da rede de saída. O valor da constante de tempo τ utilizado foi de $0{,}1\ s^{-1}$. Este ajuste de valor se deve ao fato do sinal processado pelo neurônio de saída ser um trem de pulsos diferente dos sinais de voz utilizados na rede neural. O valor da constante de aprendizado ξ foi de $5 \cdot 10^{-4}$. Este valor foi alcançado após estudo sobre a divergência do erro, ou se este se convergiria de forma muito lenta. A intensidade dos pesos da camada de saída do sistema foi gerada aleatoriamente com valores entre $-0{,}05$ e $0{,}05$. Foram escolhidos valores baixos para garantir que inicialmente os neurônios de saída não reagiriam bem a nenhum estímulo, evitando, desta forma, a convergência para mínimos locais.

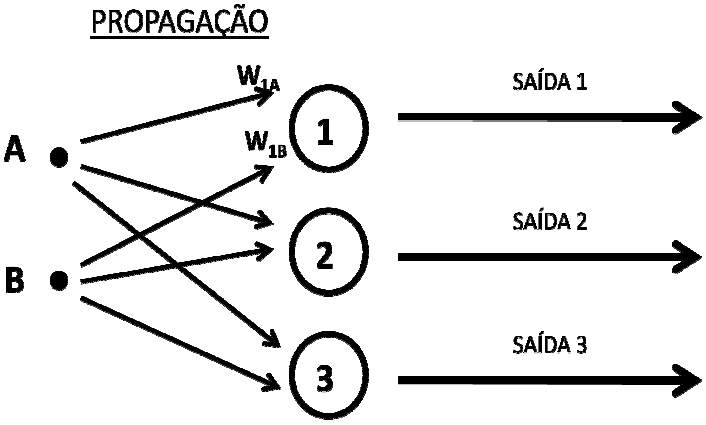

Figura 4.11 - Conexões da rede de saída.

O sinal que se deseja obter foi criado a partir dos sinais de voz na entrada do sistema. Quando determinado fonema foi apresentado à rede para seu treinamento, todas as saídas permaneceram nulas, exceto a saída que se especializou no reconhecimento deste fonema. Pela Figura (4.12), notamos como foram determinadas as saídas desejadas para os fonemas “é” e “i”. Não se tentou reconhecer o fonema desde o instante em que este começou a ser pronunciado, somente algum tempo após o seu início. Como os diferentes fonemas possuem períodos de duração diferentes, o tamanho das janelas dos sinais desejados também são diferentes.

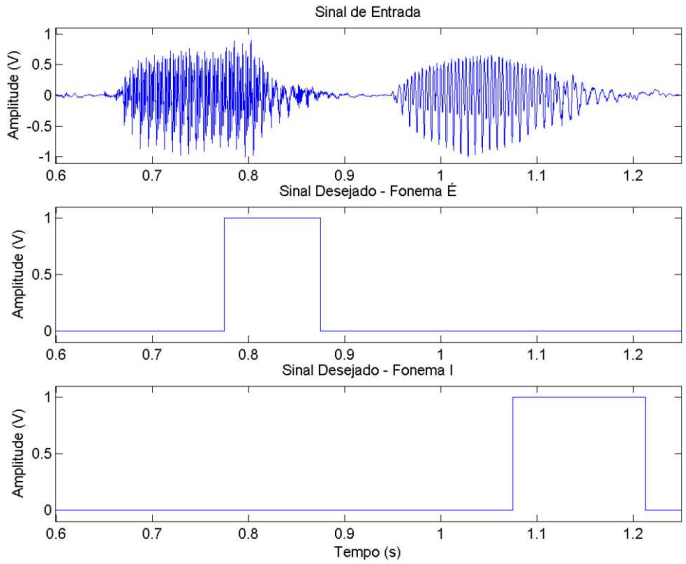

Figura 4.12 - Sinais desejados para os fonemas "é" e 'i"

No início do treinamento, ilustrado na Fig. (4.13), é possível notar que as saídas estão todas próximas de zero.

