UNIVERSIDADE DE SÃO PAULO
INTERUNIDADES EM BIOENGENHARIA – EESC/FMRP/IQSC

PAULA BELINI BARAVIEIRA

Aplicação de uma rede neural artificial para a avaliação
da rugosidade e soproidade vocal

São Carlos,
2016

PAULA BELINI BARAVIEIRA

**Aplicação de uma rede neural artificial para a avaliação
da rugosidade e soproidade vocal**

Tese de doutorado apresentada ao Programa de Pós-Graduação Interunidades em Bioengenharia - Escola de Engenharia de São Carlos / Faculdade de Medicina de Ribeirão Preto / Instituto de Química de São Carlos da Universidade de São Paulo como parte dos requisitos para a obtenção do título de doutora em Ciências.

Área de Concentração: Bioengenharia

Orientador: Prof. Dr. Arlindo Neto Montagnoli

Co-orientadora: Profa. Dra. Alcione Ghedini Brasolotto

Versão Corrigida

São Carlos,
2016

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO,
POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS
DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

B227 Belini Baravieira, Paula
/ Paula Belini Baravieira; orientador Arlindo Neto
Montagnoli; coorientadora Alcione Ghedini Brasolotto.
São Carlos, 2016.

Tese (Doutorado) - Programa de Pós-Graduação
Interunidades Bioengenharia e Área de Concentração em
Bioengenharia -- Escola de Engenharia de São Carlos;
Faculdade de Medicina de Ribeirão Preto; Instituto de
Química de São Carlos, da Universidade de São Paulo,
2016.

1. Qualidade da Voz. 2. Percepção Auditiva. 3.
Redes Neurais Artificiais. 4. Análise de Wavelet. 5.
Inteligência Artificial. 6. Disfonia. I. Título.

Programa de Pós-Graduação
Interunidades Bioengenharia
EESC / FMRP / IQSC

PAULA BELINI BARAVIEIRA

Título: "Aplicação de uma rede neural artificial para a avaliação da rugosidade e soprovidade vocal"

TESE APRESENTADA AO PROGRAMA DE PÓS-GRADUAÇÃO
INTERUNIDADES EM BIOENGENHARIA - EESC/FMRP/IQSC DA
UNIVERSIDADE DE SÃO PAULO PARA OBTENÇÃO DO TÍTULO
DE DOUTORA EM CIÊNCIAS NA ÁREA DE BIOENGENHARIA.

Defesa em 28/03/2016

Prof. Dr. Arlindo Neto Montagnoli (Orientador)

Resultado: Aprovada

Universidade Federal de São Carlos

Assinatura: Arlindo Neto Montagnoli

Profa. Dra. Kelly Cristina Alves Silvério

Resultado: APROVADA

Faculdade de Odontologia de Bauru / USP

Assinatura: Kelly Cristina Alves Silvério

Prof. Dr. Ricardo Augusto Souza Fernandes

Resultado: APROVADA

Universidade Federal de São Carlos

Assinatura: Ricardo Augusto Souza Fernandes

Profa. Dra. Maria Eugenia Dajer

Resultado: APROVADA

Universidade Federal Tecnológica do
Paraná

Assinatura: Maria Eugenia Dajer

Prof. Dr. Marcelo Suetake

Resultado: APROVADA

Universidade Federal de São Carlos

Assinatura: Marcelo Suetake

Data da homologação: ____/____/____

AGRADECIMENTOS

Agradeço primeiramente a Deus, que me inspirou e me inspira no caminho trilhado. À minha família que sempre está ao meu lado, me apoiando e dando suporte. Ao meu companheiro, André Arruda, pelo amor e apoio em todos os momentos.

Agradeço a todos os amigos que deixam minha vida mais alegre, cheia de amor e risadas. Aos meus orientadores, Arlindo Neto Montagnoli e Alcione Ghedini Brasolotto, que guiaram e ajudaram na execução deste trabalho, sendo muito mais do que orientadores, mas também amigos, conselheiros e mestres.

Agradeço às fonoaudiólogas Millena Vieira, Larissa Siqueira e Kelly Silvério que dedicaram muito de seu tempo como juízas desta pesquisa. À CAPES que financiou este projeto e à Bioengenharia/USP que me acolheu como doutoranda, abrindo um mundo de possibilidades na pesquisa interdisciplinar.

Agradeço à vida e ao universo por estes quatro anos de estudos, aprendizados, crescimento pessoal e profissional, e pela concretização de mais este sonho.

“Alguns homens veem as coisas como são, e dizem ‘Por quê?’ Eu sonho com as coisas que nunca foram e digo ‘Por que não?’”

(George Bernard Shaw)

RESUMO

BARAVIEIRA, P. B. **Aplicação de uma rede neural artificial para a avaliação da rugosidade e soproidade vocal.** 2016. 101 f. Tese (Doutorado) – Programa de Pós-Graduação Interunidades em Bioengenharia EESC/FMRP/IQSC, Universidade de São Paulo, 2016.

A avaliação perceptivo-auditiva tem papel fundamental no estudo e na avaliação da voz, no entanto, por ser subjetiva está sujeita a imprecisões e variações. Por outro lado, a análise acústica permite a reprodutibilidade de resultados, porém precisa ser aprimorada, pois não analisa com precisão vozes com disfonias mais intensas e com ondas caóticas. Assim, elaborar medidas que proporcionem conhecimentos confiáveis em relação à função vocal resulta de uma necessidade antiga dentro desta linha de pesquisa e atuação clínica. Neste contexto, o uso da inteligência artificial, como as redes neurais artificiais, indica ser uma abordagem promissora. **Objetivo:** Validar um sistema automático utilizando redes neurais artificiais para a avaliação de vozes rugosas e soprosas. **Materiais e métodos:** Foram selecionadas 150 vozes, desde neutras até com presença em grau intenso de rugosidade e/ou soproidade, do banco de dados da Clínica de Fonoaudiologia da Faculdade de Odontologia de Bauru (FOB/USP). Dessas vozes, 23 foram excluídas por não responderem aos critérios de inclusão na amostra, assim utilizaram-se 123 vozes. **Procedimentos:** avaliação perceptivo-auditiva pela escala visual analógica de 100 mm e pela escala numérica de quatro pontos; extração de características do sinal de voz por meio da Transformada *Wavelet Packet* e dos parâmetros acústicos: *jitter*, *shimmer*, amplitude da derivada e amplitude do *pitch*; e validação do classificador por meio da parametrização, treino, teste e avaliação das redes neurais artificiais. **Resultados:** Na avaliação perceptivo-auditiva encontrou-se, por meio do teste Coeficiente de Correlação Intraclassa (CCI), concordâncias inter e intrajuiz excelentes, com $p = 0,85$ na concordância interjuizes e p variando de 0,87 a 0,93 nas concordâncias intrajuiz. Em relação ao desempenho da rede neural artificial, na discriminação da soproidade e da rugosidade e dos seus respectivos graus, encontrou-se o melhor desempenho para a soproidade no subconjunto composto pelo *jitter*, amplitude do *pitch* e frequência fundamental, no qual obteve-se taxa de acerto de 74%, concordância excelente com a avaliação perceptivo-auditiva da escala visual analógica (0,80 no CCI) e erro médio de 9 mm. Para a rugosidade, o melhor subconjunto foi composto pela Transformada *Wavelet Packet* com 1 nível de decomposição, *jitter*, *shimmer*, amplitude do *pitch* e frequência fundamental, no qual obteve-se 73% de acerto, concordância excelente (0,84 no CCI), e erro médio de 10 mm. **Conclusão:** O uso da inteligência artificial baseado em redes neurais artificiais na identificação, e graduação da rugosidade e da soproidade, apresentou confiabilidade excelente (CCI > 0,80), com resultados semelhantes a concordância interjuizes. Dessa forma, a rede neural artificial revela-se como uma metodologia promissora de avaliação vocal, tendo sua maior vantagem a objetividade na avaliação.

Palavras-chave: Qualidade da Voz; Percepção Auditiva; Redes Neurais Artificiais; Análise de Wavelet; Inteligência Artificial; Disfonia.

ABSTRACT

BARAVIEIRA, P. B. **The use of an artificial neural network for evaluation of vocal roughness and breathiness.** 2016. 101 f. Tese (Doutorado) – Programa de Pós-Graduação Interunidades em Bioengenharia EESC/FMRP/IQSC, Universidade de São Paulo, 2016.

The auditory-perceptual evaluation is fundamental in the study and analysis of voice. This evaluation, however, is subjective and tends to be imprecise and variable. On the other hand, acoustic analysis allows reproducing results, although these results must be refined since the analysis is not precise enough for intense dysphonia or chaotic waves. Therefore, the will to develop measurements allowing reliable knowledge related to vocal function is not new on this research and clinical actuation field. In this context, the use of artificial intelligence such as neural networks seems to be a promising research field. **Objective:** to validate an automatic system using artificial neural networks for evaluation of vocal roughness and breathiness. **Methods:** One hundred fifty (150) voices were selected from from Clínica de Fonoaudiologia da Faculdade de Odontologia de Bauru (FOB/USP) database. These voices presented variation from neutral to intense roughness and/or breathiness. Twenty-three of them were excluded since they did not match inclusion criteria. Thus, 123 voices were used for analysis. The procedures include use of auditory-perception based on two scales: visual analog scale of 100mm and four points numerical scale. Additionally, the characteristics of voice signals were extracted by *Wavelet Packet Transform* and by analysis of acoustic parameters: jitter, shimmer, derivative amplitude and pitch amplitude. Validation of classifying system was carried out by parameterization, training, test and evaluation of artificial neural networks. **Results:** In the auditory-perceptual evaluation, excellent interrater ($p=0.85$) and intrarater ($0.87 < p < 0.93$) agreement were obtained by means of Intraclass Correlation Coefficient (ICC) testing. The artificial neural network performance has achieved the best results for breathiness in the subset composed by parameters jitter, pitch amplitude and fundamental frequency. In this case, the neural network obtained a rate of 74%, demonstrating excellent concordance with auditory-perceptual evaluation for visual analog scale (0.80 ICC) and mean error of 9mm. As for roughness evaluation, the best subset is composed by *Wavelet Packet Transform* with 1 resolution level, jitter, shimmer, pitch amplitude and fundamental frequency. For this case, a 73% rate was achieved (0.84 ICC) and mean error of 10mm was obtained. **Conclusion:** The use of artificial neural networks for roughness and breathiness evaluation present high reliability ($ICC > 0.80$), with results similar to interrater agreement. Thus, the artificial neural network reveals a promising method for vocal evaluation, bringing objective analysis as a strong advantage.

Keywords: Voice Quality; Auditory Perception; Artificial Neural Network; Wavelet Analysis; Artificial Intelligence; Dysphonia.

LISTA DE FIGURAS

Figura 1 - Representação da estrutura de um único neurônio artificial. Fonte: MacKay, 2005.....	21
Figura 2 - Representação da arquitetura de uma rede neural, modelo Multilayer perceptron, de duas camadas, onde cada nó representa um neurônio e cada linha um peso. Fonte: MacKay, 2005 (adaptado).....	22
Figura 3 - Arquitetura das RNAs empregadas na identificação dos tipos de vozes.	44
Figura 4 - Árvore binária de decomposição da Transformada Wavelet Packet.	45
Figura 5 – Exemplo do sinal de voz sem filtragem no nó (0,0), quadro B, e com filtragem por meio da técnica Wavelet Packet nos nós (4,0), quadro C, e (4,15) no quadro D.	46
Figura 6 – Observa-se no primeiro quadro a onda do sinal sonoro e no segundo quadro os picos evidenciados pela correlação, a partir dos quais se calcula a F_0 , o jitter e o shimmer.	48
Figura 7 – No primeiro quadro observa-se a onda do sinal sonoro e no segundo quadro os picos da derivada do sinal, a partir da qual se calcula a Amplitude da Derivada.	48
Figura 8 – No primeiro quadro observa-se a onda do sinal sonoro e no segundo quadro os picos da autocorrelação do sinal residual, sendo que a partir do segundo pico, circulado na imagem, se calcula a Amplitude do <i>Pitch</i>	49
Figura 9 - A janela A refere-se à frequência fundamental encontrada inicialmente pelo programa, a janela B a F_0 indicada pelo fonoaudiólogo e a parte C à F_0 encontrada pelo programa, a partir da frequência que foi indicada, com os períodos identificados corretamente na autocorrelação do sinal de voz.	49
Figura 10 – A parte A da figura exemplifica a metodologia encontrada para verificação da frequência fundamental, a F_0 encontrada inicialmente pelo programa de 82,6 Hz e a F_0 calculada manualmente de 166,7 Hz. Na parte B observa-se, após indicação da F_0 calculada pelo clínico, que o programa encontra a F_0 de 165,2 Hz e identifica corretamente os períodos do sinal de voz na autocorrelação do sinal.....	50
Figura 11 - Análise espectrográfica de uma voz feminina, crepitante intensa e sopro leve, revelando: sub-harmônicos assistemáticos na frequência de 80 Hz (marcados pelo símbolo +), frequência identificada inicialmente pela análise acústica como F_0 na figura 9; e frequência fundamental (F_0) em torno de 160 Hz, valor próximo ao encontrado no cálculo realizado pelo clínico no sinal de voz na figura 9.	51
Figura 12 - Diagrama de blocos da extração de características e fase de classificação.	52
Figura 13 – Voz com onda tipo III, avaliada pela perceptivo-auditiva com 94 de rugosidade (tipo aspereza) e 57 de sopro (escala de 0 a 100 mm). No quadro A tem-se o sinal da onda sonora, no quadro B a correlação, no quadro C a derivada do sinal e no quadro D a autocorrelação do sinal residual.	55

Figura 14 – Curva de regressão quadrática da percepção da rugosidade, soprosidade e da equação quadrática proposta pela interpolação dos dados.	58
Figura 15 – Régua com os valores utilizados para estimar-se a equação quadrática da diferença permitida, ou seja, com as notas de corte corrigidas, valores centrais e diferença permitida dentro de cada faixa de distribuição.	59
Figura 16 – Maior taxa de acerto ($\sqrt{5}y$) e confiabilidade (CCI), em três testes, da rede neural artificial de acordo com o banco de vozes utilizado para o treino e teste.	61
Figura 17 – Maior taxa de acerto ($\sqrt{5}y$) e confiabilidade (CCI), em três testes, da rede neural artificial para diferentes níveis de decomposição da Transformada Wavelet Packet (WPT) e parâmetros acústicos fixos: jitter, shimmer, derivada da amplitude, amplitude do pitch e frequência fundamental.	62
Figura 18 – Melhor taxa de acerto ($\sqrt{5}y$) e confiabilidade (CCI) da rede neural artificial, em três testes, para diferentes subconjuntos de parâmetros acústicos e sem os parâmetros extraídos da WPT.	64
Figura 19 – Melhor taxa de acerto ($\sqrt{5}y$) e confiabilidade (CCI) da rede neural artificial, em três testes, na parametrização com um nível de decomposição da Transformada Wavelet Packet juntamente com os três melhores subconjuntos de parâmetros acústicos na discriminação do grau e tipo das vozes.	66
Figura 20 – Avaliação da soprosidade pelos fonoaudiólogos, juízes da pesquisa, e pela rede neural artificial.	67
Figura 21 - Avaliação da rugosidade pelos fonoaudiólogos, juízes da pesquisa, e pela rede neural artificial.	68

LISTA DE TABELAS

Tabela 1 – Análise qualitativa dos valores do Coeficiente de Correlação Intraclassa - CCI (ROSNER, 1995).	41
Tabela 2 - Avaliação qualitativa do grau de correlação entre duas variáveis (CALLEGARI-JACQUES, 2003).	51
Tabela 3 – Notas de corte para os diferentes graus de rugosidade e soproidade com os respectivos valores de eficiência.	57
Tabela 4 - Faixas de distribuição dos graus de rugosidade e soproidade, em quatro pontos.....	57
Tabela 5 – Faixas de distribuição, dos graus de rugosidade e soproidade, a partir das notas de corte obtidas da equação quadrática proposta pela interpolação dos dados.	57
Tabela 6 – Variação do erro médio e da taxa de acerto da rede neural artificial de acordo com o banco de vozes utilizado no treino/teste, em três testes.....	61
Tabela 7 - Variação do erro médio e da taxa de acerto da RNA, em três treinos, no treino/teste variando o nível de decomposição da Transformada Wavelet Packet.	62
Tabela 8 - Variação do erro médio e da taxa de acerto da RNA, em três treinos/testes, com diferentes subconjuntos de parâmetros acústicos.....	65
Tabela 9 - Variação do erro médio e da taxa de acerto da RNA, em três treinos/testes, com uma nível de decomposição da Transformada Wavelet Packet juntamente com os três melhores subconjuntos de parâmetros acústicos na discriminação do grau e tipo das vozes.	66

LISTA DE SIGLAS

AVKI	Índice acústico de qualidade vocal
<i>BP</i>	<i>Backpropagation</i>
CCI	Coefficiente de Correlação Intraclassa
<i>CFS</i>	<i>Correlation-based Feature Selection</i>
<i>CPPS</i>	<i>Smoothed cepstral peak prominence</i>
DA	Amplitude da derivada
DSI	Índice de Severidade da disfonia
DWT	Transformada Discreta <i>Wavelet</i>
Dif_p	Diferença permitida
F_0	Frequência fundamental
<i>GNE</i>	<i>Glottal-to-noise excitation</i>
HNR	Proporção harmônico-ruído
MDVP	<i>Multi-Dimensional Voice Program</i>
mm	milímetros
<i>MLP</i>	<i>Multilayer perceptron</i>
PA	Amplitude do <i>Pitch</i>
RNA	Rede Neural Artificial
R	Voz rugosa
S	Voz soprosa
T	Período
WPT	Transformada <i>Wavelet Packet</i>

SUMÁRIO

1. INTRODUÇÃO	17
2. OBJETIVOS.....	19
3. CONSIDERAÇÕES TEÓRICAS	21
3.1 FUNDAMENTOS TEÓRICOS SOBRE A REDE NEURAL ARTIFICIAL.....	21
3.1.1 Reconhecimento de Padrões	23
4. REVISÃO BIBLIOGRÁFICA.....	25
4.1 O USO DOS SISTEMAS INTELIGENTES NA AVALIAÇÃO VOCAL	25
4.2 CORRESPONDÊNCIAS ENTRE A ANÁLISE ACÚSTICA E A AVALIAÇÃO PERCEPTIVO-AUDITIVA	33
5. MATERIAL E MÉTODOS.....	39
5.1 Casuística	39
5.2 Edição das gravações vocais	40
5.3 Avaliação perceptivo-auditiva.....	40
5.3.1 Curva de percepção-auditiva e diferença permitida.....	41
5.4 Descrição do sistema classificador.....	42
5.5 Parametrização por meio da Transformada Wavelet Packet.....	44
5.6 Parametrização por meio dos parâmetros acústicos	47
5.6 Validação do sistema classificador	52
6. RESULTADOS.....	55
6.1 Resultados da avaliação perceptivo-auditiva	56
6.2 Resultados da análise acústica.....	59
6.3 Resultados dos treinos e testes das Redes Neurais Artificiais.....	59
7. DISCUSSÃO.....	69

7.1 Perceptivo-auditiva e curva de percepção-auditiva.....	69
7.2 Análise acústica.....	71
7.3 Rede Neural Artificial	73
7.3.1 Soprosidade e sua relação com os parâmetros do melhor subconjunto de avaliação.....	73
7.3.2 Rugosidade e sua relação com os parâmetros de seu melhor subconjunto de avaliação.....	76
7.3.3 Desempenho da Rede Neural Artificial.....	77
8. CONCLUSÃO	81
REFERÊNCIAS¹	83
APÊNDICES.....	89
APÊNDICE A	91
APÊNDICE B	93
ANEXOS....	95
ANEXO A.....	97

1. INTRODUÇÃO

A avaliação perceptivo-auditiva da voz, com caracterização do grau e tipo de disfonia, é considerada o padrão-ouro da avaliação vocal realizada pelo fonoaudiólogo (MADAZIO; LEÃO; BEHLAU, 2011; MARYN; ROY, 2012; OATES, 2009). Por meio dela é possível descrever as características vocais e inferir sobre o impacto social da disfonia, visto que o julgamento da qualidade vocal é essencialmente perceptivo e intuitivo (MADAZIO; LEÃO; BEHLAU, 2011; MARYN; ROY, 2012; OATES, 2009; SHRIVASTAV, 2003). Contudo, por se tratar de uma avaliação subjetiva, está sujeita a erros e variações, mesmo entre avaliadores treinados, pois é influenciada por fatores como a experiência prévia do avaliador, conhecimento ou não dos dados de anamnese, tipo de tarefa avaliada, que pode ser de voz ou fala, (MARYN; ROY, 2012; ZRAICK; WENDEL; SMITH-OLINDE, 2005) e do protocolo utilizado (KREIMAN et al., 1993).

