

XXVIII CNMAC

USANDO REDES NEURAIS PARA CLASSIFICAÇÃO DE PADRÕES DE VOZ

Alexandre de Souza Brandão

Universidade Federal Fluminense
PGMEC – Programa de Pós-Graduação em Engenharia Mecânica
Rua Passo da Pátria, 156 - 24120-240
São Domingos Niterói – RJ – Brasil
brandaoalexandre@ig.com.br

Edson Cataldo

Universidade Federal Fluminense
Departamento de Matemática Aplicada, Centro, Niterói, Brasil
PGMEC – Programa de pós-graduação em Engenharia Mecânica
Programa de pós-graduação em Engenharia de
Telecomunicações
Rua Mário Santos Braga, s/ No - 24020-140
Centro – Niterói – RJ – Brasil
ecataldo@zipmail.com

Fabiana Rodrigues Leta

Universidade Federal Fluminense
Departamento de Engenharia Mecânica
PGMEC – Programa de Pós-Graduação em Engenharia Mecânica
Rua Passo da Pátria, 156 - 24120-240 - São Domingos – Niterói – RJ- Brasil
fabiana@vm.uff.br

Jorge Carlos Lucero

Universidade de Brasília
Departamento de Matemática
DF 70910-900
lucero@mat.unb.br

Resumo: *Redes neurais artificiais (RNA) são modelos computacionais com propriedades particulares, tais como a habilidade de aprender, de generalizar, de agrupar e de organizar dados. Uma de suas aplicações é a classificação de padrões. Neste artigo, serão usadas redes neurais artificiais para classificação de padrões de voz obtidos através de medições acústicas realizadas sobre sinais de voz digitalizados. Primeiro, classificam-se sinais de vozes normais e com uma das seguintes características de patologia: nódulo nas cordas vocais ou paralisia unilateral das cordas vocais. Depois, realiza-se uma comparação entre padrões de vozes naturais e de vozes sintetizadas. As vozes sintetizadas foram obtidas usando-se modelos mecânicos das cordas vocais e do trato vocal.*

Palavras-chave: Redes neurais artificiais, Modelos mecânicos, Classificação de padrões, Síntese de vogais, Medidas acústicas.

1. INTRODUÇÃO

Uma das principais motivações para o entendimento do mecanismo da produção de voz está no fato de que a voz humana é um dos principais meios de comunicação.

A produção da voz se inicia com uma contração-expansão dos pulmões. Cria-se, assim, uma diferença entre a pressão do ar nos pulmões e a pressão do ar na

frente da boca, causando um escoamento de ar. O escoamento passa pela laringe e, antes homogêneo, vai se transformando em uma série de pulsos (conhecidos como trem de pulsos ou sinal glotal) de ar que chegam na boca e na cavidade nasal. Os pulsos de ar são modulados pela língua, pelos dentes e lábios; isto é, pela geometria destes órgãos, de forma a produzir o que conhecemos por voz. O sinal glotal, porém, possui propriedades importantes de difícil reprodução que estão intimamente ligadas às características anatômicas e fisiológicas da laringe.

Atualmente, a teoria mais aceita para a descrição do sinal glotal é a teoria chamada de aerodinâmica-mioelástica. Esta teoria postula que o movimento, de abrir e fechar, das cordas vocais é regido pelas propriedades mecânicas dos tecidos musculares que constituem, principalmente, as cordas vocais e pelas forças aerodinâmicas que se distribuem ao longo da laringe durante a fonação. A ação neural consiste apenas em aproximar as cordas vocais de tal forma que a superfície destas vibrem.

Para facilitar o estudo do sistema de produção da voz, normalmente ele é reduzido a quatro grupos distintos, em relação à onda sonora que é produzida ou modificada pelos órgãos. O primeiro grupo, chamado de grupo da *respiração*, corresponde à produ-

ção de um fluxo de ar; que se inicia nos pulmões e termina no final da traquéia. Na faringe, encontram-se os órgãos do segundo grupo, responsáveis pela produção do sinal glotal, chamado de grupo da *vocalização*. O sinal glotal é um sinal de baixa intensidade, que necessita ser amplificado e que determinadas componentes harmônicas sofram "ênfase", de maneira que os fonemas sejam caracterizados. Este é o grupo chamado de grupo de *ressonância*. Esse fenômeno ocorre na passagem do ar pelo chamado trato vocal (porção que vai da laringe até a boca). Finalmente, as ondas de pressão são irradiadas quando chegam à boca. Esse grupo é chamado de grupo da *irradiação*.