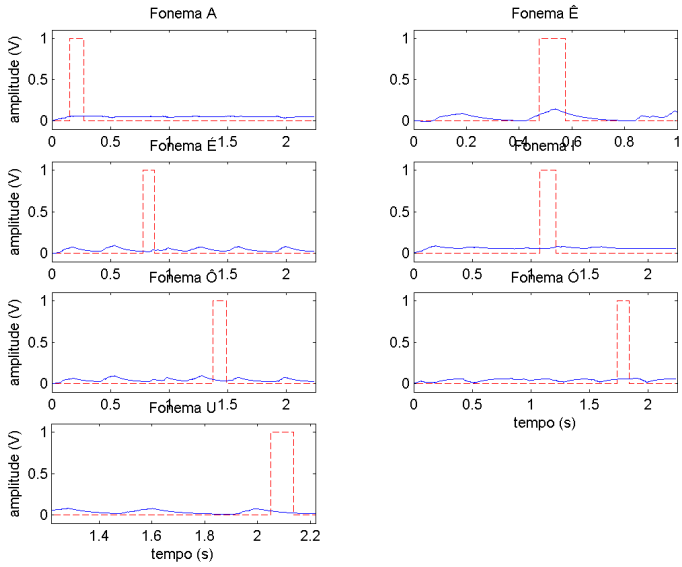

Figura 4.13 - Sinais desejados e sinais obtidos no início do treinamento

Durante o treinamento, os pesos se ajustam para intensificar os melhores estímulos e inibir os que contribuem negativamente. Pela Figura (4.14), é possível observar o desenvolvimento do treinamento após 200 épocas para os fonemas “a”, “é”, “i” e “u”.

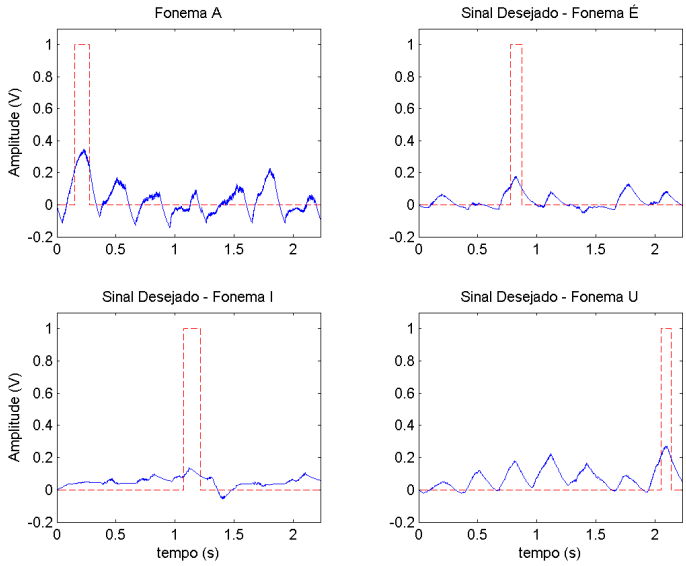

Figura 4.14 - Sinais desejados e sinais obtidos após 200 épocas

A Figura (4.15) ilustra que cada saída começa a se diferenciar após 500 épocas para os mesmos fonemas.

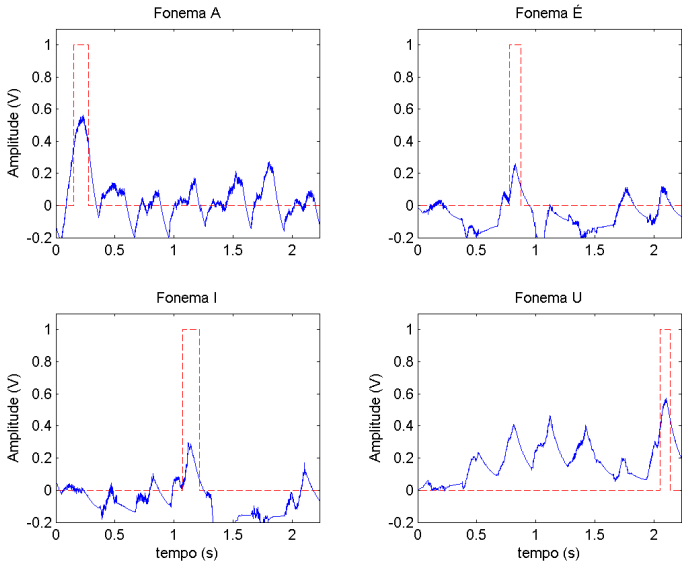

Figura 4.15 - Sinais desejados e sinais obtidos após 500 épocas

A Figura (4.16) ilustra as saídas no final do treinamento. Este resultado foi obtido após 1200 épocas. Observa-se que, para cada fonema, houve proximidade do sinal obtido com o seu respectivo desejado, justificando a interrupção do treinamento.