Na tentativa de diminuir as variações decorrentes da avaliação perceptivo-auditiva foram criadas e validadas escalas de avaliação, cuja contribuição foi desde a padronização de termos utilizados para representar cada característica vocal, até a padronização da aplicação clínica das tarefas vocais utilizadas e da forma de marcação do grau de alteração vocal (HIRANO, 1981). Dentre as escalas mais utilizadas estão a GRBAS, a CAPE-V e a escala visual-analógica do desvio vocal, as quais revelam índices bons de confiabilidade na avaliação perceptivo-auditiva (KARNELL et al., 2007; YAMASAKI et al., 2008). Além disso, pesquisas recentes encontraram que escalas contínuas, como a CAPE-V, possuem confiabilidade significativamente maior do que escalas numéricas, como a GRBAS (KARNELL et al., 2007; KREIMAN; GERRATT; ITO, 2007; YAMASAKI et al., 2008). No entanto, apesar de todos os esforços, a subjetividade continua sendo uma limitação da análise perceptivo-auditiva na avaliação vocal.

Com o intuito de suplantar estas limitações, investigam-se métodos objetivos de avaliação vocal, cuja principal vantagem é a reprodutibilidade dos resultados (FROHLICH; MICHAELIS; WERNER STRUBE, 1998). Roy et al. (2013) em uma revisão sistemática sobre os métodos de avaliação vocal revelaram, dentre 100 artigos analisados, que a maioria (60%) estudou medidas acústicas, enquanto 30% estudaram a análise perceptivo-auditiva (ROY et al., 2013). Outro dado interessante encontrado foi que a maioria dos estudos (78%) investigou a capacidade de um teste em identificar a ausência ou a presença de uma disfonia, enquanto 44% pesquisaram a habilidade

de um teste em determinar a etiologia da disfonia e apenas 18% investigaram a capacidade de um teste para discriminar o grau de alteração de uma disfonia (ROY et al., 2013).

O método objetivo mais difundido na avaliação vocal é a análise acústica linear, cujos principais parâmetros são o *jitter*, o *shimmer*, as medidas de ruído glótico e a frequência fundamental (F_0) (BEHLAU et al., 2001). Outros métodos objetivos estudados são a análise acústica não linear (DAJER et al., 2011; TITZE; BAKEN; HERZEL, 1993) e o uso de sistemas inteligentes, como as redes neurais artificiais (MARTINEZ; RUFINER, 2000; SALHI, 2010).

Quanto ao uso das redes neurais na avaliação vocal observam-se duas linhas de raciocínio, uma que se preocupa em diferenciar as vozes normais de disfônicas, enquanto que outra, além desta diferenciação procura categorizar as vozes disfônicas. Esta última diverge em dois métodos: o primeiro deles busca classificar as vozes disfônicas de acordo com o tipo de patologia causadora da disfonia (CROVATO; SCHUCK, 2007; HARIHARAN; YAACOB; AWANG, 2011; SALHI; MOURAD; CHERIF, 2008; SALHI, 2010; ULOZA et al., 2011) e o outro busca a classificação de acordo com o tipo de desvio vocal (MARTINEZ; RUFINER, 2000), sendo que este último engloba a minoria dos estudos. Além disso, alguns estudos buscaram discriminar o tipo de desvio vocal e quantificar o grau de alteração da disfonia, como ocorre na avaliação perceptivo-auditiva (RITCHINGS; MCGILLION; MOORE, 2002; SÁENZ-LECHÓN et al., 2011; SCHÖNWEILER et al., 2000). Um sistema automático capaz de fazer essa classificação vocal pode minimizar as dificuldades encontradas pela subjetividade da avaliação perceptivo-auditiva.

A aplicabilidade deste sistema pode ser ampla, pois diminui a variabilidade de opiniões quanto à identificação do tipo e grau do desvio vocal, que interferem diretamente na conduta terapêutica. Este sistema ajudaria na avaliação vocal, conduta terapêutica, acompanhamento da evolução do tratamento vocal, confirmação das impressões clínicas, emissão de laudos padronizados, além do treino de alunos na avaliação perceptivo-auditiva e de tornar os resultados de pesquisas mais confiáveis na área de análise vocal. Desta forma, o sistema inteligente agrega informações que subsidiam a avaliação perceptivo-auditiva.

2. OBJETIVOS

Objetivo geral

Propor e validar um sistema automático utilizando redes neurais artificiais para a avaliação de vozes rugosas e soprosas.

Objetivos específicos

- a) Estabelecer a curva de percepção auditiva dos diferentes graus de soprosidade e rugosidade em uma escala de 0 a 100 pontos e a mínima diferença clinicamente significativa em cada ponto desta escala.
- b) Encontrar o subconjunto de parâmetros acústicos mais eficaz na parametrização do sistema automático.
- c) Encontrar a estrutura da Transformada *Wavelet Packet* mais eficaz na parametrização do sistema automático.
- d) Avaliar o desempenho do sistema automático na graduação, em uma escala de 0 a 100 mm, da presença de rugosidade e/ou soprosidade na amostra vocal, quando presentes.

3. CONSIDERAÇÕES TEÓRICAS

3.1 FUNDAMENTOS TEÓRICOS SOBRE A REDE NEURAL ARTIFICIAL

Inspirado no cérebro humano, a rede neural artificial é um sistema inteligente capaz de aprender e reconhecer padrões nos dados apresentados em suas entradas, além de generalizar este conhecimento para a análise de novos dados (HAYKIN, 2001; MACKAY, 2005).

A rede neural artificial (RNA) é um processador que interliga de forma maciça e paralela um conjunto de células computacionais simples que são denominadas “unidades de processamento” ou “neurônios” (HAYKIN, 2001). Na rede neural, cada neurônio possui um número I de entradas, denominadas de x_i , e uma saída denominada y . Cada entrada possui um peso sináptico associado w_i ($i = 1, \dots, I$), podendo haver um parâmetro adicional denominado w_0 , chamado de “*bias*”, podendo ser observado como o peso associado a uma entrada x_0 , que é permanentemente fixada a 1 (MACKAY, 2005) (Figura 1). Destaca-se que os pesos sinápticos associados às entradas x_i são as forças de conexão entre os “neurônios” e são utilizados para armazenar o conhecimento adquirido (HAYKIN, 2001).

Figura 1 - Representação da estrutura de um único neurônio artificial. Fonte: MacKay, 2005.

Uma rede neural pode ser classificada em diferentes tipos, dependendo de sua arquitetura (número de camadas, n° de nós por camada e tipo de conexão que pode ser com ou sem realimentação), topologia (modo como os neurônios estão interconectados – todos interconectados ou não), modelo do neurônio (ex: *perceptron*), e algoritmo de treinamento que muda de acordo com o neurônio utilizado (HAYKIN, 2001).

O modelo mais comum de RNA é a *Multilayer perceptron* (MLP), que possui diversas camadas, dentre elas a camada de entrada, uma ou mais camadas neurais ocultas e a camada neural de saída (Figura 2). A camada neural oculta possui neurônios ocultos que intervêm de maneira útil entre a camada neural de entrada e a de saída, tornando a rede capaz de realizar estatísticas de ordem elevada. O tipo mais comum de treinamento da MLP é realizado pelo do algoritmo *backpropagation* (BP), que se distingue em dois passos, a propagação e a retro propagação (DUDA; HART; STORK, 2001a). Na propagação, calcula-se a saída de cada neurônio, desde a camada de entrada até a camada de saída, sem alterações nos pesos sinápticos. Enquanto que na retro propagação (*backpropagation*), calcula-se um valor de erro por meio da comparação entre a saída da rede neural com a saída desejada, que é retro propagado e permite modificações dos pesos sinápticos com o intuito de minimizar o erro quadrático médio. Este processo representa o treinamento das redes neurais, sendo realizado de forma iterativa até que o erro atinja um limiar pré-definido (DUDA; HART; STORK, 2001a; HAYKIN, 2001).

Figura 2 - Representação da arquitetura de uma rede neural, modelo *Multilayer perceptron*, de duas camadas, onde cada nó representa um neurônio e cada linha um peso. Fonte: MacKay, 2005 (adaptado).

A capacidade de aprender e armazenar o conhecimento são dois aspectos das redes neurais que se assemelham ao cérebro humano. Seus neurônios densamente interconectados funcionam, similarmente ao cérebro, de forma complexa, paralela e não linear. O processamento paralelo permite que as redes neurais sejam tolerantes a falhas, pois nesta organização a falha de alguns neurônios não prejudica de forma significativa o desempenho do restante do sistema. A

característica de não linearidade, por sua vez, permite que o sistema trate sinais de entrada inerentemente não lineares, como a voz (HAYKIN, 2001).

A rede neural possui capacidade de aprender e de se adaptar, assim consegue lidar com dados imprecisos e situações que não estão completamente definidas. Além disso, depois de treinada, possui a capacidade de generalização, ou seja, é capaz de reconhecer semelhanças quando apresentada para um novo padrão de entrada resultando num padrão de saída previsto (HAYKIN, 2001).

3.1.1 Reconhecimento de Padrões

A rede neural artificial aprende por meio de algoritmos de aprendizado, sendo que o aprendizado ocorre quando a RNA atinge uma solução generalizada para uma classe de problemas, ou seja, quando ela reconhece padrões. Reconhecer padrões é uma tarefa realizada com grande rapidez pelo ser humano, com esta habilidade reconhece-se uma face, uma voz, identifica-se pelo olfato se uma fruta está adequada ou não para o consumo, identifica-se objetos pelo tato, entre outros. Apesar de parecer simples essa é uma tarefa altamente complexa e crucial para a nossa sobrevivência (DUDA; HART; STORK, 2001b), desenvolvida ao longo das últimas dezenas de milhões de anos e para a qual evoluíram sistemas neurais e cognitivos altamente sofisticados.

Desta forma, é natural a busca pela construção de máquinas capazes de reconhecer padrões. Do reconhecimento automático da fala, identificação de impressões digitais, à identificação de uma sequência de DNA, fica claro que seria imensamente útil uma máquina capaz de reconhecer padrões de forma precisa e confiável (DUDA; HART; STORK, 2001b).

O principal objetivo do reconhecimento de padrões é identificar a classe de diferentes modelos previamente conhecidos e, para qualquer padrão percebido, escolher o modelo que melhor o representa. De forma resumida, o reconhecimento de padrões possui três fases: pré-processamento, extração de características e classificação. O pré-processamento tem o objetivo de simplificar as operações subsequentes eliminando dados irrelevantes ou distorcidos, mantendo apenas as informações relevantes. A extração de características é a redução dos dados apresentados medindo certas “características” ou “propriedades” do sinal de entrada. A classificação, por sua vez, é realizada por meio da análise dos dados apresentados e declara o sinal de entrada como pertencente a uma categoria (DUDA; HART; STORK, 2001b).

A maior parte do esforço na investigação e desenvolvimento de um sistema classificador, como as redes neurais, está na solução de subproblemas como extrair características que representem da melhor maneira possível os padrões a serem aprendidos. É importante observar se existe qualquer propriedade percebida que não seja inerente do sinal apresentado, como ruídos, e ter cuidado para não “sobre ajustar” a classificação entre as categorias. Esse mecanismo é chamado de *overfitting*, sendo que enquanto ele sobreajusta a classificação das amostras de treinamento, ele provavelmente classificará de forma errônea as novas amostras apresentadas. Assim, é necessário encontrar um ajuste intermediário para o sistema de classificação, ajustando a complexidade do modelo para que ele não seja muito simples, a ponto de não conseguir explicar as diferenças entre as categorias, nem tão complexo a ponto de classificar de forma pobre novas amostras (DUDA; HART; STORK, 2001b).

4. REVISÃO BIBLIOGRÁFICA

4.1 O USO DOS SISTEMAS INTELIGENTES NA AVALIAÇÃO VOCAL

O desenvolvimento de medidas objetivas para a avaliação da função vocal resulta de um desejo antigo de elaborar medidas que proporcionem conhecimentos mais confiáveis, em relação à função vocal, e menos subjetivos do que as avaliações que dependem de julgamentos perceptivos (HILLMAN; MONTGOMERY; ZEITELS, 1997). Neste contexto são desenvolvidos novos métodos de avaliação acústica da voz, como o uso de sistemas inteligentes.

Entre os estudos que buscaram diferenciar vozes normais de patológicas está o de Godino-Llorente e Gómez-Vilda (2004) que compararam duas abordagens de sistemas inteligentes na discriminação das vozes, sendo as redes neurais do tipo *multilayer perceptron* e as redes neurais do tipo aprendizagem por quantização vetorial (LVQ) usando vetores calculados a curto-prazo baseados no coeficiente de frequência *Mel-cepstral*. Os autores utilizaram para o treino das redes neurais a vogal sustentada /a/ de 135 indivíduos, 53 com vozes normais e 82 com disfonia, encontrando que ambas as abordagens identificaram de forma confiável as vozes disfônicas. No entanto, a metodologia de aprendizagem por quantização vetorial demonstrou ser mais confiável do que a *multilayer perceptron*, identificando corretamente 96% das vozes.

Outro estudo nesta linha foi o de Hariharan e equipe (HARIHARAN; YAACOB; AWANG, 2011) que realizaram um estudo para discriminar vozes do choro de bebês em normais ou patológicas, utilizando a rede neural associada à extração de caracteres por meio da Transformada *Wavelet Packet*. Os autores dividiram o sinal de voz em cinco níveis, por meio da Transformada *Wavelet Packet*, extraindo as medidas de energia e entropia de cada nível decomposto. Assim, quantificaram o sinal do choro dos bebês. Utilizaram estas medidas para treinar as redes neurais na discriminação entre as vozes de choros normais e patológicos, sendo que as treinaram com diferentes fatores de propagação ou parâmetro de suavização para obterem a melhor precisão de classificação. Os autores consideraram como choro patológico o choro de bebês com sinal de asfixia, a qual foi definida como a incapacidade de respirar bem dentro de 1 min após o parto do bebê. Eles analisaram 680 vozes, sendo 340 de choros normais e 340 de choros com sinais de asfixia. Estas vozes foram selecionadas de um banco de dados do Instituto Nacional de Astrofísica Ótica e Eletrônica (INAOE) – CONACYT, México. Os resultados da pesquisa foram promissores,

apresentando 99% de precisão na discriminação dos choros, mostrando a eficiência das redes neurais e da *Wavelet Packet* na discriminação dos choros dos bebês.

De Bruijn et al. (2011) buscaram classificar a ressonância vocal entre equilibrada e hipernasal. Gravaram vozes de 51 sujeitos, 18 falantes controle e 51 com hipernasalidade. Para a análise, dois fonoaudiólogos avaliaram as vozes em um trecho de fala, que posteriormente foi editada por meio do *Praat*, selecionando-se as vogais /a/, /i/ e /u/ para o treino da rede neural artificial (RNA). A RNA foi treinada utilizando-se sete coeficientes *mel-ceptrais* em combinação com os valores 0, ausência, e 1, presença de hipernasalidade. Os autores obtiveram o melhor resultado de discriminação com as vogais /i/ e /u/ com 24,9% de acerto e concluíram que este estudo é viável mas que os resultados foram pobres, assim, no momento, a RNA não é capaz de substituir um avaliador treinado.

Arjmandi e Pooyan (2012) utilizaram a Transformada *Wavelet Packet* na tentativa de diferenciar vozes patológicas de normais. Realizaram o treino/teste com um banco de dados contendo 53 indivíduos com vozes normais e 67 com vozes patológicas, sendo que para encontrar a melhor árvore realizaram uma análise qualitativa do sinal da Transformada *Wavelet* e da transformada de *Fourier* a curto prazo. Posteriormente, aplicaram a análise de componentes principais (PCA) e a análise de discriminante linear (LDA) para redução do vetor de característica e otimização do algoritmo. Os resultados apontaram a melhor estrutura da árvore de decomposição *Wavelet Packet* em seu sexto nível e alcançaram 100% de acerto na discriminação entre vozes normais e patológicas.

Cordeiro et al. (2013) pesquisaram a eficiência do envelope espectral do sinal de voz na diferenciação de vozes normais de patológicas e, posteriormente, entre vozes com desordens causadas por nódulo vocal das com edema de *Reinke*. Para a análise, utilizaram a vogal /a/ de 16 indivíduos com vozes saudáveis, 16 com edema de Reinke e 15 com nódulo vocal, todos com indivíduos de ambos os sexos. Os autores encontraram um recurso importante, na detecção das vozes patológicas, na análise da frequência e da largura da banda do primeiro pico do coeficiente de predição linear (LPC) de ordem 30. A partir desta análise, obtiveram uma discriminação de 100% entre os indivíduos saudáveis e não saudáveis e uma discriminação de 87% entre nódulos e edema de *Reinke*.