Na produção de vogais, o fluxo de ar proveniente dos pulmões é interrompido pela vibração quase-periódica das cordas vocais, conforme ilustrado na figura (1).

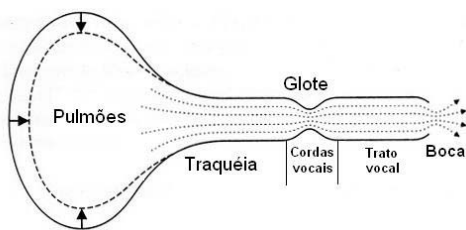


Figura 1 – Representação esquemática do sistema de produção da voz (adaptado de Titze [1]).

A partir da representação esquemática do sistema de produção de voz, pode-se observar os quatro grupos destacados anteriormente: respiração (pulmões e traquéia), vocalização (cordas vocais), ressonância (trato vocal) e radiação (boca).

Nas últimas décadas, a dinâmica das cordas vocais tem sido extensivamente estudada e alguns modelos mecânicos foram desenvolvidos. Esses modelos diferem pela representação que fazem das cordas vocais, vistas como sistemas mecânicos que modulam a passagem do ar.

Neste trabalho, usamos vozes normais (gravadas e posteriormente digitalizadas) e vozes sintetizadas. Para as vozes sintetizadas, usamos modelos mecânicos e um programa, ambos discutidos em [5,6].

2. MEDIÇÕES ACÚSTICAS SOBRE SINAIS DE VOZ

Uma das principais formas de se caracterizar o

sinal de voz é através de medições acústicas, pelo fato de que tais medidas podem revelar importantes características fisiológicas através de variações de seus valores [2,3]. Quando são obtidos conjuntos de medições acústicas aproximadamente iguais os chamados são obtidos *padrões de voz*. Uma vez obtidos os padrões de voz, *redes neurais artificiais (RNA)* podem ser usadas para classificá-los.

As medições acústicas descritas neste artigo foram obtidas através do programa de análise fonética PRAAT[8], de livre distribuição.

Os pulsos de ar do sinal glotal são formados entre os instantes de fechamento das cordas vocais, sendo o intervalo de tempo entre eles equivalente ao período, ou ciclo, do sinal de voz. A frequência fundamental do sinal de voz (inverso do período) é chamada de *pitch* e a variação dos intervalos de tempo que representam os pulsos do sinal de voz é chamada *jitter* [2,3].

A variação da amplitude do sinal de voz, obtida através da diferença entre a amplitude do sinal de voz obtida em um período e a do período seguinte é chamada *shimmer* [2,3].

Assim, *jitter* e *shimmer* são medidas da variação do período e da amplitude do sinal, ciclo-a-ciclo, respectivamente.

Há vários tipos de medidas de *jitter* e *shimmer*. As medidas que foram usadas neste trabalho serão descritas a seguir.

2.1. Medidas de Jitter

Jitter (local):

$$\text{Jitter}_{(\text{local})} = \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |P_{i+1} - P_i|}{\frac{1}{N} \sum_{i=1}^N P_i} \quad (1)$$

onde N é o número de amostras e P_i e P_{i+1} são intervalos de tempo consecutivos do sinal de voz.

Jitter (rap) e Jitter (ppq5):

$$Jitter_{(rap)} = \frac{\frac{1}{N-J+1} \sum_{i=1}^{N-J+1} \left| \frac{1}{J} \left(\sum_{j=1}^J P_{i+j-1} \right) - P_{i+0.5(J-1)} \right|}{\frac{1}{N} \sum_{i=1}^N P_i}$$

onde N é o número de amostras, P_i e P_{i+1} são intervalos de tempo consecutivos do sinal de voz e J é o tamanho de uma janela de tempo.

No caso de *Jitter (rap)* o valor de J é 3 e para o *Jitter (ppq5)* valor de J é 5.