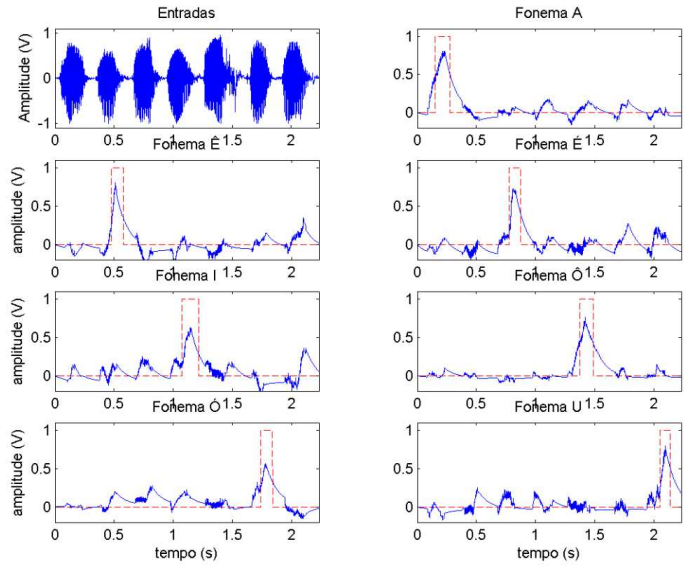

Figura 4.16 - Sinais desejados e sinais obtidos no final do treinamento

4.4 Desempenho da rede

Depois do treinamento a rede foi testada. Primeiramente passamos pela rede todas as vogais. Foram ditas separadamente e captadas pelo mesmo transdutor. Verificamos que ao se pronunciar as vogais separadamente a rede responde de maneira satisfatória. Observado o resultado da saída da rede estipulamos a amplitude e o tempo com o qual vamos considerar que a vogal foi reconhecida. Os gráficos que demonstram as saídas da rede para os fonemas “ô” e “a” estão mostrados nas Fig. (4.17) e (4.18)

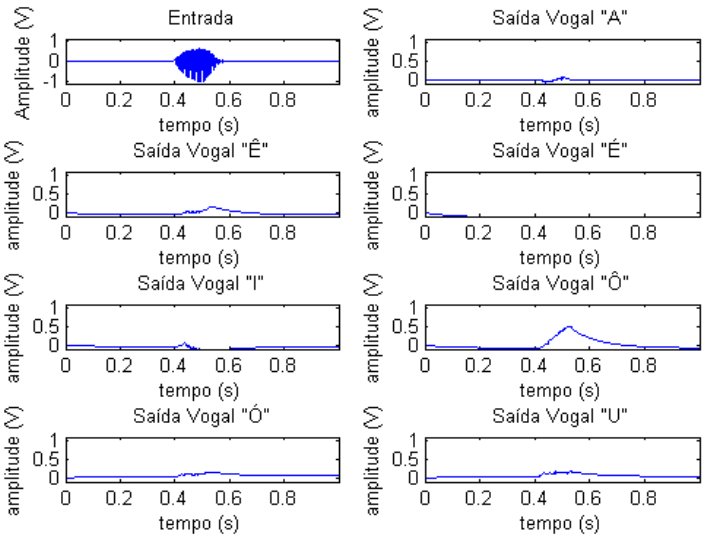

Figura 4.17 - Saída da rede para o fonema “ô”.

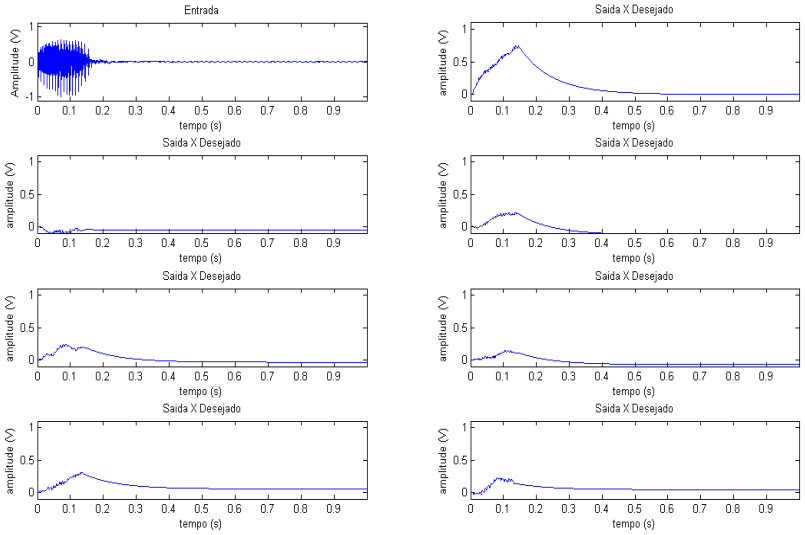

Figura 4.18 - Saída da rede para o fonema "a".