Atualmente, a maioria dos estudos de avaliação vocal realizados por meio de sistemas inteligentes, busca categorizar as vozes disfônicas de acordo com o tipo da patologia causadora da

disfonia ou um de grupo de patologias semelhantes. Neste contexto, Salhi et al. (2010) buscaram discriminar vozes normais de patológicas, subclassificando as vozes patológicas em disfonias causadas por alterações neurológicas e disfonias causadas por alterações laríngeas. Os autores avaliaram indivíduos com Parkinson, Alzheimer, dislexia e alterações laríngeas, sendo que não descrevem o banco de dados de forma clara, nem explicam qual tipo de disfonia presente nos casos apresentados. Na classificação das vozes utilizaram as redes neurais, do tipo *multilayer perceptron* com aprendizagem por *backpropagation*, associadas à Transformada Wavelet discreta e contínua geradas de uma amostra de fala. Os resultados foram de 100% de acerto na discriminação entre vozes normais e patológicas treinando as redes neurais por meio da energia contida em 7 sub-bandas de uma Wavelet até o 3º nível de decomposição. Sendo que os autores não apresentaram os resultados quanto à subclassificação das vozes patológicas em disfônicas causadas por alterações neurológicas e por alterações laríngeas.

Crovato e Schuck (2007) classificaram vozes disfônicas usando a Transformada *Wavelet Packet* e o algoritmo *Best Basis* (BBA) como redutor de dimensionalidade e seis Redes Neurais Artificiais atuando como sistemas “especializados”. Cada rede neural foi construída com três camadas *multilayer perceptron*, com 64 nós de entrada, um nó de saída e na camada intermediária um número de neurônios dependente do treino de cada grupo de patologia vocal. As vozes foram subdivididas em 5 grupos de disfônicos e um grupo de vozes saudáveis, sendo o grupo 1 composto por disfonias causadas por laringites crônicas, o grupo 2 disfonias por doenças degenerativas (câncer, granuloma e leucoplasia), grupo 3 disfonias por mobilidade incorreta das pregas vocais (paralisia unilateral e disfonias hipercinéticas), grupo 4 por alterações orgânicas nas pregas vocais (edema de *Reinke*, edema generalizado), grupo 5 por crescimentos orgânicos (nódulos, pólipos e cistos) e grupo 6 com vozes saudáveis. Cada rede neural foi treinada e associada a um dos 6 grupos, sendo alimentada pelo valor de entropia dos nós da *Best Basis Tree*. Por meio do método de validação cruzada múltipla e da técnica de variação *leave-one-out* as taxas de sucesso obtidas foram 87,5%, 95,31%, 87,5%, 100%, 96,87% e 89,06% para os grupos de 01 a 06, respectivamente.

Carvalho, Cavalcante e Cortez (2011) utilizaram as redes neurais treinadas com características extraídas da voz pela Transformada Discreta *Wavelet* (DWT) para diferenciar vozes subdivididas em quatro grupos: vozes normais, disfônicas causadas por nódulos vocais, disfônicas por edema de *Reinke* e disfonias neurológicas (Acidente Vascular Cerebral, Parkinson, Esclerose Lateral Amiotrófica, Distonia cervical, entre outros). Usando para o treino das redes neurais uma

DWT com 7 níveis de decomposição, os autores encontraram classificação correta de 93,33% para vozes normais, 86,56% para disfonias por nódulos vocais, 87,61% para disfonias por edema de *Reinke* e 89,40% para as vozes neurológicas.

Uloza et al. (2011) utilizaram outro tipo de sistema inteligente na discriminação de vozes, as máquinas de vetores de suporte ou *support vector machines*. Os autores construíram um classificador para discriminar 444 vozes em vozes saudáveis, vozes patológicas causadas por lesões nodulares nas pregas vocais, englobando nódulos, pólipos e cistos, e vozes patológicas causadas por lesões difusas, como papiloma, laringite crônica hiperplásica com queratose e carcinoma. Além disso, os autores compararam a eficiência da discriminação das vozes pela máquina de vetores de suporte com a discriminação auditiva de quatro avaliadores experientes na área de voz. Os autores obtiveram uma taxa de classificação correta de 90% entre vozes saudáveis, patológicas por lesões nodulares e patológicas por lesões difusas. Enquanto que na discriminação auditiva dos avaliadores a taxa de classificação correta foi de 60%. Desta forma, os autores concluíram que o sistema de classificação automático foi mais eficiente que a discriminação auditiva dos avaliadores.

Alguns pesquisadores buscam a diferenciação das vozes disfônicas por meio da classificação do tipo de desvio vocal, como Martínez e Rufiner (2000) que por meio das redes neurais buscaram a diferenciação de vozes normais, bitonais e roucas. Os autores utilizaram as redes neurais as quais foram treinadas por meio de parâmetros extraídos da Transformada Rápida de *Fourier* (FFT), coeficiente *Cepstrais*, escala *Mel-Cepstral*, *Delta-Cepstral* e *Delta Mel-cepstral*. Os resultados foram promissores, com porcentagens elevadas de classificação correta das vozes. A análise cepstral foi a técnica de processamento que obteve melhor desempenho, com 91% de acerto, e a Transformada rápida de *Fourier* a com pior desempenho com 63,5% de acertos.

Linder et al. (2008) treinaram redes neurais, com arquitetura de rede multineural e propagação adaptativa, melhorado com um algoritmo de *backpropagation*, com o intuito de diferenciar vozes saudáveis de rugosas e soprosas. Além disso, diferenciar as vozes normais de disfônicas por um parâmetro que chamaram de *H* (*hoarseness*), onde $H=0$ era considerado as vozes rugosas e soprosas avaliadas pela GBRAS em 0 (voz sem desvio) ou 1 (desvio vocal leve) e $H=1$ englobava as vozes com grau 2 (desvio vocal moderado) e grau 3 (desvio vocal intenso). O intuito da criação desta análise foi instruir profissionais da saúde que não são especialistas na área da voz a encaminharem quando necessário o paciente para tratamento específico, os autores indicam que vozes consideradas com $H=1$ devem ser encaminhadas para tratamento. 120 amostras vocais

foram avaliadas perceptivo-auditivamente por 31 clínicos da área da voz, com confiabilidade inter-avaliadores moderada. Além disso, foi realizada análise acústica do *jitter*, *shimmer*, desvio padrão, frequência fundamental (F_0) e *glottal-to-noise excitation* (GNE), os quais foram correlacionados com os dados perceptivos-auditivos. Ambos, análise acústica e perceptivo-auditiva foram utilizados para treino das redes neurais. Os resultados apontaram 39,2% de classificação correta entre voz rouca e soprosa, 6,25% de acerto maior que o resultado ao acaso. Na discriminação entre voz normal e disfônica pelo parâmetro H a porcentagem de classificação correta foi de 80%. Os autores concluíram que os resultados da distinção entre vozes normais e disfônicas foi bom e pretendem criar um programa comercializado para os não-especialistas em voz saberem quando realizar encaminhamentos e para triagens vocais.

Alguns trabalhos tentaram discriminar o grau de desvio vocal, como Schönweiler et al. (2000) que por meio do treino das redes neurais, estatística multivariada e mapas combinados buscaram discriminar vozes soprosas e rugosas e o grau de desvio vocal destas. Para a parametrização dos sistemas utilizaram parâmetros vocais acústicos correspondentes à uma escala perceptivo-auditiva de quatro pontos, sendo “0” sem desvio vocal, “1” desvio leve, “2” moderado e “3” intenso. Para o estudo selecionou-se 117 vozes com e sem desvio vocal, as quais foram analisadas por 31 avaliadores por meio da avaliação perceptivo-auditiva. As redes neurais foram treinadas com parâmetros acústicos avaliados com “importância relativa” alta nas árvores de regressão, sendo o log das medidas: desvio padrão da frequência fundamental, *jitter*, *shimmer*, proporção harmônico-ruído e grau de quebras de sonoridade, todos calculados pelo MDVP (*Multi-Dimensional Voice Program*). Os autores encontraram os melhores resultados com as redes neurais *feedforward* de três camadas, constituídas por 50 redes de trabalho simultâneo. A taxa de acerto foi de 40% na classificação combinada da soproidade e rugosidade, subindo para 65% quando se analisaram elas separadamente. Observa-se que a melhor classificação foi para a soproidade sem desvio vocal a desvio vocal moderado e para a rugosidade na classificação do desvio vocal de grau leve. Segundo os autores estes resultados estão dentro da faixa relatada por outros estudos utilizando redes neurais artificiais nas áreas de biologia e medicina clínica. Assim, conclui-se que os resultados incentivam ainda mais a pesquisa de redes neurais *feedforward* para análise acústica vocal.

Ritchings et al. (2002) propõe um sistema baseado nas redes neurais para a avaliação objetiva da qualidade vocal em pacientes em recuperação de vários estágios de câncer de laringe.

Para isso os autores selecionaram 77 vozes masculinas, cuja amostra vocal foi a vogal sustentada /i/, as quais foram avaliadas auditivamente por um fonoaudiólogo por meio de uma escala numérica chamada SALT, que possui 7 pontos, sendo “0” as vozes com menor grau de desvio e “7” com maior grau de desvio vocal. O sistema proposto extraiu parâmetros de curto e longo prazo no domínio do tempo e no domínio da frequência que foram mais relevantes e os usou para treinar e testar a rede neural artificial, do tipo *multilayer perceptron*. Os autores realizaram diversas combinações destes parâmetros e obtiveram os melhores resultados com uma combinação de parâmetros de curto e longo prazo, sendo capaz de distinguir os 7 graus de desvio vocal com uma precisão de 92% de acerto. Assim, os autores concluem que este sistema poderia ser utilizado como uma ferramenta de avaliação, auxiliando na avaliação perceptivo-auditiva por meio da escala SALT durante a avaliação clínica da qualidade da voz.

Sáenz-Lechón et al. (2011) realizaram um estudo com o objetivo de classificar automaticamente vozes rugosas (R) de soprosas (B) de acordo com a escala GRBAS de avaliação perceptivo-auditiva. Para isso, um sistema foi treinado para extrair o primeiro coeficiente de frequência *Mel-cepstral* de uma vogal sustentada. Em seguida, um classificador foi treinado para estimar os correspondentes graus rugosidade e soprosidade. O estudo foi desenvolvido com 226 amostras vocais, as quais foram avaliadas por um juiz experiente na área, por meio da escala GRBAS e subdividas em normais e disfônicas, ambos contendo soprosidade e rugosidade. Os autores obtiveram 62% de classificação correta para vozes rugosas e soprosas de grau “0” (sem desvio vocal), 30% e 42% para vozes R e B, respectivamente para o grau 1 (desvio vocal leve), 35% (R) e 33% (B) para o grau 2 (desvio moderado) e, finalmente, 48% (R) e 41% (B) para o grau 3 (desvio intenso). Os autores consideraram que os resultados obtidos revelaram uma correlação significativa entre a avaliação subjetiva e automática, indicando a viabilidade da avaliação objetiva da qualidade de voz por meio de medidas perceptivamente significativas.

Stráník, Cmejla e Vokrál (2014) avaliaram a soprosidade em um sistema automático (implementado pelo *Toolbox* de processamento de fala do programa *Matlab*) baseado na escala GRBAS, na qual 0 corresponde a ausência de desvio vocal, 1 desvio vocal leve, 2 moderado e 3 intenso. Para o estudo selecionaram 593 trechos de fala, contendo 34 palavras, os quais foram avaliados por 5 especialistas da área. Na parametrização os autores testaram 92 parâmetros, dos quais quatro foram indicados na seleção de atributos por meio do programa *Weka*, sendo: *Glottal-to-Noise Excitation ratio*, *Cepstral Peak Prominence Pearson* da autocorrelação do pico, índice de

soprosidade, e a razão entre as energias das frequências altas, médias e baixas. Os resultados apontaram acerto de 77% na avaliação das vozes soprosas, com concordância de 0,59 pelo teste estatístico *Kappa*, e os autores concluíram que é promissora a avaliação da soprosidade por meio de um sistema automático baseado em medidas acústicas.

Moro-Velázquez et al. (2015) pesquisaram um sistema automático, baseado no Modelo de Misturas Gaussianas, para avaliar o grau geral e a rugosidade de amostras vocais de acordo com a escala perceptiva de 4 pontos GRBAS. Os autores utilizaram duas parametrizações, sendo uma baseada nos coeficientes cepstrais *mel-frequency* e outra por um conjunto de parâmetros derivados dos espectros da modulação. O estudo foi desenvolvido com um banco de dados, com amostra da vogal sustentada /a/ e uma leitura do texto “*Rainbow passage*”, com 53 vozes normais e 657 vozes patológicas. O melhor resultado foi obtido usando para a parametrização os parâmetros do espectro da modulação, sendo encontrado 81,6% de eficiência no grau geral e 84,7% na rugosidade. Além disso, a concordância foi qualificada como boa pelo teste *Kappa*, no qual encontrou-se valor de 0,73 para o grau geral e 0,76 para a rugosidade.

4.2 CORRESPONDÊNCIAS ENTRE A ANÁLISE ACÚSTICA E A AVALIAÇÃO PERCEPTIVO-AUDITIVA

A avaliação perceptivo-auditiva da voz e a análise acústica são metodologias complementares de avaliação (LOPES et al., 2012). A avaliação perceptivo-auditiva é considerada o padrão-ouro da avaliação vocal realizada pelo fonoaudiólogo (MADAZIO; LEÃO; BEHLAU, 2011; MARYN; ROY, 2012; OATES, 2009), no entanto depende de julgamentos perceptivos. Em contrapartida a análise acústica é um método objetivo, com reprodutibilidade de resultados (FROHLICH; MICHAELIS; WERNER STRUBE, 1998), mas que precisa ser aprimorado.

Para que a análise acústica seja confiável é necessário compreender suas limitações, neste intuito Titze (TITZE, 1995) propôs uma classificação dos tipos de onda do sinal de voz em três tipos, sinal do tipo 1, 2 e 3. No sinal do tipo 1 encontram-se as vozes cujo sinal é quase periódico e não há mudanças qualitativas na análise do segmento, e se há sub-harmônicos as suas energias possuem uma magnitude inferior à energia da frequência fundamental. No tipo 2 estão os sinais com alterações qualitativas (bifurcações) na análise do segmento ou com sub-harmônicos, cujas energias se aproximam da energia da frequência fundamental (F_0). E no tipo 3 estão os sinais sem uma estrutura periódica aparente. Sendo que apenas sinais do tipo 1 permitem uma análise acústica confiável das medidas de perturbação, como *jitter*, *shimmer*, frequência fundamental, proporção harmônico-ruído.

Em consonância com esta proposição Rabinov et al. (1995), em um estudo onde comparou a acurácia da avaliação perceptivo-auditiva com a do *jitter* na avaliação de vozes ásperas, concluiu que a avaliação perceptivo-auditiva foi mais confiável que a análise acústica do *jitter* conforme houve aumento do desvio vocal.

O *jitter* e o *shimmer* são medidas de perturbação do sinal sonoro a curto-prazo, ou seja, medem perturbações que ocorrem ciclo-a-ciclo na fonação. O *jitter* mensura as perturbações referentes ao período do ciclo e altera-se quando há falta de controle na vibração das pregas vocais, enquanto que o *shimmer* mensura as perturbações referentes à amplitude da onda do sinal de voz e altera-se quando há diminuição da resistência glótica, por exemplo: paralisia de pregas vocais e fendas (BEHLAU et al., 2001). A correlação entre estes parâmetros e os achados perceptivo-auditivos é controversa. Alguns estudos descrevem o *jitter* como correlacionado com a aspereza (BEHLAU et al., 2001), enquanto outros o correlacionam com a rouquidão (WOLFE; MARTIN,

1997), ou com a soprosidade (DEJONCKERE, 1995), ou ainda não o correlaciona nem com a rugosidade nem com a soprosidade (BHUTA; PATRICK; GARNETT, 2004; MADAZIO; LEÃO; BEHLAU, 2011). Com o *shimmer* não é diferente, é correlacionado com a rouquidão e soprosidade (BEHLAU et al., 2001; LOPES et al., 2012) e não é correlacionado com nenhum destes dois parâmetros (BHUTA; PATRICK; GARNETT, 2004; MADAZIO; LEÃO; BEHLAU, 2011). Com tantas discrepâncias a confiabilidade destes parâmetros é insatisfatória, com medidas de baixa sensibilidade e especificidade, especialmente em relação aos sinais com aperiodicidade intensa (BROCKMANN et al., 2011). Apesar disso, o uso destas medidas é promissor quando associado com sistemas automáticos de reconhecimento de padrões (WANG; JO, 2007).

Apesar da análise acústica ser um método objetivo de avaliação ela envolve uma série de fatores na gravação vocal que interferem na qualidade da captação do sinal, da análise e consequentemente na correlação com parâmetros perceptivo-auditivos. Fatores como o microfone utilizado e o ambiente de gravação são bastante conhecidos, no entanto Brockmrna et al. (2011) alertam, que também influenciam os valores do *jitter* e do *shimmer*, a vogal utilizada, o gênero e a intensidade vocal, que foi o fator de maior influência na captação destes parâmetros. Baseado nestes achados os autores sugerem um protocolo padronizado de gravação vocal com a vogal /a/ e intensidade vocal mínima de 80 dB.

Bhuta et al. (2011) correlacionaram dados da avaliação perceptivo-auditiva por meio da escala GRBAS, com a análise acústica realizada por meio do programa *Multi-Dimensional Voice Program* (MDVP). O estudo foi realizado com vozes de 37 pacientes (12 homens/ 25 mulheres). A escala GRBAS analisa o grau geral de desvio vocal (G), a rugosidade (R), a soprosidade (S), a astenia (A) e a tensão da voz (S), por outro lado o MDVP analisa 19 parâmetros acústicos da voz. Os autores encontraram correlação entre 3 parâmetros acústicos do MDVP com os parâmetros perceptivos da GRBAS. O grau geral de desvio vocal apresentou correlação com o índice de turbulência da voz, com a proporção harmônico-ruído e com o índice de fonação suave. A rugosidade apresentou correlação com a proporção harmônico-ruído (NHR) e a soprosidade com o índice de fonação suave. A partir destes resultados os autores sugerem que futuros estudos desenvolvam uma medida para quantificar a voz a partir da GRBAS e dos 3 parâmetros clinicamente relevantes do MDVP.