Jitter (ddp):

$$Jitter_{(ddp)} = \frac{\frac{1}{N-3} \sum_{i=1}^{N-3} \left| |P_{i+3} - P_{i+2}| - |P_{i+1} - P_i| \right|}{\frac{1}{N} \sum_{i=1}^N P_i} \quad (3)$$

onde N é o número de amostras, P_i , P_{i+1} , P_{i+2} e P_{i+3} são intervalos de tempo consecutivos do sinal de voz.

2.2. Medidas de Shimmer

Shimmer (local):

$$Shimmer_{(local)} = \frac{\frac{1}{N-1} \sum_{i=1}^N |A_{i+1} - A_i|}{\frac{1}{N} \sum_{i=1}^N A_i} \quad (4)$$

onde N é o número de amostras e A_i e A_{i+1} são amplitudes do sinal de voz correspondentes a intervalos de tempo consecutivos.

Shimmer (apq3), Shimmer (apq5) e Shimmer (apq11):

$$Shimmer_{(apq3)} = \frac{\frac{1}{N-J+1} \sum_{i=1}^{N-J+1} \left| \frac{1}{J} \left(\sum_{j=1}^J A_{i+j-1} \right) - A_{i+0.5(J-i)} \right|}{\frac{1}{N} \sum_{i=1}^N A_i}$$

(5)

onde N é o número de amostras e A_i e A_{i+1} são amplitudes do sinal de voz correspondentes a intervalos de tempo consecutivos.

No caso de *Shimmer apq3* o valor de J é 3, no caso de *Shimmer apq5* o valor de J é 5 e no caso de *Shimmer apq11* o valor de J é 11.

Shimmer(ddd):

$$Shimmer_{(ddd)} = \frac{\frac{1}{N-3} \sum_{i=1}^{N-3} \left| |A_{i+3} - A_{i+2}| - |A_{i+1} - A_i| \right|}{\frac{1}{N} \sum_{i=1}^N A_i} \quad (6)$$

onde N é o número de amostras e A_i , A_{i+1} , A_{i+2} e A_{i+3} são as amplitudes do sinal de voz correspondentes a intervalos de tempo consecutivos.

2.3. Pitch (frequência fundamental média)

Além das medidas de Jitter e Shimmer temos ainda o *Pitch* que é a frequência cujo valor corresponde ao recíproco da média dos intervalos de tempo considerados.

$$Pitch_{(médio)} = \frac{1}{\frac{1}{N} \sum_{i=1}^N P_i} \quad (7)$$

3. RESULTADOS DAS MEDIÇÕES ACÚSTICAS EM VOZES NATURAIS

Foram usadas como amostras dez sinais de voz obtidos de pacientes com padrão de voz normal, doze sinais de voz obtidos de pacientes com padrão de voz afetado por nódulo nas cordas vocais e oito sinais de voz obtidos de pacientes com padrão de voz afetado por paralisia unilateral (paralisia em uma das cordas vocais).

O fonema utilizado para análise foi a vogal sustentada /e/. Esta vogal foi escolhida para as medições por apresentar mais nítidas as medições descritas na seção anterior.

As tabelas simplificadas I, II e III exemplificam os conjuntos de medições obtidas da forma explicada acima. Com relação à nomenclatura das

amostras, SN1 significa Sinal Normal n° 1, SPN1 significa Sinal Patologia Nódulo n° 1 e SPP1 significa Sinal Patologia Paralisia n° 1. Os diferentes números são apenas para identificar as amostras.

Tabela 1 – Medições para vozes normais.

Arquivo (*.wav)	SN1	SN2	SN10
Jitter (local):	0,27%	0,32%	0,26%
Jitter (rap):	0,14%	0,18%	0,16%
Jitter (ppq5):	0,15%	0,19%	0,15%
Jitter (ddp):	0,42%	0,54%	0,47%
Shimmer (local):	1,27%	4,48%	2,83%
Shimmer (apq3):	0,70%	2,51%	1,62%
Shimmer (apq5):	0,79%	2,87%	1,74%
Shimmer (apq11):	0,92%	3,11%	1,92%
Shimmer (dda):	2,11%	7,53%	4,85%
Intensidade (dB):	69,08	74,44	63,32
Pitch (Hz):	188,6	123,11	199,38

Tabela 2 – Medições para vozes com nódulo.