Cada neurônio de saída responde diferentemente ao sinal captado. Os padrões de cada fonema estão de alguma forma interpretados nas conexões que chegam aos neurônios de saída. Utilizamos como entrada da rede a energia advinda de bandas críticas de energia. Estes parâmetros fornecem informações cruciais no reconhecimento de padrões, contudo fonemas como “ê” e “ô” tem algumas dessas características em comum. Podemos ver na Fig. (4.17) que o neurônio de saída responsável pelo reconhecimento do fonema “ê” reconheceu alguns parâmetros do fonema “ô”

Os neurônios de saída reconhecem erroneamente alguns fonemas, porém percebe-se na Fig. (4.19) que o neurônio do fonema correspondente tem sua saída sobressalente. Na Figura (4.19) temos o fonema “ó” como entrada e o pulso nos fonemas “ó” e “é”, observa-se que o neurônio de saída “ó” permanece mais tempo acima do limiar determinado.

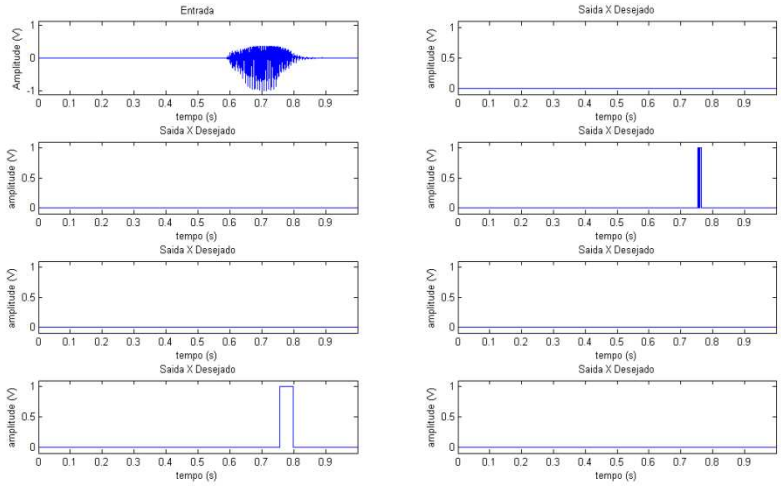

Figura 4.19 - Saída da rede para o fonema "ó".

Comparamos duas vogais separadamente. Os fonemas “a” e “u” foram falados e passados pela rede 20 vezes. Podemos visualizar a resposta da rede para o fonema “a” na Fig. (4.20) e o fonema “u” na Fig. (4.21). A taxa de acerto pode ser vista na Tab. (4.1).

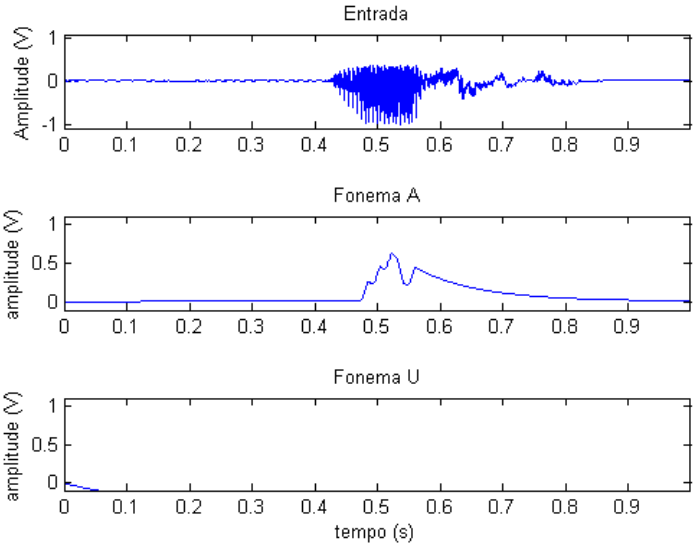

Figura 4.20 - Saída da rede para o fonema "a".