Lopes et al. (2012) em um trabalho avaliando 71 crianças de 3 a 9 anos verificaram se havia correlação entre a avaliação perceptivo-auditiva (realizada por uma escala visual analógica) com a

análise acústica da F_0 , *jitter*, *shimmer* e *glottal-to-noise excitation* – GNE). Encontraram por meio do teste de *Pearson* correlação (de fraca a moderada) do *shimmer* e do GNE com o grau geral de desvio vocal, rugosidade, sopro e parâmetros de instabilidade vocal. Sendo que o GNE foi o único parâmetro capaz de distinguir a intensidade do desvio vocal.

Pereira et al. (2006) realizaram uma pesquisa com 21 indivíduos com *Parkinson* e 15 saudáveis, esses últimos sem histórico de doenças neurológicas ou otorrinolaringológica, não-fumantes e sem relato de uso abusivo de álcool. O objetivo do estudo foi avaliar o parâmetro acústico Amplitude do pitch (PA), também denominado como autocorrelação do sinal residual, nos indivíduos com doença de *Parkinson*. O autor encontrou diminuição significativa dos valores do PA nos indivíduos com *Parkinson* em comparação com os saudáveis, concluindo que este parâmetro possui boa especificidade para diferenciar estes grupos.

Freitas et al. (2015) correlacionam a avaliação perceptivo-auditiva, por meio da escala GRBRAS, com a análise acústica de quatro diferentes programas de análise acústica. Os resultados apontam diferenças na correlação dos parâmetros perceptivos. Por meio do *Dr. Speech* (Tiger Electronics, Seattle, WA) os autores encontram o *shimmer* correlacionado com a rugosidade e o o parâmetro proporção-harmônico-ruído (HNR) com a sopro; através do *Multidimensional Voice Program* (Kay Elemetrics, NJ, USA) encontraram o *jitter* correlacionado a sopro e o *shimmer* com a rugosidade; no PRAAT (University of Amsterdam, The Netherlands) a rugosidade não aparece correlacionado com nenhum parâmetro acústico e a sopro correlacionada com o HNR; e no *Voice Studio* (Seegnal, Oporto, Portugal) encontraram o *shimmer* correlacionado tanto com a rugosidade como com a sopro. Estes resultados mostram a labilidade destes parâmetros de acordo com o programa acústico utilizado e ressalta o cuidado ao utilizá-los.

Como pode ser observado, há uma controvérsia em relação à utilidade destas medidas acústicas de forma isolada (MADAZIO; LEÃO; BEHLAU, 2011), assim observa-se que atualmente a tendência é procurar um conjunto de parâmetros acústicos que correspondam de forma mais confiável ao tipo e grau de disfonia (AWAN; ROY, 2005; HAKKESTEEGT et al., 2008; MA; YIU, 2006; MADAZIO; LEÃO; BEHLAU, 2011; WUYTS et al., 2000; YU et al., 2001).

Ma e Yiu (2006) avaliaram a precisão de um conjunto de parâmetros na discriminação da intensidade do desvio vocal comparativamente ao grau geral de desvio vocal graduado por meio da avaliação perceptivo-auditiva (G da escala GRBAS). Os parâmetros selecionados incluíram

medidas aerodinâmicas (tempo máximo de fonação, pressão intraoral, fluxo de ar e pressão na fonação) perfil de extensão vocal (fonetograma) e análise acústica de medidas de perturbação (F_0 , *jitter*, *shimmer* e proporção harmônico-ruído). Participaram do estudo 112 indivíduos com disfonia e 41 com vozes normais. Os resultados apontaram que na discriminação do grau da disfonia os melhores parâmetros foram o tempo máximo de fonação na vogal /a/, o pico de pressão intraoral na sílaba /pi/, o perfil de extensão vocal e o dos parâmetros acústicos o *jitter*. Concluíram que o uso destes quatro parâmetros combinados discriminaram corretamente 67,3% dos níveis de desvio vocal.

Hakkesteege et al. (2008) tiveram por objetivo investigar a eficiência do Índice de Severidade da disfonia (DSI), o qual analisa a voz por meio de um conjunto de parâmetros acústicos, e compará-lo com a escala perceptivo-auditiva GRBAS (G- grau geral de desvio vocal, R- rugosidade, B-soprosidade, A- astenia e S-tensão). Para o cálculo do DSI utiliza-se a frequência fundamental mais alta (F_0 -high em Hz), a menor intensidade (I-low em dBSPL), o tempo máximo de fonação (TMF em segundos) e o *jitter* (%). Para sua construção realiza-se o cálculo a seguir: $DSI = 0,13 \times TMF + 0,053 \times F_0 \text{ Higt} - 0,26 \times I \text{ low} - 1,18 \times Jitter (\%) + 12,4$. Os resultados variam de +5 à -5, sendo que +5 corresponde a vozes normais e -5 vozes com disfonias intensas. Participaram do trabalho 294 sujeitos com disfonias e 118 com vozes normais, os quais foram avaliados pela GRBAS e pelo DSI. Os resultados apontaram que os pacientes disfônicos apresentaram um escore de DSI menor que o grupo controle e DSI significativamente menor quando a grau perceptivo-auditivo foi maior. Na discriminação entre indivíduos disfônicos e não-disfônicos encontrou-se uma sensibilidade de 0,72 e especificidade de 0,75, considerando-se o valor de corte do DSI de 3,0. Os autores concluíram que o DSI é um instrumento útil para mensurar objetivamente a gravidade da disfonia.

Nemr et al. (2015) estudaram a correlação entre o DSI, a avaliação perceptivo-auditiva (CAPE-V) e o gênero, de indivíduos com e sem disfonia. Participaram do estudo 66 indivíduos, sendo 24 disfônicos e 42 com vozes sem alterações, os quais foram avaliados por um fonoaudiólogo especialista na área. O DSI foi capaz de diferenciar os indivíduos disfônicos dos normais, sendo que no grupo com disfonia encontrou-se correlação moderada do DSI com o grau geral da disfonia (CAPE-V) e com a soprosidade, além de uma correlação fraca com a rugosidade. No grupo sem alterações vocais encontraram tempo máximo de fonação maior nos homens, sendo que em ambos os grupos foi encontrada frequência fundamental mais aguda nas mulheres. Os autores concluíram

que o DSI é capaz de distinguir indivíduos com e sem disfonia, que há correlação com a escala CAPE-V e não há correlação com o gênero.

Madazio et al. (2011) identificaram, por meio do Diagrama do desvio fonatório, quais as características discriminavam vozes em predominantemente rugosas, soprosas e tensas. O Diagrama do desvio fonatório é representado em um plano bidimensional, sendo no eixo horizontal medidas de perturbação (*jitter*, *shimmer* e correlação) e no eixo vertical medidas de ruído (*glottal-to-noise ratio* – GNE). Um total de 196 amostras vocais normais e disfônicas foram submetidas à avaliação perceptivo-auditiva do tipo e grau de desvio vocal, e à análise acústica do *jitter*, *shimmer*, correlação, GNE e interpretação do diagrama do desvio fonatório de acordo com a localização nos quadrantes, forma e densidade. Os resultados apontaram, entre os parâmetros acústicos, o GNE como o único capaz de diferenciar as vozes rugosas, soprosas e tensas. Além disso, os autores encontraram diferenças significativas, quanto à localização no diagrama, entre vozes normais e disfônicas, entre soprosas e rugosas e entre soprosas e tensas.

Maryn e colaboradores desenvolveram uma série de estudos correlacionando o grau geral da disfonia, avaliado pela perceptiva-auditiva, com conjuntos de parâmetros acústicos, utilizando amostras de vogal sustentada e de fala combinadas. Em uma pesquisa com 251 sujeitos, publicada em 2010, testaram a correlação do grau geral da disfonia com 13 parâmetros acústicos, os quais foram extraídos pelo programa *Praat* e eram baseados em medidas da perturbação da frequência, da amplitude, análises do espectro e do cepstro. Por meio de uma análise de regressão múltipla os autores encontraram um modelo multiparamétrico com seis parâmetros acústicos na correlação com o grau geral da voz, são eles: *Smoothed cepstral peak prominence* (CPPS), *slope of the long-term average spectrum* (Slope), *tilt of the trend line through the long-term average spectrum* (Tilt), *shimmer local* (*shimmer* %), *shimmer local dB* (*shimmer* em dB), e proporção harmônico-ruído (HNR). Encontrando o CPPS, medida ceptral, o principal contribuinte para a predição da qualidade global voz. Os resultados apontaram coeficiente de correlação de 0,78 e área da curva ROC = 0,89, reforçando a hipótese de que o uso combinado de amostras de vogal sustentada e de fala são eficientes no diagnóstico do grau geral da voz. Os autores chamaram o modelo encontrado de Índice acústico de qualidade vocal (AVQI) e sugeriram o aprimoramento deste algoritmo acústico encontrado (MARYN et al., 2010).

Em 2015, Maryn e Weenink estudaram o parâmetro acústico “*Smoothed cepstral peak prominence*” (CPPS), principal componente do Índice acústico de qualidade vocal (AVQI), visto

que houve uma implementação deste parâmetro no programa *Praat*. Além disso, também investigaram a relação do AVQI com o nova e a antiga versão do CPPS. Para esta pesquisa os autores selecionaram 289 vozes, com amostras de vogal sustentada e fala. Encontraram alta correlação entre as versões antiga e nova do AVQI, com coeficientes de determinação e correlação próximos a 100%, concluindo que ambas as versões são comparáveis e podem medir a severidade da disfonia facilmente.

Outros estudos seguiram uma linha diferente de análise acústica, fundamentando-se não apenas nas medidas acústicas baseadas no tempo (como frequência fundamental, *jitter* e *shimmer*), mas também em medidas baseadas na análise do espectro/cepstro do sinal sonoro. Awan and Roy (2005) procuraram um conjunto de medidas acústicas (baseadas tanto no tempo quanto no espectro/cepstro) para classificação do tipo de voz em normal, soprosa, rouca e áspera. Participaram do estudo 134 mulheres, sem desvio vocal e com disfonia funcional, diversificadas quanto ao tipo e grau do desvio vocal. Os autores produziram um modelo baseado em 5 medidas acústicas: LOGSHIM (logaritmo do *shimmer*), INVSQRTF₀ (raiz quadrada inversa da F₀), INVSQRTSIG (raiz quadrada inversa do *pitch* sigma), DFTR (razão da Transformada rápida de Fourier) e CPP/EXP (razão entre a amplitude do pico cepstral e a amplitude esperada do cepstro como determinado através da regressão linear). Por meio deste conjunto de medidas os autores classificaram corretamente os tipos de vozes com 79,9% de precisão. Concluíram que o uso de medidas baseadas no tempo associadas a medidas baseadas no espectro/ cepstro é uma abordagem promissora. As medidas baseadas no espectro são calculadas independentemente da identificação de cada ciclo, assim, são capazes de quantificar disfonias intensas com sinais vocais aperiódicos de forma mais confiável, reduzindo possíveis imprecisões no cálculo destes sinais. Por outro lado, as medidas tradicionais baseadas no tempo são mais compreendidas, possuindo um apoio da literatura para sua utilização tanto quanto sua relação com as impressões perceptivas da voz.

5. MATERIAL E MÉTODOS

Este estudo foi desenvolvido com gravações vocais contidas no banco de dados do Laboratório de Voz da Clínica de Fonoaudiologia da Faculdade de Odontologia de Bauru - Universidade de São Paulo (FOB/USP), as quais fazem parte dos procedimentos realizados rotineiramente com pacientes atendidos nesta clínica. Foi aprovado pelo Comitê de Ética em Pesquisa da Faculdade de Odontologia de Bauru - Universidade de São Paulo (FOB-USP) sob parecer nº 161.349.

As gravações contidas no banco de dados da Clínica de Fonoaudiologia são realizadas em ambiente acusticamente tratado e armazenadas digitalmente. Durante o registro vocal os pacientes permanecem sentados com um microfone de cabeça da marca AKG, modelo C444PP, o qual é posicionado lateralmente a 60° e a 5 cm da comissura labial, conforme padronizado por Titze (TITZE, 1995). O sinal sonoro é registrado por meio do programa *Sound Forge 10.0*, com uma taxa de amostragem de 44.100 Hz, canal Mono em 16 bit, em um computador *Intel*®, *Pentium*® 4, CPU 2.040 GHz e 256 MB de RAM, monitor *LG Flatron E7015 17"*, placa de som modelo Audigy II, marca *Creative*.

5.1 Casuística

Foram selecionadas 150 gravações vocais da vogal sustentada /a/, das quais 123 permaneceram na pesquisa. As amostras englobaram indivíduos de ambos os sexos, com idade superior a 18 anos. As amostras vocais englobaram desde vozes neutras até vozes com presença em grau intenso de rugosidade e/ou soprosidade, sendo consideradas vozes rugosas aquelas que apresentam impressão de irregularidade na fonte sonora, podendo ter elementos de rouquidão, aspereza e crepitação; e vozes soprosas aquelas que apresentam escape de ar audível na voz (BEHLAU, 2004).

Após a seleção, as vozes foram mantidas ou excluídas desta pesquisa de acordo com a avaliação de quatro juízes e com os critérios de inclusão: qualidade vocal caracterizada como predominantemente rugosa ou soprosa, com concordância entre pelo menos 3 juízes; e diferença máxima de 10 milímetros (mm) na avaliação perceptivo-auditiva (EADIE et al., 2010) de pelo menos três dos quatro juízes.

5.2 Edição das gravações vocais

A amostra vocal utilizada foi a vogal sustentada /a/, a qual foi editada, por meio do programa *Sound Forge* 10.0, para padronização da análise. A edição consistiu na eliminação do primeiro segundo da emissão e seleção dos 3 segundos subsequentes, sendo que se fosse observada alguma irregularidade abrupta buscava-se um outro trecho mais estável da emissão. Desta forma, eliminaram-se os principais trechos de instabilidade vocal e obteve-se uma amostra vocal mais homogênea para o treino do sistema automático e uma avaliação mais confiável.

5.3 Avaliação perceptivo-auditiva

A avaliação perceptivo-auditiva foi realizada por quatro fonoaudiólogos, especialistas em voz, previamente treinados. O treinamento consistiu em avaliações individuais de amostras vocais da vogal sustentada /a/, das quais os fonoaudiólogos foram orientados a graduar a soproidade (escape de ar audível na voz) e a rugosidade (irregularidade na vibração das pregas vocais), sendo que na rugosidade os fonoaudiólogos deveriam indicar qual era o componente predominante: crepitação, aspereza ou rouquidão.

A avaliação foi realizada em duas etapas distintas, na primeira as vozes foram avaliadas por meio de uma escala visual analógica (ANEXO A) que varia entre 0 e 100 mm, onde “0 mm” indica ausência de desvio vocal e “100 mm” o máximo desvio vocal possível; e na segunda as vozes foram avaliadas por uma escala numérica de quatro pontos (ANEXO B), na qual “0” representa ausência do parâmetro avaliado, “1” presença em grau leve, “2” moderado e “3” intenso.

Tanto na fase de treinamento como na de avaliação as análises vocais foram realizadas em ambiente acusticamente tratado, por meio de fones de ouvido da marca *Behringer*, modelo HPX2000, em um computador com placa de som do modelo *Audigy II* (marca *Creative*). A escolha deste método deu-se para evitar vieses como ruídos no ambiente de avaliação, diferentes computadores, placas de som e fones de ouvido.

As avaliações por meio da escala visual analógica ocorreram semanalmente com duração de uma hora, por seis meses. Depois de cada avaliação, as vozes graduadas com diferença maior de 10 mm foram discutidas pelos juízes, que podiam ouvir novamente a amostra vocal e mudar a graduação da soproidade/rugosidade na escala visual analógica. Esse método possibilitou maior

confiabilidade entre os fonoaudiólogos e um treino contínuo de suas percepções. Após um intervalo de 30 dias realizou-se a avaliação por meio da escala numérica de quatro pontos, a qual foi realizada individualmente, sendo possível ouvir a voz novamente para saneamento de dúvidas intrapessoais.

Quanto a análise estatística foi aplicado o teste Coeficiente de Correlação Intraclass (CCI), que avaliou a concordância inter e intrajuiz, sendo que para essa última foram repetidas 10% da amostra. O CCI é uma medida de confiabilidade dos avaliadores/juízes, utilizada em estudos com variáveis contínuas, e seu cálculo é realizado pela razão da variância entre unidades de análise e a variância total, derivadas da análise de variância do teste Anova (LUQUETTI; LAGUARDIA, 2009). Na análise qualitativa do CCI considerou-se os intervalos descritos na tabela 1 (ROSNER, 1995). Como resultado final da análise de cada uma das vozes foi utilizado a mediana da perceptivo-auditiva dos quatro avaliadores.

Tabela 1 – Análise qualitativa dos valores do Coeficiente de Correlação Intraclass - CCI (ROSNER, 1995).

Quando o CCI for igual a:	A concordância é caracterizada como:
0 --- 0,4	Pobre
0,4 --- 0,75	Satisfatória
> 0,75	Excelente

Os resultados da escala visual analógica foram utilizados para o treino das redes neurais artificiais, pois esta permite ao clínico avaliar pequenos incrementos de rugosidade e soprosidade, sendo mais precisa na avaliação. Além disso, foi utilizada para o estabelecimento da curva de percepção auditiva juntamente com a escala numérica de quatro pontos.

5.3.1 Curva de percepção-auditiva e diferença permitida

A curva de percepção-auditiva prevê o estabelecimento de um valor clinicamente significativo mínimo, para cada ponto da escala visual analógica, por meio do qual se estabeleceu a taxa de acerto, em porcentagem, das RNAs.

Primeiramente, estipulou-se a nota de corte para a graduação da rugosidade e soprosidade por meio da curva ROC, a qual é baseada nos valores de sensibilidade, especificidade e eficiência. As notas de corte foram estimadas a partir da regra da máxima eficiência, considerando-se os

maiores valores de sensibilidade e especificidade, conforme mostra-se na seção de Resultados 6.1. A eficiência corresponde à área sob a curva ROC, sendo que quanto mais próximo do valor 1 mais precisos são os valores de corte traçados.

Na segunda etapa realizou-se a interpolação dos valores encontrados para as notas de corte para a correção destes, por meio de uma equação que representa a relação entre as escalas visual analógica e numérica. Em seguida, estimou-se uma equação para calcular a mínima diferença clinicamente significativa (seção 6.1), a qual foi chamada de diferença permitida ($Diff_p$), para cada valor específico da escala visual analógica de 0 a 100 mm.

5.4 Descrição do sistema classificador

O principal entrave para o sucesso na classificação de sinais geralmente está associado à parametrização. A redução da dimensão do sinal a uma coleção menor de parâmetros que ainda represente adequadamente os sinais originais muitas vezes é o processo mais importante. Posteriormente, com esses parâmetros adequados, torna-se mais fácil implementar um classificador, o qual muitas vezes é realizado por uma rede neural artificial.