Arquivo (*.wav)	SPN1	SPN2	SPN12
Jitter (local):	0,23%	0,78%	0,41%
Jitter (rap):	0,13%	0,47%	0,22%
Jitter (ppq5):	0,13%	0,47%	0,30%
Jitter (ddp):	0,38%	1,41%	0,65%
Shimmer (local):	2,17%	4,31%	3,45%
Shimmer (apq3):	1,16%	2,38%	1,84%
Shimmer (apq5):	1,33%	2,77%	2,03%
Shimmer (apq11):	1,64%	2,90%	2,64%
Shimmer (dda):	3,49%	7,13%	5,53%
Intensidade (dB):	68,34	58,83	73,74
Pitch (Hz):	198,05	163,27	207,49

Tabela 3 – Medições para vozes com paralisia.

Arquivo (*.wav)	SPP1	SPP2	SPP8
Jitter (local):	0,60%	0,45%	2,28%
Jitter (rap):	0,34%	0,26%	1,38%
Jitter (ppq5):	0,36%	0,27%	1,61%
Jitter (ddp):	1,01%	0,76%	4,13%
Shimmer (local):	4,50%	4,64%	11,53%
Shimmer (apq3):	2,61%	2,59%	6,61%
Shimmer (apq5):	2,91%	3,05%	7,34%
Shimmer (apq11):	3,19%	3,16%	8,21%
Shimmer (dda):	7,82%	7,77%	19,84%
Intensidade (dB):	75,41	63,23	64,07
Pitch (Hz):	156,78	140,33	172,6

4. REDES NEURAIS ARTIFICIAIS

As redes neurais artificiais [7] vêm ganhando espaço no ramo da modelagem matemática, provando serem sistemas capazes de resolver problemas em situações onde é difícil criar modelos explícitos. As RNA possuem a habilidade de inferir relações não-lineares complexas e também a capacidade de resolver os problemas sem a necessidade de definição de listas de regras. Neste contexto, a proposta deste artigo é

mapear as características não-lineares das medições acústicas através de redes neurais artificiais, buscando a classificação entre três tipos de padrões de voz (normal, nódulo e paralisia) e, também, a comprovação de que a voz sintetizada pelo programa SINTESE desenvolvido por Cataldo e Nicolato [2,3] possui um padrão que pode ser classificado como normal.

Uma rede neural artificial consiste num modelo computacional em camadas, de nódulos (neurônios) processadores, interligadas por conexões ponderadas (pesos). Cada neurônio aplica uma função, chamada função de ativação, ao somatório de suas entradas mais um polarizador (conexão ponderada de entrada unitária).

Neste experimento utiliza-se uma rede neural artificial alimentada adiante (*feed forward network*) desenvolvida no ambiente MATLAB. Em redes neurais deste tipo, os dados fluem das unidades de entrada para as unidades de saída de forma direta, ou seja, não há realimentação.

A Figura (2) mostra a RNA projetada para este trabalho.

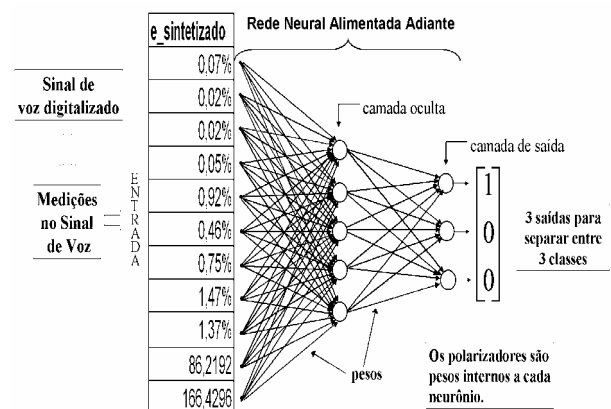


Figura 2 – Rede Neural Artificial.