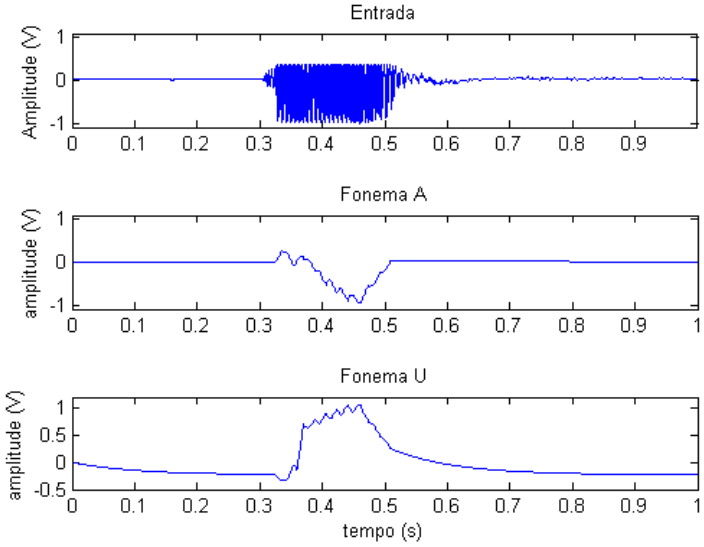

Figura 4.21 - Saída da rede para o fonema "u".

Tabela 4.1 - Taxa de acerto para "a" e "u".

Vogal	A	U
Taxa de acerto	75%	80%

A vogais pronunciadas separadamente tiveram reconhecimento satisfatório, porém quando passamos as palavras pela rede algumas situações indesejadas foram percebidas. Quando ocorria um ditongo a rede não reconhecia a semivogal. Podemos ver na Fig. (4.22) a palavra “dois”, a rede conseguiu reconhecer o fonema “ô”, porém a semivogal “i”, que se apóia na vogal “ô”, passa despercebida pela rede.

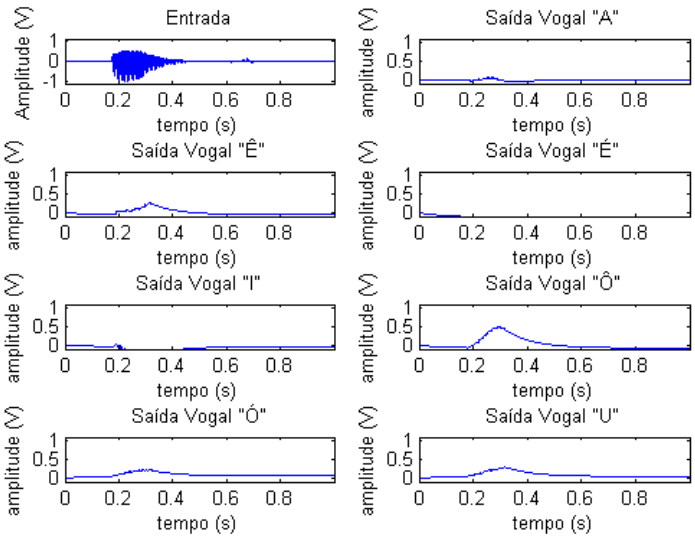

Figura 4.22 - Saída da rede para a palavra "dois".

As semivogais tem o mesmo som das vogais mas sua intensidade sonora é inferior a da vogal. O treinamento aconteceu com exemplos normalizados do fonema “i”. Quando a palavra foi normalizada, a diferença de intensidade entre vogal e semivogal permaneceu a mesma. Com isso a amostra do fonema “i” não teve intensidade suficiente para sensibilizar a saída.

5 Considerações Finais

5.1 Conclusões

Este projeto explorou todas as etapas do desenvolvimento de um sistema de reconhecimento de palavras, desde o estudo dos fonemas que serviriam de base de dados até o estudo de seu desempenho, passando por detalhes como a estrutura da rede neural e o sistema de classificação para tomar decisões sobre a palavra pronunciada a partir dos fonemas que foram reconhecidos. Um estudo inicial dos fonemas foi de grande importância para que se compreendesse o comportamento da rede perante alguns resultados.