Um problema relacionado ao processamento linear de sinais é a observação de um determinado parâmetro A corrompido por ruído aditivo, mas que depende deste parâmetro por meio de uma função arbitrária $f(t,A)$, em um intervalo $[0,T]$, ou seja: $y(t) = F(t,A) + n(t)$. Considerando um problema clássico de classificação de sinais onde tem-se $y(t)$, que pode pertencer a diversas categorias $\{C_j\}_{j=1}^M$, conforme o valor do parâmetro A , o qual, por exemplo, pode representar a amplitude de um sinal $s(t) = A \cdot f(t)$, cuja informação está na sua amplitude (KOVÁCS, 1996). Nesse trabalho pretende-se diagnosticar as condições vocais a partir do sinal de voz de um indivíduo, sendo que para cada categoria diagnosticada j é escolhida uma coleção de séries temporais representativas:

$$\{y_{i,j}(t)\}_{i=1}^{N_j} \quad j = 1, 2 \dots M. \quad (1)$$

Nesse contexto, pode-se tentar uma abordagem direta com redes neurais, treinando-se uma rede neural multicamada que terá como entrada todos os N pontos das séries temporais associadas a cada um dos sinais, objetivando classificá-las em cada uma das M categorias correspondentes. Contudo, essa abordagem impõe uma dimensão muito grande à rede neural, além de não evidenciar características importantes da voz.

Um dos paradigmas mais comuns para utilização de redes neurais em processamento de sinais pode ser resumido conforme descrito a seguir (KOVÁCS, 1996). Dado um conjunto de sinais que devem ser classificados em M categorias: $\{y_{i,j}(t)\}_{i=1}^{N_j} j = 1, 2 \dots M$, executam-se os seguintes passos:

I) Busca-se uma parametrização adequada. A particular parametrização escolhida determinará a eficiência e o desempenho do classificador. Essa etapa também é conhecida como obtenção do vetor de características do classificador.

II) Escolhe-se um conjunto de sinais representativos para o treinamento do classificador.

III) Obtém-se o vetor de característica associado a uma classificação pré-determinada.

IV) Opta-se por uma RNA avaliando suas características, como a dimensão das camadas e desempenho.

V) Treina-se a RNA.

VI) Avalia-se seu desempenho com um outro conjunto independente ou utiliza-se uma técnica estatística, como a validação cruzada.

O treinamento das redes neurais artificiais, bem como seus algoritmos de treinamento, foram configuradas pelas ferramentas do *Neural Network Toolbox* disponível no *Matlab*® (marca: *Math Works Inc.*) (DEMUTH; BEALE; HAGAN, 2007). Além disso, no processo de parametrização deste estudo foram utilizados parâmetros acústicos e a Transformada *Wavelet Packet* (WPT) como parte do vetor de características, sendo empregado tanto o uso isolado de cada uma das técnicas como a associação destas.

Quanto a configuração da RNA realizou-se inúmeros testes com diversas arquiteturas, número de camadas e de neurônios, obtendo-se a melhor arquitetura na identificação das vozes rugosas e soprosas contendo: na camada de entrada um número de neurônios correspondente ao número de parâmetros treinados/testados na RNA, os quais foram extraídos da análise acústica e Transformada *Wavelet Packet*; duas camadas neurais intermediárias com 15 e 4 neurônios, respectivamente; e dois neurônios na camada de saída, um correspondente à soproidade e outro à rugosidade. A arquitetura das RNAs pode ser visualizada por meio da Figura 3.

Figura 3 - Arquitetura das RNAs empregadas na identificação dos tipos de vozes.

5.5 Parametrização por meio da Transformada *Wavelet Packet*

A Transformada *Wavelet Packet* é uma generalização da Transformada Discreta *Wavelet* (DWT), em outras palavras, ela é um refinamento da DWT no domínio da frequência e é baseada no teorema matemático de Daubechies, suporte 2 (GOSWANI; CHAN, 1999). Na *Wavelet* o sinal é dividido em coeficiente de aproximação (filtro passa-baixa) e detalhe (filtro passa-alta), sendo que na DWT apenas os coeficientes de aproximação continuam a ser subdivididos enquanto que na WPT ambos os coeficientes, de aproximação e detalhe, continuam sendo subdivididos, decompondo o sinal em forma de árvore binária (Figura 4) (GOSWANI; CHAN, 1999).

Figura 4 - Árvore binária de decomposição da Transformada *Wavelet Packet*.

Para a extração das características vocais do sinal sonoro por meio de *WPT* foi utilizada a abordagem baseada na descrição de Goswami e Chan (GOSWANI; CHAN, 1999), onde realiza-se a decomposição do sinal da *WPT* até o quarto nível de decomposição obtendo 16 sub-bandas (Figura 5), cada uma com tamanho aproximado de 8000 pontos. O tamanho de cada sub-banda foi calculado por: $n_i = \frac{n_{sample}}{2^{n_{levels}}}$. Supondo um sinal de 3 segundos com taxa de amostragem igual

44.100 amostras/segundo, temos $n_{sample} = 3(44100)$. Assim $n_{sample} = \frac{3(44100)}{2^4}$; chegando no $n_i = 8268$ pontos por sub-banda.

Para cada sub-banda do quarto nível de decomposição calculou-se o teor médio de energia dos pontos de cada sub-banda da *WP* tais como:

$$v_i^{dp} = \frac{1}{n_i} \sum_{j=1}^{n_i} p_{i,j}^2, \quad i = 1, 2, \dots, 16 \text{ and } n_i = 8268, \quad (2)$$

onde v_i^{wp} é o i th elemento no vetor de características *WP*; n_i é o número de amostras de cada sub-banda, e $p_{i,j}$ são os pontos *WP* j -ésimo na i -ésima sub-banda. Como resultado, o vetor *WP* é representado por:

$$v^{wp} = \{v_1^{wp}, v_2^{wp}, \dots, v_{16}^{wp}\}^t. \quad (3)$$

Figura 5 – Exemplo do sinal de voz sem filtragem no nó (0,0), quadro B, e com filtragem por meio da técnica *Wavelet Packet* nos nós (4,0), quadro C, e (4,15) no quadro D.

A *Wavelet Packet* foi calculada pelas ferramentas disponíveis no *Matlab* (marca: *Math Works Inc.*) cuja otimização, baseada no critério de entropia, descrito por Coifman e Wickerhauser, seleciona a decomposição mais adequada do sinal. Isso significa que se analisa cada nó da árvore de decomposição e quantifica-se a informação a ser adquirida através da realização de cada divisão. Esta *Wavelet Toolbox* do *Matlab* utiliza um algoritmo de filtragem adaptativa, baseado no trabalho de Coifman e Wickerhauser (COIFMAN; WICKERHAUSER, 1992).

5.6 Parametrização por meio dos parâmetros acústicos

Na extração dos parâmetros acústicos, como parte do vetor de características, utilizou-se o Programa Análise de Voz, versão atualizada de 2014 (MONTAGNOLI, 1998). Os parâmetros acústicos analisados foram: frequência fundamental (F_0), *jitter* (coeficiente de perturbação de frequência), *shimmer* (coeficiente de perturbação de amplitude), amplitude do *pitch* (PA) (DAVIS, 1978) e derivada da amplitude (DA).

Para o cálculo do *jitter*, *shimmer* e frequência fundamental (F_0) realizou-se a correlação cruzada entre o sinal de voz e uma amostra dele próprio, sendo esta amostra com tamanho correspondente ao período (T) do sinal de voz analisado. Por meio dessa correlação cruzada os picos de correlação máxima ficam claramente evidenciados, tornando-se fácil a localização dos períodos do sinal (Figura 6). No cálculo específico do *Jitter* e do *Shimmer*, foram utilizados cinco períodos de uma janela amostral de 186 ms do sinal de voz, mantendo-se uma equivalência com o método proposto por Davis (DAVIS, 1978), enquanto que a F_0 foi calculada pela razão $1/T$ (ms).

No cálculo da amplitude da derivada (DA) foi realizado a derivada do sinal de voz e calculada a média de cinco picos da derivada de primeira ordem do sinal (Figura 7), enquanto que para o cálculo da Amplitude do *pitch* (PA) realizou-se a autocorrelação do sinal residual e mensurou-se a amplitude do segundo pico gerado por esta autocorrelação (DAVIS, 1978) (Figura 8). O sinal residual é resultado da filtragem inversa do sinal de voz, na qual filtra-se a interferência do trato vocal e do envelope espectral gerado pela glote (DAVIS, 1978; MONTAGNOLI, 1998), sendo composto pela F_0 e todos harmônicos com a mesma amplitude; enquanto que a autocorrelação é o cruzamento desse sinal residual com ele mesmo.

Destaca-se que todos os parâmetros acústicos foram extraídos da vogal sustentada /a/ editada em três segundos, mesma amostra da avaliação perceptivo-auditiva, e que o resultado obtido foi adquirido pela média dos valores encontrados em janelas amostrais de 186 ms. Todas as janelas foram inspecionadas visualmente e foram incluídas apenas as janelas que estavam detectando de forma correta os picos da derivada e da correlação do sinal, ou seja, que estavam identificando os períodos do sinal de voz corretamente. Além disso, o *software* permite o clínico indicar um valor aproximado da frequência fundamental, a qual ele corrige automaticamente, possibilitando a identificação correta dos períodos de sinais de vozes com sub-harmônicos (figura 9). O sub-harmônico é uma bifurcação de frequência fundamental, ou seja, ocorre quando há uma

ou mais linhas entre os harmônicos (múltiplos da F_0) (LEÃO, 2008). Os sub-harmônicos são comumente encontrados em vozes rugosas (LEÃO, 2008), especialmente as do tipo crepitante, e quando ocorrem abaixo da F_0 podem ser confundidos com essa, comprometendo a análise acústica. Com este procedimento essas dificuldades foram superadas.

Figura 6 – Observa-se no primeiro quadro a onda do sinal sonoro e no segundo quadro os picos evidenciados pela correlação, a partir dos quais se calcula a F_0 , o *jitter* e o *shimmer*.

Figura 7 – No primeiro quadro observa-se a onda do sinal sonoro e no segundo quadro os picos da derivada do sinal, a partir da qual se calcula a Amplitude da Derivada.

Figura 8 – No primeiro quadro observa-se a onda do sinal sonoro e no segundo quadro os picos da autocorrelação do sinal residual, sendo que a partir do segundo pico, circulado na imagem, se calcula a Amplitude do *Pitch*.

Figura 9 - A janela A refere-se à frequência fundamental encontrada inicialmente pelo programa, a janela B a F_0 indicada pelo fonoaudiólogo e a parte C à F_0 encontrada pelo programa, a partir da frequência que foi indicada, com os períodos identificados corretamente na autocorrelação do sinal de voz.

A metodologia estabelecida para o clínico indicar a F_0 ao programa de análise acústica foi: primeiramente, observar se os picos da autocorrelação do sinal haviam sido identificados corretamente (exemplo: Figura 6). Caso esta identificação estivesse errada, como exemplificado na figura 8 (janela A e B), foi realizada uma medida do período no sinal de voz, que é apresentado no domínio do tempo, em milissegundos (ms). Esse cálculo foi realizado pela diferença entre o final e o início de um período (T) e assim, aplicou-se a equação $\left(\frac{1}{T}\right) 1000$, encontrando-se a F_0 em

Hz (Figura 10). Posteriormente analisou-se o sinal de voz por meio da espectrografia, na qual se verificou o valor da F_0 e o valor do sub-harmônico para confirmação, respectivamente, do período calculado manualmente na janela do sinal de voz e da F_0 indicada inicialmente pela análise acústica, com valor próximo ao do sub-harmônico (Figura 11).

Figura 10 – A parte A da figura exemplifica a metodologia encontrada para verificação da frequência fundamental, a F_0 encontrada inicialmente pelo programa de 82,6 Hz e a F_0 calculada manualmente de 166,7

Hz. Na parte B observa-se, após indicação da F_0 calculada pelo clínico, que o programa encontra a F_0 de 165,2 Hz e identifica corretamente os períodos do sinal de voz na autocorrelação do sinal.

Figura 11 - Análise espectrográfica de uma voz feminina, crepitante intensa e soprosa leve, revelando: sub-harmônicos assistemáticos na frequência de 80 Hz (marcados pelo símbolo +), frequência identificada inicialmente pela análise acústica como F_0 na figura 9; e frequência fundamental (F_0) em torno de 160 Hz, valor próximo ao encontrado no cálculo realizado pelo clínico no sinal de voz na figura 9.

A correlação dos parâmetros acústicos com a avaliação perceptivo-auditiva foi realizada por meio do teste de *Pearson*, para testar quais eram mais apropriados para a parametrização da RNA. Para qualificar a associação foi utilizado o coeficiente de correlação (R), cuja medida representa a associação entre duas variáveis quantitativas, e a avalia de forma qualitativa, quando presente, conforme apresentado na Tabela 2.

Tabela 2 - Avaliação qualitativa do grau de correlação entre duas variáveis (CALLEGARI-JACQUES, 2003).

Quando /R/ for igual a:	a correlação é caracterizada como:
0	Nula
0 ---- 0.3	Fraca
0.3 ---- 0.6	Regular
0.6 ---- 0.9	Forte
0.9 ---- 1	Muito forte
1	Plena ou perfeita

5.6 Validação do sistema classificador

Utilizou-se para o treino do classificador a *aprendizagem supervisionada*, definida desta forma por conhecermos a natureza da amostra antes de analisá-la pelas redes neurais, assim cada amostra que é apresentada ao classificador foi previamente categorizada (*class label*) (DUDA; HART; STORK, 2001b, 2001c). Nesta pesquisa apresentamos ao sistema classificador as amostras editadas da vogal sustentada /a/, sendo que para cada amostra foi apresentado o respectivo resultado da avaliação perceptivo-auditiva realizada pelos fonoaudiólogos, quanto ao grau do desvio vocal e natureza do mesmo [neutro, soproso e/ou rugoso (crepitante, rouco e/ou áspero)]. Assim, por meio da *backpropagation* (BP), o sistema classificador ajusta os pesos sinápticos para chegar a uma classificação mais próxima da apresentada, minimizando os erros (DUDA; HART; STORK, 2001a; HAYKIN, 2001).

Além disso, para minimizar os custos computacionais do sistema na definição dos padrões também se apresentou ao classificador o vetor de características, que são as características específicas do sinal de voz extraídas pela *Wavelet* e as características extraídas pela análise acústica. O sistema proposto pode ser resumido na figura 12.

Figura 12 - Diagrama de blocos da extração de características e fase de classificação.

A fase de teste foi simultânea à fase de treino, para a efetuação de ambas utilizou-se o método de validação cruzada *Leave-one-out*, no qual se extrai uma voz do conjunto de dados, treina-se com o restante do conjunto e testa-se com esta única voz que não participou do treino (RON KOHAVI, 1995). Este método foi repetido com todas as vozes do conjunto, sendo registradas as respostas do sistema classificador nos testes.

Na avaliação do desempenho e validação do sistema classificador foram comparadas, de forma pareada, os resultados da RNA com os da avaliação perceptivo-auditiva, por meio da taxa de acerto, em porcentagem. Considerou-se acerto quando a diferença entre o resultado apresentado pela rede neural e a avaliação perceptivo-auditiva foi menor ou igual a mínima diferença clinicamente significativa, a qual foi chamada de diferença permitida e representadas pela equação $Dif_p = \sqrt{5y}$. Destaca-se que o processo para encontrar a fórmula da diferença permitida está descrita no item resultados.

Para complementar a análise do desempenho das RNAs foram calculados a ‘média absoluta da diferença’, diferença entre os resultados do sistema automático e os da avaliação perceptivo-auditiva, e aplicado o teste Coeficiente de Correlação Intraclass (Tabela 1), que avaliou a concordância entre a RNA e o valor da perceptivo-auditiva utilizada no treino desta.

6. RESULTADOS

Das 150 vozes avaliadas 27 foram excluídas: três por estarem gravadas com uma taxa de amostragem diferente de 44.100 Hz, tornando-se inadequadas para a extração dos parâmetros acústicos e da *WPT*; uma pela impossibilidade da extração correta dos parâmetros acústicos devido ao grau intenso de desvio vocal (Figura 13) e sem periodicidade aparente (TITZE, 1995); e 23 foram excluídas por não responderem aos critérios de inclusão da amostra na avaliação perceptivo-auditiva, das quais três vozes foram excluídas por falta de consenso quanto à característica predominante do desvio vocal, onze por diferença maior que 10 mm entre dois juízes ou mais, quatro por não terem uma característica vocal predominante e quatro por não terem o parâmetros rugosidade ou soprosidade como predominante, sendo três predominantemente tensas e uma predominantemente instável. Desta forma, 123 vozes foram utilizadas na realização deste estudo.

Figura 13 – Voz com onda tipo III, avaliada pela perceptivo-auditiva com 94 de rugosidade (tipo aspereza) e 57 de soprosidade (escala de 0 a 100 mm). No quadro A tem-se o sinal da onda sonora, no quadro B a correlação, no quadro C a derivada do sinal e no quadro D a autocorrelação do sinal residual.

6.1 Resultados da avaliação perceptivo-auditiva

Quanto à avaliação perceptivo-auditiva as vozes foram classificadas: 12 como quase neutras, com variação média de 10 mm de R ou S; 38 predominantemente soprosas; e 73 predominantemente rugosas, das quais 34 tinham a rouquidão como parâmetro predominante da rugosidade, 16 tinham a aspereza e 23 a crepitação. A concordância inter e intrajuiz foram classificadas como excelente (Tabela 1), ambas avaliadas por meio do CCI, encontrando-se $p = 0,85$ na concordância interjuizes e p variando de 0,87 a 0,93 nas concordâncias intrajuiz.

Em relação ao estabelecimento da curva de percepção auditiva primeiramente foram encontradas as notas de corte das diferentes intensidades de soprosidade e rugosidade, as quais foram semelhantes para ambos parâmetros (Tabelas 2 e 3). A partir da interpolação das notas de corte encontrou-se a curva de percepção-auditiva da rugosidade e soprosidade, revelando curvas de função quadrática (Figura 14). Pela semelhança da percepção entre a rugosidade e soprosidade, calculou-se a média entre suas notas de corte (grau 0, 1, 2 e 3) e estabeleceu-se uma nova curva de função quadrática (Figura 14), a partir da qual extraiu-se a equação:

$$y = 5,25x^2 + 3,75x, \quad (4)$$

onde y é o valor encontrado na escala visual analógica (0 a 100 mm) e x o grau do parâmetro avaliado (graus 0, 1, 2 e 3). A partir desta função encontrou-se as seguintes faixas de distribuição para ambos os parâmetros: grau 0 – 0 a 9 mm; grau 1 – 9 a 29 mm; grau 2 de 29 a 59 mm; e grau 3 – 59 a 99 mm. Observa-se que através destas faixas de distribuição não se atinge a pontuação máxima de 100 mm da escala visual analógica, desta forma, corrigiu-se a equação encontrada por meio da aproximação para valores múltiplos de dez, gerando-se a seguinte equação quadrática:

$$y = 5x^2 + 5x, \quad (5)$$

onde y é o valor encontrado na escala visual analógica (0 a 100 mm) e x o grau do parâmetro avaliado nos graus 0, 1, 2 e 3. Desta forma, as notas de corte da rugosidade e soprosidade passaram a ter valores (Tabela 4) com início em 0 mm e final em 100 mm, respeitando a variação de pontos da escala visual analógica.