O algoritmo de aprendizado da RNA é o de retropropagação e consiste, basicamente, no seguinte: dado um conjunto de pesos e polarizadores inicial, a rede neural recebe como amostras os valores das medições descritas anteriormente (na forma de um vetor de entrada). Cada vetor de entrada representa as características que se deseja distinguir entre as amostras de voz. Durante o treinamento da rede, o resultado correto para cada entrada, é mostrado para a rede, ou seja, os pesos da

rede são ajustados, através da *regra delta* [7], de modo que a rede mostre o resultado correto para cada uma das entradas conhecidas. Os pesos são ajustados após cada apresentação de um determinado vetor de entrada conhecido. A função de ativação precisa ser diferenciável (pois o algoritmo de retropropagação é baseado na derivada desta função) e não-decrescente (para que sua derivada não troque de sinal, pois isso comprometeria a convergência do algoritmo). Os neurônios da camada oculta possuem funções de ativação do tipo sigmóide (possuem a forma $f(v) = 1/(1 + e^{-av})$ $a > 0$) [7], formando combinações de não-linearidades para possibilitar o mapeamento de relações não-lineares. Os neurônios da camada de saída possuem funções de ativação do tipo linear, pois, neste caso, não é desejável o efeito de saturação da função sigmóide sobre os valores da saída. Maiores detalhes sobre redes neurais e o algoritmo de retropropagação podem ser encontrados em [7].

O ajuste da rede é testado colocando-se vetores de entrada que não foram usados para treinar a rede. A idéia é que vozes de um determinado padrão tenham vetores de medições com distâncias euclidianas pequenas entre si. Desse modo, vetores de entrada que não foram usados para treinar a rede, mas que pertencem ao mesmo padrão, serão interpolados entre os vetores que a rede conhece e classificados como tais.

A rede é treinada com vetores-alvo binários S , com componentes S_k ; $k = 1, 2, 3$, cujas componentes são dadas por:

$S_k = 1$ - quando a amostra pertence à classe C_k

$S_k = 0$ - quando a amostra não pertence à classe C_k

A partir desta notação, na rede projetada, a classe é representada por um vetor alvo de dimensão três, descrito da seguinte forma:

$$S = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \text{ padrão normal} \quad ; \quad S = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \text{ padrão}$$

nódulo

$$S = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \text{ padrão paralisia}$$

Uma rede neural classificadora de múltiplas camadas treinada com o algoritmo de retropropagação, com um conjunto finito de exemplos independentemente e identicamente distribuídos (i.i.d.), leva a uma aproximação assintótica das probabilidades de classe a posteriori subjacentes, desde que o tamanho do conjunto de treinamento seja suficientemente grande e que o processo de aprendizagem por retropropagação não fique preso em um mínimo local da função de erro do valor de saída da rede neural em relação ao valor da saída desejada. O número de neurônios na camada oculta depende de quão bem se deseja que a rede mapeie ou absorva a variabilidade do conjunto de amostras disponível para treinamento.

Como, no nosso caso, não se conhece, a priori, e a variabilidade do conjunto de amostras e o banco de amostras de voz conta, por enquanto, com apenas trinta amostras, este número de neurônios foi definido experimentalmente. Ou seja, a rede foi treinada inicialmente com três neurônios na camada oculta, depois com quatro, depois com cinco, e assim por diante observando-se o índice de acertos em relação às amostras deixadas de fora do treinamento. A configuração de camada oculta com cinco neurônios foi aquela em que se obteve maior índice de acertos. A idéia é começar com um número reduzido de neurônios na camada oculta parte do princípio de que o número de neurônios deve ser sempre pequeno em relação ao número de amostras para não sobre-parametrizar o modelo da rede (caso em que não se obtém melhoria apesar do aumento do número de neurônios) [7].

Foi utilizada apenas uma camada oculta, pois se pode provar que uma rede neural com apenas uma camada oculta é capaz de realizar qualquer mapeamento entrada-saída.

O número de amostras deixadas de fora do treinamento foi baseado na seguinte relação [9]:

$$r_{ótimo} = 1 - \frac{\sqrt{2W - 1} - 1}{2(W - 1)} \quad (8)$$

onde $r_{ótimo}$ representa a fração do conjunto de amostras disponível que será destinada ao treinamento da rede neural e W é o total de pesos da rede. Para o projeto da rede neste trabalho os dados são mostrados na tabela 4.