A ferramenta utilizada foi o MATLAB, e o seu desempenho foi considerado satisfatório. Todo o trabalho foi organizado por meio de matrizes, e por isso a ferramenta proporcionou muitas facilidades. Alguns laços, porém, foram inevitáveis, o que tornou o desempenho do código possivelmente inferior ao que poderia ser desenvolvido em uma linguagem compilada.

Apesar de sistemas de reconhecimento da fala humana já serem conhecidos, e até mesmo utilizados comercialmente, procuramos desenvolver um estudo com base em uma rede neural pulsada, com pouco pré-processamento, tendo em vista que o sinal de voz é variável com o tempo, então é lógico ter um processador que atua de acordo com esta variável. Durante o desenvolvimento do projeto, foi considerada a distribuição aleatória dos pesos sinápticos, já que as conexões entre os neurônios possuem intensidades diferentes.

Foi testado, durante a etapa de treinamento da rede, o algoritmo iterativo de minimização de erro quadrático, e sua eficiência pôde ser confirmada. Pudemos compreender a importância de seguir os fundamentos deste algoritmo, como a inicialização aleatória dos pesos na saída e a escolha do coeficiente de aprendizagem, evitando, por isso, a não convergência ou os mínimos locais.

O sistema de classificação não pode ser utilizado. Como o desempenho da rede perante as palavras não foi satisfatório, a utilização do sistema de classificação tornou-se sem propósito.

5.2 Trabalhos futuros

Uma possível continuação do presente trabalho seria melhorar o desempenho da rede, focando no aprimoramento do sistema de reconhecimento. Apresentar para a rede, no momento do treinamento, sons nasais como o “u”, observado na palavra “um”, e também semivogais como o “i”, observado na palavra “dois”.

Outra possibilidade seria explorar o sistema de classificação que foi desenvolvido, porém não profundamente utilizado. Verificar seu potencial em reconhecer palavras à partir do momento em que o sistema de reconhecimento estiver mais robusto.

O estudo das consoantes e sua utilização no sistema que foi desenvolvido é um caminho natural que o projeto pode trilhar, tornando o sistema mais robusto e possibilitando, naturalmente, a expansão do banco de palavras.

Um caminho que também pode ser seguido no aprimoramento do trabalho é a otimização do código, através da substituição de *laços* por funções com processamento mais rápido, ou a utilização de *Toolboxes* que são disponibilizadas pelo fabricante do MATLAB. Outra alternativa para alcançar este objetivo seria a migração do código para uma linguagem compilada.

6 Referências Bibliográficas

1. SILVA, A. G. D. **Reconhecimento de Voz para Palavras Isoladas**. Recife. 2009.
2. BRAGA, L. P. **Reconhecimento de voz dependente de locutor utilizando Redes Neurais Artificiais**. Recife. 2006.
3. **interaula**. Disponivel em: <http://www.interaula.com/versao1.3/portugues/por0000-1-1.htm>. Acesso em: 25 outubro 2010.
4. Disponivel em: <http://www.inf.ufrgs.br/~danielnm/docs/exame_quali_daniel_muller.pdf>.
5. CUADROS, C. D. R. et al. **Comparação entre as técnicas de MFCC e ZCPA para reconhecimento robusto do locutor em ambientes rudosos**. Rio de Janeiro, RJ.
6. BRAGA, A. D. P.; LUDERMIR, T. B.; CARVALHO, A. C. D. L. F. **Redes Neurais Artificiais:** Teoria e Aplicações. Rio de Janeiro: LTC-LIVROS TÉCNICOS E CIENTÍFICOS EDITORA S.A., 2000.
7. Disponivel em: <http://amora2008cerebro.pbworks.com/Para-que-serve-os-neuronios>. Acesso em: 02 set. 2010.
8. RUIA, L. R. **FÍSICA: SOM E AUDIÇÃO HUMANA**. Rio Grande do Sul.
9. LATHI, B. P. **Modern Digital and Analog Communication Systems**. New York, EUA: Oxford Press, 1998.
10. FANT, G. **Acoustic theory of speech production: with calculations based on X-Ray studies of Russian**. [S.l.]. 1960.
11. TIMOSZCZUK, A. P. **Reconhecimento Automático do Locutor com Redes Neurais Pulsadas**. São Paulo. 2004.
12. HAYKIN, S. **Redes Neurais:** Princípios e prática. São Paulo, SP: Bookman Companhia Editora, 2001.