Tabela 3 – Notas de corte para os diferentes graus de rugosidade e soproosidade com os respectivos valores de eficiência.

Grau	Rugosidade				Soproosidade			
	Notas de Corte	Sensibilidade	Especificidade	Eficiência	Notas de Corte	Sensibilidade	Especificidade	Eficiência
1	8,5	0,879	1,000	0,940	8,5	0,921	1,000	0,961
2	28,5	0,847	0,818	0,833	33,5	0,750	0,974	0,862
3	59,5	1,000	0,915	0,958	52	1,000	0,868	0,934

Tabela 4 - Faixas de distribuição dos graus de rugosidade e soproosidade, em quatro pontos.

Grau de presença do parâmetro avaliado	Faixa de distribuição	
	Rugosidade	Soproosidade
0	0 - 8,5	0 - 8,5
1	8,5 - 28,5	8,5 - 33,5
2	28,5 - 59,5	33,5 - 52,5
3	59,5 - 100	52,5 - 100

Tabela 5 – Faixas de distribuição, dos graus de rugosidade e soproosidade, a partir das notas de corte obtidas da equação quadrática proposta pela interpolação dos dados.

Faixas de distribuição da Rugosidade e Soproosidade	
Grau	Faixas de distribuição
0	0 – 10 mm
1	11 – 30 mm
2	31 – 60 mm
3	61 – 100 mm

Figura 14 – Curva de regressão quadrática da percepção da rugosidade, soprosidade e da equação quadrática proposta pela interpolação dos dados.

Para gerar a equação do cálculo do diferença permitida (Dif_p) foram consideradas: as notas de corte da equação quadrática, ou seja 10, 30 e 60 mm; os valores centrais da faixa de distribuição de cada grau: 5, 20, 45 e 80 mm; e o erro em cada grau para que o valor continue na mesma faixa de distribuição: ± 5 , ± 10 , ± 15 e ± 20 mm (Figura 15). Exemplificando, na faixa de grau 0, que varia de 0 a 10 mm, o valor central da faixa é 5 mm e a variação permitida, para que continue na mesma faixa, é de ± 5 mm. Assim, se um fonoaudiólogo avalia a voz em 5 mm o outro pode marcar até 10 mm (+ 5) ou até 0 mm (- 5) para ser considerado acerto. Na faixa de grau 1, o valor central é 20 mm, assim se um fonoaudiólogo avalia a voz em 20 mm, o outro poderá avaliar até 30 mm (+10 mm) ou 10 mm (-10 mm), sendo utilizado o mesmo raciocínio para os demais graus.

Figura 15 – Régua com os valores utilizados para estimar-se a equação quadrática da diferença permitida, ou seja, com as notas de corte corrigidas, valores centrais e diferença permitida dentro de cada faixa de distribuição.

A partir da interpolação destes dados gera-se a equação quadrática:

$$y = \frac{x^2}{5}, \quad (6)$$

onde y representa os valores centrais de cada faixa de distribuição e x representa o valor da diferença permitida de acordo com estes valores centrais. Invertendo-se a equação apresentada obteve-se a equação da diferença permitida:

$$Dif_p = \sqrt{5y}, \quad (7)$$

onde y é o valor encontrado na avaliação perceptivo-auditiva e Dif_p é a diferença permitida, ou seja, a mínima diferença clinicamente significativa, para o valor específico encontrado.

6.2 Resultados da análise acústica

Na seleção dos parâmetros mais adequados à parametrização da RNA aplicou-se o teste de correlação de *Pearson* entre cada parâmetro acústico e a avaliação perceptivo-auditiva da rugosidade e da soproidade. Os resultados não mostraram correlação entre nenhum parâmetro acústico isolado com os parâmetros perceptivos (valor de p : 0,233 a 0,908). No entanto, a combinação do *jitter*, *shimmer*, PA e F_0 formaram um subconjunto relevante na discriminação e graduação das vozes (descrito no próximo item, 6.3).

6.3 Resultados dos treinos e testes das Redes Neurais Artificiais

Com o intuito de se estimar a taxa de acerto aleatória da rede neural artificial realizou-se 1.000 testes com cada uma das 123 vozes, sendo gerado um valor randômico de 0 a 100, simulando

uma avaliação aleatória na escala visual analógica, e comparado com a análise perceptivo-auditiva, mediana dos quatro juízes. Destes testes obteve-se 21% de acerto, representando a taxa de acerto ao acaso, se a rede neural artificial não fosse utilizada.

Diversos treinos e testes foram realizados com as RNAs com o intuito de encontrar os melhores resultados. Primeiramente, treinou-se os bancos de vozes separadamente, de acordo com o tipo vocal: predominantemente soprosas separadamente das rugosas, posteriormente subdividiu-se também as rugosas quanto ao tipo de rugosidade predominante na amostra: rouco, áspero ou crepitante. Selecionou-se esta abordagem na tentativa de facilitar o treino das RNAs, visto que estas vozes possuem características acústicas diferentes (BEHLAU et al., 2001). Em seguida o treino foi realizado com todo o banco de vozes junto, tendo a vantagem de um número maior de amostras para o treino/teste. A parametrização utilizada foi a entropia do quarto nível de decomposição da Transformada *Wavelet Packet* e os parâmetros acústicos: *jitter*, *shimmer*, derivada da amplitude (DA), amplitude do *pitch* (PA) e frequência fundamental (F_0), conforme descrito na metodologia deste trabalho.

Na avaliação do desempenho da RNA foram realizados três treinos/testes para cada parametrização testada, calculando-se a taxa de acerto, por meio da fórmula da diferença permitida ($\sqrt{5y}$), a confiabilidade pelo teste estatístico CCI e o erro médio, pela média do módulo da diferença entre os resultados da RNA e os da avaliação perceptivo-auditiva, considerando-se a mediana dos juízes. Após a análise destes resultados observou-se que o treino de todas as vozes juntas foi a melhor forma de composição do banco de dados para os treinos/testes, considerando-se a melhor taxa de acerto para ambos, rugosidade (67%) e soprosidade (64%) (Figura 16), e o menor erro médio entre os três testes (12 mm para rugosidade e soprosidade) (Tabela 5). Além disso, a variação da taxa de acerto nos 3 testes realizados com cada parametrização foi mais consistente no treino com todas as vozes juntas (Tabela 5).

Figura 16 – Maior taxa de acerto ($\sqrt{5y}$) e confiabilidade (CCI), em três testes, da rede neural artificial de acordo com o banco de vozes utilizado para o treino e teste.

Tabela 6 – Variação do erro médio e da taxa de acerto da rede neural artificial de acordo com o banco de vozes utilizado no treino/teste, em três testes.

Variação do erro médio e da taxa de acerto da RNA para diferentes composições dos dados de treino/teste								
Banco de vozes	Erro médio (mm)				Taxa de acerto			
	Soprosidade		Rugosidade		Soprosidade		Rugosidade	
	Menor	Maior	Menor	Maior	Menor	Maior	Menor	Maior
Soprosas	12	15	12	16	47%	59%	33%	57%
Rugosas	12	13	12	15	57%	57%	67%	68%
Roucas	12	14	8	10	48%	52%	60%	65%
Ásperas	13	19	13	19	46%	54%	54%	58%
Crepitantes	14	15	11	17	38%	53%	53%	63%
TODAS	12	12	11	12	63%	64%	67%	69%

Após determinar a melhor forma de composição do banco de dados, constituído por todas as vozes juntas, foram testadas diferentes parametrizações da RNA. Para verificar qual o melhor nível de decomposição da Transformada *Wavelet Packet* na avaliação das vozes treinou-se a RNA com diferentes níveis de decomposição da WPT. Observou-se maior taxa de acerto e confiabilidade (CCI) na WPT com 1 nível de decomposição, sendo que o desempenho da RNA diminuiu gradualmente com o aumento do número de níveis de decomposição (figura 17), ocorrendo o mesmo quanto ao erro médio da RNA (tabela 6).

Figura 17 – Maior taxa de acerto ($\sqrt{5y}$) e confiabilidade (CCI), em três testes, da rede neural artificial para diferentes níveis de decomposição da Transformada *Wavelet Packet* (WPT) e parâmetros acústicos fixos: *jitter*, *shimmer*, derivada da amplitude, amplitude do *pitch* e frequência fundamental.

Tabela 7 - Variação do erro médio e da taxa de acerto da RNA, em três treinos, no treino/teste variando o nível de decomposição da Transformada *Wavelet Packet*.

Variação do erro médio e da taxa de acerto da RNA para diferentes níveis de decomposição da Transformada <i>Wavelet Packet</i>								
Nível de decomposição da WPT	Erro médio				Taxa de acerto			
	Soproisidade		Rugosidade		Soproisidade		Rugosidade	
	Menor	Maior	Menor	Maior	Menor	Maior	Menor	Maior
1	10	11	11	11	65%	66%	65%	74%
2	11	12	11	12	63%	65%	67%	68%
3	11	12	13	13	63%	63%	67%	69%
4	12	12	12	13	62%	64%	59%	69%
5	12	14	14	15	59%	63%	60%	60%
6	12	16	14	15	49%	62%	55%	56%
7	12	15	16	18	55%	60%	50%	52%

Após a definição da melhor árvore da WPT, que foi com 1 nível de decomposição, realizou-se uma pesquisa para identificar o melhor subconjunto de parâmetros acústicos na graduação da rugosidade e da soproosidade. Para esta pesquisa treinou-se a rede neural artificial com diferentes combinações de parâmetros acústicos e sem a Transformada *Wavelet Packet*, sendo:

- *Jitter*, *shimmer*, derivada da amplitude (DA), amplitude do *pitch* (PA) e frequência fundamental (F_0);
- *Jitter*, *Shimmer*, DA e PA;
- *Jitter*, *shimmer*, PA e F_0 ;
- *Jitter*, *shimmer*, DA e F_0 ;
- *Jitter*, *shimmer* e PA;
- *Jitter*, *shimmer* e DA;
- *Jitter* e *shimmer*;
- DA, PA e F_0 ;
- *Jitter*, DA, PA e F_0 ;
- *Jitter*, PA e F_0 ;
- *Jitter*, DA e F_0 ;
- *Shimmer*, DA, PA e F_0 ;
- *Shimmer*, PA e F_0 ;
- *Shimmer*, DA e F_0 ;

Os resultados variaram de 43% a 74% de acerto na soproosidade e de 42% a 72% de acerto na rugosidade de acordo com o conjunto de parâmetros utilizados, enquanto que o CCI variou de 0,32 a 0,80 na soproosidade e de 0,55 a 0,88 na rugosidade (figura 18). Além disso, o erro médio da RNA variou de 9 a 19 mm na soproosidade e 10 a 18 mm na rugosidade (tabela 7). Os três melhores resultados foram obtidos nos subconjuntos: *Jitter*, *shimmer*, PA e F_0 ; *Jitter*, PA e F_0 ; e *Shimmer*, DA, PA e F_0 .

Taxa de acerto e confiabilidade da RNA com diferentes subconjuntos de parâmetros acústicos

Figura 18 – Melhor taxa de acerto ($\sqrt{5y}$) e confiabilidade (CCI) da rede neural artificial, em três testes, para diferentes subconjuntos de parâmetros acústicos e sem os parâmetros extraídos da WPT.

** Subconjuntos de parâmetros acústicos com os melhores resultados da RNA na discriminação do grau e tipo de voz.

Tabela 8 - Variação do erro médio e da taxa de acerto da RNA, em três treinos/testes, com diferentes subconjuntos de parâmetros acústicos.

Variação do erro médio e da taxa de acerto da RNA em parametrizações com diferentes subconjuntos de parâmetros acústicos								
Parametrização	Erro médio				Taxa de acerto			
	Soprosidade		Rugosidade		Soprosidade		Rugosidade	
	Menor	Maior	Menor	Maior	Menor	Maior	Menor	Maior
<i>Jitter, Shimmer, DA, PA e F₀</i>	11	11	11	11	64%	65%	72%	72%
<i>Jitter, Shimmer, DA e PA</i>	12	12	11	12	61%	63%	60%	67%
<i>Jitter, shimmer, DA e F₀</i>	16	17	15	15	43%	50%	51%	55%
<i>jitter, shimmer, PA e F₀ **</i>	10	11	10	11	65%	71%	68%	72%
<i>Jitter, shimmer e PA</i>	12	13	12	12	60%	65%	60%	65%
<i>Jitter, shimmer e DA</i>	16	17	13	15	48%	50%	54%	59%
<i>Jitter e shimmer</i>	16	16	14	15	45%	46%	54%	57%
<i>DA, PA e F₀</i>	10	11	10	13	66%	70%	61%	70%
<i>Jitter, DA, PA e F₀</i>	10	11	11	12	63%	69%	58%	70%
<i>Jitter, PA e F₀ **</i>	9	10	11	11	68%	74%	69%	70%
<i>Jitter, DA e F₀</i>	15	16	15	18	47%	52%	42%	56%
<i>Shimmer, DA, PA e F₀ **</i>	10	11	10	13	67%	71%	60%	72%
<i>Shimmer, PA e F₀</i>	10	11	11	11	63%	68%	65%	67%
<i>Shimmer, DA e F₀</i>	14	19	14	14	45%	56%	55%	57%

** Subconjuntos de parâmetros acústicos com os melhores resultados da RNA na discriminação do grau e tipo de voz.

Para finalizar os treinos/testes realizou-se a combinação do melhor resultado da Transformada *Wavelet Packet* com os três melhores resultados dos subconjuntos de parâmetros acústicos. Desta forma foram testados os seguintes subconjuntos de parâmetros:

- 1 nível de decomposição na WPT, *Jitter, shimmer, PA e F₀*;
- 1 nível de decomposição na WPT, *Jitter, PA e F₀*;
- 1 nível de decomposição na WPT, *Shimmer, DA, PA e F₀*;

Os treinos/testes apontaram melhor resultado para a rugosidade no subconjunto: 1 nível de decomposição na WPT, *Jitter, shimmer, PA e F₀*, na qual encontrou-se 73% de acerto, confiabilidade excelente – CCI = 0,84, erro médio de 10 mm e taxa de acerto consistente nos três testes, enquanto que para a soprosidade o desempenho foi semelhante nesse subconjunto. No subconjunto: 1 nível de decomposição na WPT, *Jitter, PA e F₀*, houve um decréscimo pequeno do desempenho da RNA, e no terceiro subconjunto houve diminuição de aproximadamente 10% no desempenho para a rugosidade. Em relação à soprosidade o desempenho da RNA foi semelhante para o subconjuntos 1 nível de decomposição na WPT, *Jitter, shimmer, PA e F₀*; e inferior nos

subconjuntos: 1 nível de decomposição na WPT, *Jitter*, PA e F₀; e 1 nível de decomposição na WPT, *Shimmer*, DA, PA e F₀, na adição da WPT com 1 nível de decomposição nos testes. O resultado detalhado destes testes encontram-se na figura 19 e tabela 8.

Figura 19 – Melhor taxa de acerto ($\sqrt{5y}$) e confiabilidade (CCI) da rede neural artificial, em três testes, na parametrização com um nível de decomposição da Transformada *Wavelet Packet* juntamente com os três melhores subconjuntos de parâmetros acústicos na discriminação do grau e tipo das vozes.

Tabela 9 - Variação do erro médio e da taxa de acerto da RNA, em três treinos/testes, com uma nível de decomposição da Transformada *Wavelet Packet* juntamente com os três melhores subconjuntos de parâmetros acústicos na discriminação do grau e tipo das vozes.

Erro médio e variação da taxa de acerto da RNA em parametrizações com 1 nível de decomposição da WPT combinada com os três melhores subconjuntos de parâmetros acústicos								
Parametrização	Erro médio				Taxa de acerto			
	Soproisidade		Rugosidade		Soproisidade		Rugosidade	
	Menor	Maior	Menor	Maior	Menor	Maior	Menor	Maior
WPT1, <i>jitter</i> , <i>shimmer</i> , PA e F ₀	11	11	10	11	63%	70%	70%	73%
WPT1, <i>Jitter</i> , PA e F ₀	10	11	10	11	63%	71%	70%	70%
WPT1, <i>Shimmer</i> , DA, PA e F ₀	11	11	11	11	64%	64%	63%	66%

Após a análise de todos os resultados o melhor desempenho da rede neural artificial para a soprosidade ocorreu no subconjunto: *jitter*, amplitude do *pitch* e frequência fundamental, no qual obteve-se taxa de acerto de 74%, concordância excelente com a avaliação perceptivo-auditiva da escala visual analógica (0,80 no CCI) e erro médio de 9 mm. Para a rugosidade o melhor subconjunto foi: Transformada *Wavelet Packet* com 1 nível de decomposição, *jitter*, *shimmer*, amplitude do *pitch* e frequência fundamental, no qual obteve-se 73% de acerto, concordância excelente (0,84 no CCI), e erro médio de 10 mm. Destaca-se que a maior taxa de acerto para a rugosidade foi de 74% e ocorreu no subconjunto: 1 nível de decomposição na WPT, *jitter*, *shimmer*, amplitude do *pitch*, derivada da amplitude e frequência fundamental. Apesar disso, a consistência da rede neural artificial foi menor neste subconjunto, cujo desempenho variou 65% e 74% (tabela 6); além disso, a confiabilidade também foi inferior (CCI = 0,80) e houve aumento do erro médio para 11 mm. Assim, apesar de maior taxa de acerto em 1%, no restante dos aspectos avaliados, este subconjunto foi inferior ao subconjunto que obteve 73% de acerto. Além disso, ao comparar-se o desempenho da RNA com o dos fonoaudiólogos, juízes dessa pesquisa, observou-se semelhança entre os resultados, tanto na avaliação da soprosidade (figura 20) como na rugosidade (figura 21).

Figura 20 – Avaliação da soprosidade pelos fonoaudiólogos, juízes da pesquisa, e pela rede neural artificial.

Figura 21 - Avaliação da rugosidade pelos fonoaudiólogos, juízes da pesquisa, e pela rede neural artificial.