Tabela 4 – Total de pesos da rede neural projetada

Pesos da entrada	Vetores de 11 componentes \times 5 neurônios ocultos	55
Pesos da saída	5 neurônios ocultos \times 3 neurônios saída	15
Polarizadores	1 por neurônio (5 ocultos e 3 saída)	8
	Total de pesos (W)	78

Substituindo-se $W=78$ em (8) obtém-se $r_{ótimo} = 0.9256$. Assim 92,56% das amostras serão para treinar a rede e o restante para teste. Como há trinta amostras, vinte e sete delas serão utilizadas para treinamento e três para teste.

A rede neural artificial foi treinada com duzentas épocas de treinamento, ou seja, o algoritmo de retropropagação se repete duzentas vezes ou até que ocorra algum dos seguintes critérios de parada:

- Erro atingiu uma tolerância pré-determinada.
- Erro cai a uma taxa suficientemente pequena.

A cada *época* (apresentação completa de todo o conjunto de amostras de treinamento para a RNA), testa-se a generalização reparando-a, se a mesma for suficientemente boa. A tabela 5 apresenta um resumo da Rede Neural Artificial criada.

Tabela 5 – Resumo da Rede Neural Artificial

Tipo	Alimentada adiante
Algoritmo	retro-propagação
Camadas ocultas	1
Neurônios na camada oculta	5
Neurônios na camada de saída	3
Função de ativação (camada oculta)	Sigmóide
Função de ativação (camada de saída)	Linear

5. RESULTADOS OBTIDOS

5.1. Classificação de padrões de vozes naturais (não sintetizadas)

Conforme foi dito na seção 3, do total de trinta amostras, correspondentes à soma das amostras apresentadas nas tabelas 1, 2 e 3, vinte e sete foram usadas no treinamento da rede neural e três foram deixadas para testes. Assim, foram extraídas as colunas: 5 da tabela 1, 8 da tabela 2 e 7 da tabela 3. Estas colunas/amostras foram escolhidas por apresentarem valores em algumas de suas componentes mais acentuados em relação às suas respectivas tabelas/classes. Isto foi feito como uma forma de testar a capacidade de generalização da rede neural artificial, ou seja, oferecer um pouco mais de dificuldade ao teste da RNA. Porém, as três colunas para o teste da RNA também poderiam ter sido retiradas aleatoriamente. A tabela 6 apresenta a classificação correta para valores não usados no treinamento.

Tabela 6 – Classificação correta para valores não usados no treinamento.

Medições	SN5	SPN8	SPP7	alexandre
Jitter (local):	0,7	1,28	4,49	0,28
Jitter (rap):	0,39	0,72	2,59	0,11
Jitter (ppq5):	0,48	0,94	2,82	0,15
Jitter (ddp):	1,17	2,15	7,76	0,34
Shimmer (local):	3,27	7,6	31,5	4,18
Shimmer (ppq3):	1,73	4,28	19,11	2,25
Shimmer (ppq5):	1,94	4,89	17,34	2,56
Shimmer (ppq11):	2,48	5,63	20,08	2,99
Shimmer (dda):	5,2	12,85	57,32	6,73
Intensidade (dB):	56,76	60,55	58,43	73,48
Pitch (Hz):	181,51	171,47	81,43	128,12
RNA				
Saída 1	3,846963059	-9,26E-07	-5,58E-07	1,000001036
Saída 2	-2,846966271	0,838263978	4,17E-09	-2,46E-07
Saída 3	3,21E-06	0,161736948	1,000000554	-7,91E-07
Resultado	normal	nódulo	paralisia	normal

Uma vez treinada a rede, seus resultados devem aproximar-se dos vetores alvos correspondentes ao seu padrão, ou seja, cada componente de um vetor saída deve ser aproximadamente 0 ou 1 sendo o limiar de decisão 0,5. Por exemplo, o resultado da rede a uma entrada para ser entendido como normal deve ter a primeira componente do vetor coluna maior que 0,5 e as outras duas menores que 0,5.

5.2. Classificação de padrões de vozes sintetizadas

Paralelamente, foram criadas amostras de voz sintetizadas (vogal /e/ sustentada) através do

programa SINTESE [6], variando a pressão dos pulmões, variando a tensão nas cordas vocais, variando ambos e sem qualquer variação. Medidas acústicas foram obtidas desses quatro sinais de voz sintetizados. Usamos as opções do programa SINTESE que considera, para as cordas vocais, um sistema mecânico (massa-mola-amortecedor) de um grau de liberdade e de dois graus de liberdade, conforme ilustrado na Figura (3).