7. DISCUSSÃO

7.1 Perceptivo-auditiva e curva de percepção-auditiva

O treino adequado das redes neurais artificiais, quando se utiliza a abordagem supervisionada, depende de dados consistentes e fidedignos da avaliação utilizada para a supervisão. Nesta pesquisa, para o treino supervisionado da RNA foi empregada a avaliação perceptivo-auditiva, a qual apresentou, por meio do teste CCI, resultados excelentes de confiabilidade tanto intra como interjuízes, revelando uma avaliação perceptivo-auditiva robusta e consistente.

A partir da avaliação perceptivo-auditiva estabeleceu-se uma curva de percepção auditiva, estudo motivado pela dificuldade encontrada atualmente na avaliação da concordância intra e interjuízes na escala visual analógica de 100 mm, cujos métodos não consideram a não-linearidade da percepção dos desvios vocais, entre outras dificuldades. O teste Coeficiente de Correlação Intraclass (CCI), utilizado nesta pesquisa, deve ser empregado com cuidado de acordo com as amostras vocais avaliadas. Seu cálculo está vinculado a variância total dos dados apresentados (LUQUETTI; LAGUARDIA, 2009), assim se a amostra da pesquisa não englobar vozes com os diversos graus de desvio vocal, os quais estendem-se de 0 a 100 milímetros (mm), corre-se o risco de obter-se uma avaliação errônea da concordância, por meio do CCI, que será calculada baseado apenas na variação dos dados apresentados e não a real variação de 100 pontos. Outro método empregado atualmente para verificar a concordância intra e interjuízes considera empiricamente, em uma escala de 0 a 100 mm, o valor fixo de 10 mm como mínima diferença clinicamente significativa (EADIE et al., 2010), ou seja, em diferenças menores que 10 mm considera-se que há concordância entre os avaliadores, independente do grau do desvio vocal, desconsiderando a não-linearidade da percepção-auditiva e gerando resultados contestáveis de concordância. Desta forma, com a análise da percepção-auditiva para os diferentes desvios vocais espera-se encontrar um método mais eficiente de avaliação da concordância.

Os resultados da análise da percepção-auditiva dos fonoaudiólogos apontaram uma curva quadrática de percepção dos desvios vocais, o que significa, explicando de maneira sintetizada, que é necessário duplicar o desvio vocal para que ele seja percebido de forma linear. A percepção não-linear do desvio vocal é condizente com o comportamento do ouvido humano, que também percebe

a variação de outras grandezas como a intensidade e a frequência do sinal sonoro não-linearmente (MOORE, 2012). A percepção auditiva da frequência estende-se de 20 a 20.000 Hz, havendo maior sensibilidade na percepção dos sons de 1.000 a 4.000 Hz (BONALDI et al., 2004), enquanto que a percepção da intensidade varia de 10^{-11} a 10^{-5} metros (WALKER, 2008), em outras palavras, a amplitude varia em uma faixa de intensidade compreendida entre de 0 dB (limiar da audição) e 120/140 dB (sons desconfortáveis e que podem causar lesão), lembrando que a percepção da intensidade também varia de acordo com a frequência (BONALDI et al., 2004).

Tanto o aumento da frequência como o da intensidade são logarítmicos e obedecem a Lei de Weber de estímulo/sensação, que quantifica a relação entre a intensidade do estímulo físico e o seu efeito percebido (FERNANDES, 2010). Desta forma, o aumento da frequência é percebido pela relação entre as frequências e não pelo intervalo entre elas, exemplificando: o intervalo entre 100 e 200 Hz é percebido como igual ao intervalo entre 200 e 400 Hz ou ainda entre 400 e 800 Hz, apesar destes aumentos serem exponenciais o ouvido humano percebe apenas a relação entre as frequências, que no caso é igual a dois e representa o intervalo de uma oitava (FERNANDES, 2010). O mesmo ocorre com a percepção do aumento da intensidade, ou seja, quanto mais aumentamos a intensidade do sinal sonoro menos sensível fica nossa percepção, assim, para se ter a sensação de que o aumento é linear a intensidade deve ser aumentada exponencialmente (FERNANDES, 2010).

A curva encontrada na percepção dos desvios vocais, portanto, é condizente com a não-linearidade nas percepções da intensidade e frequência vocal, revelando estreita relação entre estas variáveis e reforçando a validade dos resultados encontrados. Além disso, acredita-se que essa curva encontrada para a rugosidade e soproidade seja a mesma para os demais parâmetros perceptivos, sugerindo-se pesquisas futuras para a confirmação dessa curva de percepção.

Quanto à diferença permitida (Dif_p), sua equação foi estimada baseando-se nos valores encontrados na curva quadrática da avaliação perceptivo-auditiva, no entanto, para valores menores que 5 nota-se que a Dif_p é muito pequena, podendo chegar a zero. Assim, adotou-se o valor mínimo de 5 mm para a diferença, ou seja, se $Dif_p < 5$ este valor retorna a cinco. Desta forma, evita-se falsos negativos no julgamento do erro/acerto entre avaliações.

Sabe-se que as pesquisas que envolvem avaliações perceptivo-auditivas, geralmente tem mais de um juiz (BELE, 2005; ENGLERT et al., 2015; FINK et al., 2015; MADAZIO; LEÃO; BEHLAU, 2011; MARYN; KIM; KIM, 2015; MARYN et al., 2010; NEMR et al., 2012;

TEIXEIRA; BEHLAU, 2015), não havendo um consenso quanto ao número de juízes indicado. Além disso, cada pesquisa apresenta o valor final da perceptiva de um modo diferente, algumas utilizam a média dos juízes (FINK et al., 2015), outras a mediana destes, e algumas utilizam a avaliação do juiz com melhor confiabilidade intrajuiz no teste estatístico (TEIXEIRA; BEHLAU, 2015). Observa-se, entre os artigos levantados, que poucos descreveram como obtiveram o valor final da perceptivo-auditiva.

Com base nestas variações metodológicas sugere-se que a utilização da diferença permitida proposta seja realizada de acordo com o valor final da perceptivo-auditiva utilizado por cada pesquisa, ou seja, se a pesquisa apresentou a média entre os juízes a Dif_p deverá ser calculado a partir dos valores médios, se utilizou a mediana deverá utilizar o valor da mediana para o cálculo e se utilizou as notas do juiz com melhor confiabilidade intrajuiz deverá utilizar estas notas. Desta forma, averiguar se os juízes acertaram ou erraram, de acordo com a equação proposta, dependerá da metodologia abordada por cada pesquisador.

Sugerem-se estudos futuros, com um grupo maior de avaliadores para confirmar a curva quadrática encontrada. Além disso, sugere-se pesquisas com vozes sintetizadas, que possa mensurar a percepção do fonoaudiólogo na mudança de incrementos semelhantes de desvio vocal. Assim, tem-se a possibilidade de traçar de forma mais detalhada e objetiva a não-linearidade da percepção dos desvios vocais.

7.2 Análise acústica

Na análise acústica, conforme descrito por Titze, apenas sinais do tipo I poderiam ser avaliados de forma confiável pela análise acústica linear quando se trata de medidas dependentes da identificação do período do sinal, com as medidas de perturbação *jitter* e *shimmer* (TITZE, 1995). No entanto, neste trabalho a análise acústica foi semiautomática, ou seja, todas as janelas amostrais foram conferidas visualmente e foram selecionadas para o cálculo dos parâmetros apenas aquelas cujos períodos do sinal sonoro foram identificados corretamente. Esta metodologia permitiu a mensuração de ondas tipo I, II e III de forma confiável, sendo apenas uma voz com disфонia intensa excluída do banco de dados.

A possibilidade de uma análise acústica semiautomática permitiu a análise confiável de ondas tipo II e tipo III, o que não ocorre nos programas de análise automática. Além disso, os programas tradicionais não permitem que o clínico visualize quais períodos foram identificados do sinal sonoro, o que diminui a confiabilidade dos valores fornecidos na análise acústica. Desta forma, o programa semiautomático mostra-se mais confiável e eficiente na avaliação de vozes tipo II e tipo III. Diante disso, sugere-se que pesquisas futuras investiguem essa nova proposta de análise acústica.

Quanto a seleção dos parâmetros acústicos mais apropriados na parametrização da RNA foi aplicado o teste de correlação de *Pearson*, o qual não apresentou resultados promissores, visto que nenhum dos parâmetros acústicos, relacionados de forma isolada, obtiveram correlação com a rugosidade ou soproidade. Este resultado é esperado e concorda com os encontrados na literatura, como os estudos de Bhuta et al. (2004) e Madazio et al. (2011), os quais não encontram correlação do uso isolado do *jitter* e do *shimmer* com a avaliação perceptivo-auditiva. No entanto, sabe-se que há controvérsias na relação do *jitter* e do *shimmer* com parâmetros perceptivos, encontrando-se pesquisas que obtiveram correlação do *jitter* com aspereza (BEHLAU et al., 2001), rouquidão (WOLFE; MARTIN, 1997) ou soproidade (DEJONCKERE, 1995; VAZ FREITAS et al., 2015). Enquanto que para o *shimmer* há pesquisas que o encontraram correlacionado com a rouquidão e soproidade (BEHLAU et al., 2001; LOPES et al., 2012) ou apenas com a rouquidão (VAZ FREITAS et al., 2015). A discrepância destes resultados pode estar envolvida com uma série de fatores, a começar pela forma de gravação vocal (DEJONCKERE, 2010) e as características distintas das diversas avaliações perceptivo-auditivas e acústicas. Além disso, sabe-se que a voz é um som complexo, a qual dificilmente apresenta uma única característica de desvio, por isso a dificuldade na busca da correlação de parâmetros acústicos com perceptivos. Quando disfônica a voz frequentemente apresenta um conjunto de características desviantes, sendo as mais comuns a rugosidade, a soproidade e a tensão (LEÃO, 2008), assim estudos que se dispõem a estudar um único tipo de voz certamente avaliará vozes com predomínio da característica vocal buscada (BEHLAU et al., 2001; DEJONCKERE, 1995; LEÃO, 2008; LOPES et al., 2012; MADAZIO, 2009; WOLFE; MARTIN, 1997), e não exclusivamente com o tipo vocal em questão. Somado a isso, acredita-se que uma única característica vocal tenha mais de uma pista acústica relacionada à sua percepção (MADAZIO, 2009).

Apesar da complexidade do sinal sonoro e seus desvios, estudos apontam que o uso de parâmetros acústicos, como o *jitter* e o *shimmer*, é promissor quando associado com sistemas automáticos de reconhecimento de padrões (WANG; JO, 2007), e também quando estes parâmetros acústicos são empregados juntos na identificação de parâmetros perceptivos (HAKKESTEEGT et al., 2008; MA; YIU, 2006; MADAZIO; LEÃO; BEHLAU, 2011; WUYTS et al., 2000; YU et al., 2001). Seguindo este raciocínio realizou-se testes com diversos subconjuntos de parâmetros para o treino/teste da RNA e, como apontado na literatura, o uso de diversos parâmetros acústicos em conjunto foi promissor.

7.3 Rede Neural Artificial

Na análise do desempenho da RNA, na avaliação da rugosidade e da soproidade, considerou-se a taxa de acerto, a confiabilidade (CCI), o erro médio e a consistência da RNA em cada subconjunto treinado e testado. Desta forma, o melhor desempenho da rede neural artificial na avaliação da rugosidade ocorreu no subconjunto composto pela Transformada *Wavelet Packet* com 1 nível de resolução, *jitter*, *shimmer*, amplitude do *pitch* (PA) e frequência fundamental (F_0), enquanto que para a soproidade o melhor desempenho ocorreu com o subconjunto composto pelo *jitter*, amplitude do *pitch* e F_0 . Observa-se que o *jitter*, PA e F_0 foram parâmetros promissores na avaliação de ambos os parâmetros perceptivos.

Para compreender os parâmetros que compuseram o subconjunto com melhor desempenho na RNA, na soproidade e na rugosidade, é necessário entender a fisiologia dessas vozes e seus marcadores perceptivos e acústicos. Destaca-se que nenhum parâmetro sozinho é capaz de identificar e avaliar a soproidade ou a rugosidade, mas cada um com sua particularidade contribui na identificação dessas vozes quando inseridos e correlacionados com outros parâmetros em um subconjunto.

7.3.1 Soproidade e sua relação com os parâmetros do melhor subconjunto de avaliação

Na soproidade ocorre a passagem de ar não-sonorizado pelas pregas vocais concomitante a produção vocal (BEHLAU et al., 2001), assim, ela é caracterizada pela sensação de ar na voz. A voz soprosa se diferencia das demais pela presença elevada de ruído, tanto nas frequências graves

como nas altas (LEÃO, 2008), sendo que diversos trabalhos a descrevem com presença de ruído apenas nas frequências agudas, acima de 4Khz (BIELAMOWICZ et al., 1996; DEJONCKERE; LEBACQ, 1996; HAMMARBERG, 2000; SANTOS et al., 2005). Além disso, a voz soprosa frequentemente está associada a vozes mais agudas (HAMMARBERG et al., 1980; WOLFE; MARTIN, 1997) e ao registro de falsete (STEINHAEUER et al., 2004), onde a produção da voz é realizada com as pregas vocais alongadas formando uma fenda paralela.

Com base nas informações apresentadas sobre a soprosidade, esperava-se que o uso da Transformada *Wavelet Packet* fosse promissor na avaliação da voz soprosa, o que não foi confirmado pelos resultados encontrados. Ao contrário, ao acrescentar-se a WPT com 1 nível de decomposição observou-se diminuição do desempenho da RNA para 71%, com aumento do erro médio para 10 mm. Além disso, observou-se que o desempenho da RNA diminuiu gradualmente com o aumento do número de níveis de decomposições da WPT, cada nível adicionado decompõe o sinal de voz, subdividindo novamente cada folha pelos filtros passa-baixa e passa-alta (GOSWANI; CHAN, 1999). Dessa forma, a energia do ruído em frequências agudas, que teoricamente discriminaria a voz soprosa, também é subdividida em novas folhas, e com folhas menores o risco dessa energia oscilar entre uma folha e outra aumenta, é possível que com isso perca-se o padrão de discriminação da vozes soprosas. Assim, a WPT com um nível de decomposição foi a que se mostrou mais promissora, no entanto não trouxe contribuições adicionais ao subconjunto de parâmetros utilizados.

Quanto as medidas de perturbação a curto prazo, apenas o *jitter* compôs o subconjunto com melhor desempenho da RNA na avaliação da soprosidade. Este resultado não era esperado, visto que acreditava-se que tanto o *jitter* como o *shimmer* compusessem o melhor subconjunto de discriminação, conforme observado nas novas abordagens propostas na literatura (HAKKESTEEGT et al., 2008; MADAZIO; LEÃO; BEHLAU, 2011; MADAZIO, 2009; NEMR et al., 2015). No entanto, estas abordagens foram propostas para avaliar o grau geral da disfonia e não um parâmetro específico, assim mudanças são esperadas. Sabe-se que o *jitter* mede as perturbações ciclo-a-ciclo do período do sinal sonoro, que teoricamente é descrito como relacionado à falta de controle na vibração das pregas vocais, mais esperado nas vozes rugosas, enquanto que o *shimmer* altera-se quando há falta de resistência glótica (BEHLAU et al., 2001), assim, mais esperado nas vozes soprosas. Apesar disso, como dito anteriormente, diversos estudos encontram resultados diferentes na correlação destes parâmetros com os perceptivo-auditivos,

como o estudo de Dejonckere (1995) e Freitas et. al (2015), que também encontram o *jitter* correlacionado com a soprosidade (DEJONCKERE, 1995). Além disso, Freitas et. al (2015) encontram o *shimmer* correlacionado apenas à rugosidade, por meio do programa MDVP, sendo que ao correlacionar a rugosidade e soprosidade com parâmetros acústicos por meio de outros programas encontram resultados diversos. Assim, os autores concluem que estas correlações isoladas não permitem o estabelecimento de relações diretas e questionam a possibilidade desse processo com os parâmetros acústicos atuais (VAZ FREITAS et al., 2015). Dessa forma, a falta do *shimmer* no melhor subconjunto para identificar a soprosidade aumenta o questionamento sobre o real correlato fisiológico deste parâmetro, correlato que não é claro, visto a divergência de resultados encontrados na literatura.

Quanto à Amplitude do *Pitch* (PA), Davis (1978) o relaciona com vogais e consoantes fricativas, sonoras e surdas, encontrando PA com valor alto nas vogais, valor diminuído nas fricativas sonoras e PA com valor zero nas fricativas surdas. O autor correlaciona os valores de PA com a periodicidade do período ciclo-a-ciclo e com o ruído (DAVIS, 1978). Assim, quanto maior a PA obtida, maior a periodicidade ciclo-a-ciclo no período e menores as interferências causadas por ruídos. Acredita-se que a razão pela PA ter sido um bom parâmetro na avaliação das vozes seja porque as vozes soprosas possuem uma presença elevada de ruídos, causando a diminuição dos valores da PA. Este resultado concorda com os achados por Pereira (PEREIRA et al., 2006), que encontraram valores estatisticamente menores em indivíduos com *Parkinson* comparados com indivíduos com vozes saudáveis. Os autores correlacionaram a diminuição da PA com a hipocinesia laríngea que ocorrem nos quadros de *Parkinson*, no entanto, não realizaram a avaliação perceptivo-auditiva das vozes.

A soprosidade, como visto, está frequentemente associada a frequências mais agudas. Exemplificando, observa-se o estudo de Santos (2012) que estudou o envelhecimento vocal. A autora encontrou, em vozes masculinas, o aumento da frequência fundamental correlacionado ao aumento da soprosidade (SANTOS, 2012), que neste caso, pode ser explicado pela possível perda de massa das pregas vocais, comum nas presbilaringes, que ocasiona fenda fusiforme e aumento da F_0 . Dessa forma observa-se que há correlação fisiológica entre estes parâmetros, e portanto, é esperado que a F_0 contribua na discriminação das vozes soprosas dentro do subconjunto de parâmetros selecionados.

7.3.2 Rugosidade e sua relação com os parâmetros de seu melhor subconjunto de avaliação

Quanto à voz rugosa, sua ocorrência indica irregularidade na vibração das pregas vocais (DEJONCKERE, 2010), o que auditivamente ocasiona a percepção de uma qualidade vocal do tipo ruidosa (BEHLAU et al., 2001). A rugosidade diferencia-se das demais vozes pela maior ocorrência de sub-harmônicos e ruídos em frequências graves (LEÃO, 2008), sendo que o ruído é mais intenso nas frequências graves pois o componente aperiódico da vibração das pregas vocais é mais intenso nesta região de frequências (KLATT; KLATT, 1990).