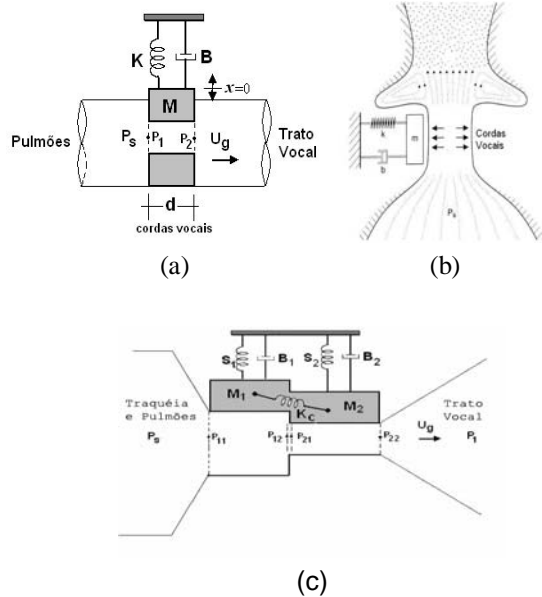


Figura 3 – (a) Modelo mecânico de segunda ordem para as cordas vocais. (b) Sistema vocal para o Modelo de Flanagan e Landgraf [4]. (c) Modelo mecânico de dois graus de liberdade para as cordas vocais. (adaptado de Titze [1]).

Para o trato vocal, o programa considera um circuito acústico considerando o trato dividido em vários tubos cilíndricos.

Essa opção foi escolhida, pois Cataldo et al [6] mostra que é possível obter síntese de vogais, de ótima qualidade, considerando, para as cordas vocais sistemas de baixa ordem, variando, porém, corretamente, determinados parâmetros em relação ao tempo. No programa, foram considerados quatro casos. Dois modelos para as cordas vocais foram usados e para cada modelo foram considerados dois casos: com variação da pressão subglotal e sem variação da pressão subglotal. Os detalhes podem ser obtidos em [6]. A Tabela 7 mostra os resultados das medições para as amostras sintetizadas e a Tabela 8 mostra a

classificação correta para as quatro vozes sintetizadas.

Tabela 7 – Resultados das medições para as amostras sintetizadas.

Arquivo (*.wav)	e_sintetizado1	e_sintetizado2	e_sintetizado3	e_sintetizado4
Jitter (local):	0,07%	1,30%	0,95%	0,02%
Jitter (rap):	0,02%	0,10%	0,09%	0,01%
Jitter (ppq5):	0,02%	0,19%	0,15%	0,02%
Jitter (ddp):	0,05%	0,29%	0,27%	0,03%
Shimmer (local):	0,92%	2,24%	2,08%	0,72%
Shimmer (apq3):	0,46%	0,93%	0,69%	0,44%
Shimmer (apq5):	0,75%	1,26%	1,13%	0,72%
Shimmer (apq11):	1,47%	2,17%	2,15%	1,44%
Shimmer (dda):	1,37%	2,79%	2,06%	1,31%
Intensidade (dB):	86,21	83,84	85,2	86,43
Pitch (Hz):	166,42	154,14	159,41	166,95

Tabela 8 – Classificação correta para as 4 vozes sintetizadas.

Medições	e_sintetizado1	e_sintetizado2	e_sintetizado3	e_sintetizado4
Jitter (local):	0,07	1,3	0,95	0,02
Jitter (rap):	0,02	0,1	0,09	0,01
Jitter (ppq5):	0,02	0,19	0,15	0,02
Jitter (ddp):	0,05	0,29	0,27	0,03
Shimmer (local):	0,92	2,24	2,08	0,72
Shimmer (apq3):	0,46	0,93	0,69	0,44
Shimmer (apq5):	0,75	1,26	1,13	0,72
Shimmer (apq11):	1,47	2,17	2,15	1,44
Shimmer (dda):	1,37	2,79	2,06	1,31
Intensidade (dB):	86,2192	83,8494	85,2	86,43
Pitch (Hz):	166,4296	154,14	159,41	166,95
RNA				
Saída 1	1,000001036	1,000001036	1,000001036	1,000001036
Saída 2	-1,94E-06	-2,46E-07	-2,47E-07	-3,22E-06
Saída 3	9,04E-07	-7,90E-07	-7,89E-07	2,19E-06
Resultado	normal	normal	normal	normal

6. CONCLUSÕES

Uma rede neural artificial foi modelada para classificar padrões de vozes: normais, com nódulo nas cordas vocais e com paralisia unilateral e sintetizadas. Os resultados obtidos demonstram sua eficiência nesta classificação, conseguindo interpolar com sucesso as três entradas deixadas de fora do treinamento entre as entradas dos seus respectivos grupos.

A RNA treinada com este conjunto de amostras foi capaz de reconhecer corretamente três tipos de vozes (amostras deixadas de fora do treinamento) e amostras de vozes sintetizadas.

A precisão da RNA na sua tarefa de classificação depende da variabilidade dos valores das amostras usadas como exemplos para o seu treinamento, como também do número de amostras, da seguinte forma: o número de amostras deve ser não apenas suficiente, mas deve também ser composto por amostras, cujos valores estejam suficientemente distribuídos entre os valores possíveis do sistema

para permitir que a rede o modele com a precisão desejada. Isso justifica o fato de a RNA projetada neste artigo ter conseguido classificar com sucesso, apesar do número pequeno de amostras, ou seja, não se tem noção da variabilidade de cada grupo de vozes. Porém, provavelmente, as vinte e sete amostras usadas possuem valores suficientemente distribuídos para permitir que a rede interpole corretamente.

Outro fato a ser notado é que a quantidade de medições diferentes, que compõem o vetor de entrada da rede neural, aumenta a sua capacidade classificadora. Essas medições são equivalentes a atributos de um objeto (amostra de voz). Assim, quanto maior a quantidade de atributos para comparação, maior a capacidade de discernir entre duas classes de objetos.

Além disso, oferecemos mais uma ferramenta para auxiliar na determinação de se uma voz sintetizada pode ser considerada como normal.

O nosso próximo objetivo é modificar os modelos mecânicos, usados no programa SINTESE [6], para obter sinais de vozes sintetizadas com características de patologia das cordas vocais e usar redes neurais artificiais para classificá-los.

7. AGRADECIMENTOS

Agradecemos à Fonoaudióloga Lígia Mattos, pela atenção dispensada e por ter nos cedido as amostras de sinais de voz.

8. REFERÊNCIAS

- [1] TITZE, I. R., 1994, **Principles of Voice Production**, PrenticeHall, EnglewoodCliff New Jersey.
- [2] ROSA, M. O. **Análise Acústica da Voz para Pré-diagnóstico de Patologias da Laringe**. 219p. Dissertação (Mestrado em Engenharia Elétrica). Escola de Engenharia de São Carlos, USP, São Paulo, 1998.
- [3] VIEIRA, M. N. **Automated Measures of Dysphonias and the Phonatory Effects of Asymmetries in the Posterior Larynx**. Tese (Doutorado em Engenharia Elétrica). Universidade de Edinburg 1997.
- [4] ISHIZAKA, K. E FLANAGAN, J. **Synthesis of voiced sounds from two-mass model of the vocal cords**, Bell Syst. Tech. Journal, Vol. 51, pp. 1233-1268, 1972.
- [5] CATALDO E., LETA F. R., LUCERO J., NICOLATO L., **Voiced Sounds Synthesis Using Mechanical Models**. In: 11TH International Workshop on Systems, Signals and Image Processing, 2004, Póznan. Anais do IWSSIP'04. 2004.
- [6] CATALDO E., NICOLATO L., **Modelagem Matemática da Produção da Voz e sua Aplicação à Síntese Articulatoria**, Monografia de Iniciação Científica apresentada ao Instituto Nacional de Matemática Pura e Aplicada para o Prêmio de Iniciação Científica, 2004.
- [7] HAYKIN, S. **Redes Neurais Princípios e Prática**, 2ª Edição, Porto Alegre: Bookman, 2001.
- [8] PRAAT – **Programa para análise fonética**, www.praat.org.