Em relação aos subtipos da rugosidade tem-se a rouquidão, aspereza e crepitação. A rouquidão é a alteração vocal mais comum, possui qualidade ruidosa e frequentemente possui frequência fundamental e intensidade reduzidos. Sua ocorrência está relacionada a lesões de massa nas pregas vocais, como nódulos, pólipos, cistos, edemas e neoplasias (BEHLAU et al., 2001). Por outro lado, a aspereza possui uma qualidade mais tensa e desagradável (RUSZ et al., 2015), com redução da massa vibrante e rigidez na mucosa das pregas vocais, o que gera uma frequência mais aguda. A aspereza pode ocorrer devido a leucoplasias, retrações cicatriciais pós-cirúrgicas ou sulcos vocais (BEHLAU et al., 2001). Na crepitação nota-se um ruído pulsátil, sendo descrita com uma qualidade vocal rouca com sensação adicional de impulsos repetitivos (ISHI et al., 2008), é caracterizada por uma voz com frequência fundamental grave, intensidade fraca e grande aperiodicidade na vibração das pregas vocais. A voz crepitante ocorre em pregas vocais grossas e encurtadas, podendo ocorrer devido ao uso vocal inadequado em registro basal, edemas de *Reinke* ou ocorrer fisiologicamente em vozes masculinas mais graves (BEHLAU et al., 2001).

Conhecendo a voz rugosa compreende-se a participação da Transformada *Wavelet Packet* com 1 nível de decomposição no melhor subconjunto de avaliação da rugosidade. A WPT com 1 nível de decomposição filtra a voz de acordo com a frequência em duas folhas, uma contém a entropia das frequências graves (filtro passa-baixa) e outra das frequências agudas (filtro passa-alta). Assim, todo ruído proveniente da rugosidade pode ser identificado na folha que contém as frequências graves, diferenciando a rugosidade da soproidade. Outro dado que vale a pena discutir é que com o aumento do número de níveis de decomposição da WPT o desempenho da RNA diminuiu, semelhante ao que ocorreu na avaliação da soproidade, mantendo-se a hipótese discutida para a voz soproosa de que com o aumento do número de folhas da WPT houve maior oscilação da entropia nas folhas e, com isso, perda do padrão para avaliação das vozes.

A participação do *jitter* e do *shimmer* no melhor subconjunto da rugosidade era esperado, visto que estes representam perturbações a curto prazo do sinal de voz, ou seja, perturbações ciclo-a-ciclo, que podem ocorrer em virtude do desequilíbrio fonatório. Como mencionado anteriormente, o uso destes parâmetros é controverso quando correlacionados de forma isolada com parâmetros perceptivos, no entanto, são promissores e agregam valor quando utilizados conjuntamente com outros parâmetros, demonstrado em novas propostas metodológicas como o Diagrama de desvio fonatório (MADAZIO; LEÃO; BEHLAU, 2011; MADAZIO, 2009) e o Índice de severidade da disfonia (DSI) (HAKKESTEEGT et al., 2008; NEMR et al., 2015).

A participação da amplitude do *pitch* (PA) no melhor subconjunto de avaliação da voz rugosa também era esperada, visto que ela altera-se com a falta de periodicidade do período ciclo-a-ciclo e com a presença de ruídos (DAVIS, 1978), ajudando a identificar o grau dos desvios vocais.

Em relação à contribuição da frequência fundamental na avaliação da rugosidade, sabe-se que essa é influenciada pela natureza da lesão das pregas vocais responsável pela disfonia. Como mencionado anteriormente, a rouquidão ocorre frequentemente por lesões de massa nas pregas vocais, o que leva à diminuição da F_0 , na crepitação também é comum F_0 grave, por outro lado a aspereza ocorre por rigidez do tecido da prega vocal levando à um aumento da F_0 . Além disso, encontrou-se na literatura, em uma pesquisa sobre o envelhecimento vocal, correlação do aumento da rugosidade com a diminuição da F_0 em mulheres (SANTOS, 2012). Dessa forma, justifica-se a contribuição da F_0 na identificação das vozes rugosas.

7.3.3 Desempenho da Rede Neural Artificial

O desempenho da rede neural artificial foi semelhante para a avaliação da soproidade e da rugosidade. Na avaliação da soproidade obteve-se taxa de acerto de 74%, concordância excelente com a avaliação perceptivo-auditiva da escala visual analógica (0,80 no CCI) e erro médio de 9 mm; enquanto que na avaliação da rugosidade obteve-se taxa de acerto de 73%, concordância excelente (0,84 no CCI) e erro médio de 10 mm. Além disso, observa-se que a taxa de acerto da RNA em testes aleatórios, ou seja o acerto ao acaso da RNA, foi cerca de 21%.

Quanto aos trabalhos encontrados na literatura, poucos discriminaram a qualidade vocal pelo tipo de voz e seus graus de desvio, sendo que nenhum entre eles utilizou a escala visual analógica de 100 pontos para o treino e teste dos sistemas automáticos. A escala utilizada pela

maioria foi a GRBAS de quatro pontos, na qual “0” representa uma voz sem desvio vocal, “1” desvio leve, “2” moderado e “3” intenso (MORO-VELÁZQUEZ et al., 2015; SÁENZ-LECHÓN et al., 2011; SCHÖNWEILER et al., 2000; STRÁNÍK; ČMEJLA; VOKŘÁL, 2014); enquanto que uma pesquisa utilizou uma escala com 7 pontos, na qual “0” representa as vozes com menor desvio e “7” com maior grau de desvio vocal (RITCHINGS; MCGILLION; MOORE, 2002).

A porcentagem de acerto destas pesquisas variaram consideravelmente, sendo que apenas uma obteve taxa de acerto maior que a nossa, no entanto com níveis de concordância com a avaliação perceptivo-auditiva inferior à que atingimos. Entre as pesquisas que tiveram desempenho inferior a nossa observa-se o estudo de Schönweiler et al. (2000) que encontraram 40% de acerto nas RNAs na avaliação da rugosidade e soproidade combinadas, aumentando essa taxa para 65% quando analisou estas vozes separadamente. Sáenz-Lechón et al. (2011) avaliaram as vozes rugosas e soprosas, pela escala GRBAS, obtendo taxa de acerto de 44% na rugosidade e 45% na soproidade, com melhores resultados para a classificação das vozes sem desvio vocal (grau 0), nas quais obtiveram 62% de classificação correta para vozes rugosas e soprosas.

Stráník, Cmejla e Vokrál (2014) encontraram resultados semelhantes aos nossos. Os autores avaliaram a soproidade, baseado na escala GRBAS, utilizando trechos de fala para a classificação automática das vozes. Os autores alcançaram 77% de acerto na avaliação das vozes soprosas, no entanto, com concordância de 0,59 (concordância moderada) pelo teste estatístico *Kappa*, resultado inferior ao encontrado nesta pesquisa (0,80 – CCI, concordância excelente- tabela 1). Destaca-se que o teste *Kappa* tem a mesma função estatística do CCI, com a diferença de que o Kappa é utilizado em escalas categóricas/nominais e o CCI em escalas com dados contínuos.

Moro-Velázquez et al. (2015) avaliaram o grau geral e a rugosidade, baseado na escala GRBAS, utilizando um sistema automático chamado de Modelo de Misturas Gaussianas. Os autores atingiram 81,6% de eficiência no grau geral e 84,7% na rugosidade, resultados superiores aos nossos. No entanto, apresentaram concordância inferior ao do nosso estudo, com valor de 0,73 para o grau geral e 0,76 para a rugosidade, ambos realizados pelo teste *Kappa* e considerados graus de concordância bons.

Ritchings et al. (2002) avalia o grau geral de desvio vocal pela RNA, de acordo com a escala de 7 pontos, sendo capaz de distinguir os 7 graus de desvio vocal com uma precisão de 92% de acerto segundo os autores, não sendo aplicado testes de concordância. Apesar da alta porcentagem de acerto encontradas pelos autores, lembra-se que este estudo foi baseado no grau

geral de disfonia, e não em parâmetros perceptivos específicos, não podendo assim ser diretamente comparado com os nossos resultados.

O Coeficiente de Correlação Intraclass (CCI) é utilizado atualmente na maioria dos trabalhos da área de voz para avaliar a concordância, intra e interjuízes, de dados contínuos em avaliações perceptivo-auditivas (CRNAITO et al., 2012; EADIE et al., 2010; KREIMAN; GERRATT; ITO, 2007; MARYN; KIM; KIM, 2015; NEMR et al., 2012; YAMASAKI et al., 2008). Neste trabalho empregou-se o CCI tanto na concordância intra e interjuízes como para avaliar a concordância dos resultados apresentados pela RNA com os valores utilizados para seu treino, que é a mediana dos quatro fonoaudiólogos na avaliações perceptivo-auditivas de cada voz. Como mencionado, na soproidade a concordância da RNA foi de 0,80 enquanto que a concordância interjuízes, dos fonoaudiólogos, foi de 0,84; para a rugosidade a concordância foi de 0,84 entre a RNA e os dados de treino; e de 0,85 entre os fonoaudiólogos. Comparando os resultados da rede neural artificial com a avaliação dos fonoaudiólogos observou-se que a rede se comportou como um quinto juiz da pesquisa, visto que seu desempenho foi semelhante ao dos juízes na avaliação perceptivo-auditiva, tanto na soproidade (figura 20) como na rugosidade (figura 21).

Quanto ao erro médio da RNA, sabe-se que na literatura considera-se empiricamente o valor de 10 mm como a mínima diferença clinicamente significativa (EADIE et al., 2010), assim em diferenças menores que 10 mm considera-se que há concordância entre os avaliadores. Dessa forma, o erro médio encontrado tanto na soproidade (9 mm) como na rugosidade (10 mm) foi considerado bom e dentro dos limites de variação esperados. Destaca-se que o erro médio foi calculado a partir da média do erro da RNA, em mm, de todas as amostras vocais utilizadas no estudo, assim contendo desde vozes quase neutras a vozes com grau intenso de rugosidade/soproidade, portanto, não sendo possível aplicar a equação quadrática encontrada, que possui um resultado diferente em cada ponto da RNA.

Acredita-se que para o aperfeiçoamento da avaliação vocal por meio da RNA é importante dar continuidade a este trabalho, testando novos parâmetros, sejam eles: acústicos tradicionais, provenientes de avaliações não-lineares, aerodinâmicos, entre outros; capazes de contribuir com a discriminação e graduação da soproidade e da rugosidade vocais. Além disso, acredita-se que aumentar o número de amostras de cada subtipo vocal também possa melhorar o desempenho da RNA.

Observando-se conjuntamente todos os resultados tem-se que a RNA foi passível de treino tanto quanto o ouvido humano, apresentando um desempenho equivalente aos fonoaudiólogos juízes desta pesquisa. Dessa forma, esta pesquisa indica novos caminhos na avaliação vocal, possibilitando uma avaliação tão eficiente quanto a perceptivo-auditiva, mas com a vantagem da objetividade e reprodutibilidade dos resultados. Além disso, destaca-se a contribuição do estudo da curva de percepção-auditiva, a qual revelou a natureza não-linear da percepção dos desvios vocais, que possibilitou o desenvolvimento da equação da diferença permitida, a qual traz consigo uma nova forma de computar a concordância intra e interjuízes, superando as limitações dos métodos aplicados atualmente.

8. CONCLUSÃO

O uso das redes neurais artificiais, na identificação e avaliação do grau da rugosidade e da soproidade, apresentou confiabilidade excelente ($CCI > 0,80$), revelando uma relação significativa com a avaliação perceptivo-auditiva. Assim, o uso das redes neurais artificiais na avaliação vocal é uma metodologia viável e promissora, tendo sua maior vantagem a objetividade na avaliação e reprodutibilidade de resultados.

Quanto ao melhores parâmetros acústicos e da Transformada *Wavelet Packet* na parametrização da rede neural artificial encontrou-se na avaliação da soproidade um subconjunto composto pelos parâmetros: *jitter*, amplitude do *pitch* e frequência fundamental, com desempenho de 74% de acerto, confiabilidade excelente (0,80 – CCI) e erro médio de 9 mm; e na avaliação da rugosidade o melhor subconjunto foi: Transformada *Wavelet Packet* com 1 nível de decomposição, *jitter*, *shimmer*, amplitude do *pitch* e frequência fundamental, com desempenho de 73% de acerto, confiabilidade excelente (0,84 – CCI) e erro médio de 10 mm.

Em relação à percepção do grau da rugosidade e soproidade encontrou-se uma curva quadrática, o que significa que é necessário duplicar o desvio vocal para que ele seja percebido de forma linear. A percepção não-linear do desvio vocal encontrada é condizente com o comportamento do ouvido humano, que também percebe a variação de outras grandezas como a intensidade e a frequência do sinal sonoro não-linearmente. Assim, a mínima diferença clinicamente significativa não é representada por um valor único, ela é diferente de acordo com o grau da soproidade/rugosidade do sinal de voz, variando de acordo com a equação: $Dif_p = \sqrt{5y}$, onde y refere-se a graduação, entre 0 e 100 mm, do grau do parâmetro perceptivo avaliado e Dif_p à diferença permitida, ou seja, a mínima diferença clinicamente significativa.

REFERÊNCIAS¹

- ARJMANDI, M. K.; POOYAN, M. An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. **Biomedical Signal Processing and Control**, v. 7, n. 1, p. 3–19, jan. 2012.
- AWAN, S. N.; ROY, N. Acoustic prediction of voice type in women with functional dysphonia. **Journal of Voice**, v. 19, n. 2, p. 268–282, 2005.
- BEHLAU, M. et al. Avaliação de voz. In: BEHLAU, M. (Ed.). . **Voz: o livro do especialista**. Rio de Janeiro: Revinter, 2001. p. 85–245.
- BEHLAU, M. Consensus Auditory- Perceptual Evaluation of Voice (CAPE-V), ASHA 2003. Refletindo sobre o novo/New reflexions. **Rev SBFa**, v. 9, n. 3, p. 187–9, 2004.
- BELE, I. V. Reliability in perceptual analysis of voice quality. **Journal of Voice**, v. 19, n. 4, p. 555–573, 2005.
- BHUTA, T.; PATRICK, L.; GARNETT, J. D. Perceptual evaluation of voice quality and its correlation with acoustic measurements. **Journal of Voice**, v. 18, n. 3, p. 299–304, set. 2004.
- BIELAMOWICZ, S. et al. Comparison of Voice Analysis Systems for Perturbation Measurement. **Journal of Speech Language and Hearing Research**, v. 39, n. 1, p. 126, 1 fev. 1996.
- BONALDI, L. V. et al. Origem da audição: um capítulo na evolução do equilíbrio. In: **Bases anatômicas da audição e do equilíbrio**. São Paulo: Livraria Santos Editora Ltda, 2004. p. 3–6.
- BROCKMANN, M. et al. Reliable jitter and shimmer measurements in voice clinics: the relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. **Journal of voice : official journal of the Voice Foundation**, v. 25, n. 1, p. 44–53, jan. 2011.
- CALLEGARI-JACQUES, S. M. Correlação linear simples. In: **Bioestatística: Princípios e aplicações**. [s.l.] Artmed, 2003. p. 84–95.
- CANNITO, M. P. et al. Perceptual structure of adductor spasmodic dysphonia and its acoustic correlates. **Journal of Voice**, v. 26, n. 6, p. 818.e5–818.e13, 2012.
- CARVALHO, R. T. S.; CAVALCANTE, C. C.; CORTEZ, P. C. Wavelet transform and artificial neural networks applied to voice disorders identification. **2011 Third World Congress on Nature and Biologically Inspired Computing**, p. 371–376, 2011.
- COIFMAN, R. R.; WICKERHAUSER, M. V. Entropy Based Algorithms for Best Basis Selection. **IEEE Transactions on Information Theory**, v. 32, n. 2, p. 712–718, 1992.
- CORDEIRO, H. T.; FONSECA, J. M.; RIBEIRO, C. M. LPC Spectrum First Peak Analysis for Voice Pathology Detection. **Procedia Technology**, v. 9, p. 1104–1111, 2013.

¹De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023

CROVATO, C. D. P.; SCHUCK, A. The Use of Wavelet Packet Transform and Artificial Neural Networks in Analysis and Classification of Dysphonic Voices. **IEEE Transactions on Biomedical Engineering**, v. 54, n. 10, p. 2006–2008, 2007.

DAJER, M. E. et al. Vocal Dynamic Visual Pattern for voice characterization. **Journal of Physics: Conference Series**, v. 332, p. 012026, 2011.

DAVIS, S. B. **Acoustic characteristics of normal and pathological voices** *Speech and Language: Research and Theory*, 1978.

DE BRUIJN, M. et al. Artificial neural network analysis to assess hypernasality in patients treated for oral or oropharyngeal cancer. **Logopedics, phoniatrics, vocology**, v. 36, n. 4, p. 168–74, dez. 2011.

DEJONCKERE, P. Principal components in voice pathology. **J Voice**, v. 4, p. 96–105, 1995.

DEJONCKERE, P. Assessment of Voice and Respiratory Function. In: REMACLE, M.; ECKEL, H. E. (Eds.). **Surgery of Larynx and Trachea SE - 2**. [s.l.] Springer Berlin Heidelberg, 2010. p. 11–26.

DEJONCKERE, P. H.; LEBACQ, J. Acoustic, perceptual, aerodynamic and anatomical correlations in voice pathology. **ORL; journal for oto-rhino-laryngology and its related specialties**, v. 58, n. 6, p. 326–32, jan. 1996.

DEMUTH, H.; BEALE, M.; HAGAN, M. **Neural Network Toolbox 5 - User's Guide**. Disponível em: <<http://www.mathworks.com/access/helpdesk/help/toolbox/nnet/>>. Acesso em: 20 mar. 2012.

DUDA, R. O.; HART, P. E.; STORK, D. G. Multilayer Neural Networks. In: DUDA, R. O.; HART, P. E.; STORK, D. G. (Eds.). **Pattern Classification**. 2. ed. New York: wiley, 2001a. p. 282–347.

DUDA, R. O.; HART, P. E.; STORK, D. G. Introduction. In: DUDA, R. O.; HART, P. E.; STORK, D. G. (Eds.). **Pattern Classification**. 2. ed. New York: Wiley, 2001b. p. 1–19.

DUDA, R. O.; HART, P. E.; STORK, D. G. Maximum likelihood and Bayesian estimation. In: DUDA, R. O.; HART, P. E.; STORK, D. G. (Eds.). **Pattern Classification**. New York: Wiley, 2001c. p. 84–159.

EADIE, T. L. et al. The role of experience on judgments of dysphonia. **Journal of Voice**, v. 24, n. 5, p. 564–573, 2010.

ENGLERT, M. et al. Perceptual Error Identification of Human and Synthesized Voices. **Journal of voice : official journal of the Voice Foundation**, 31 ago. 2015.

FERNANDES, J. C. **Acústica, ruídos e perda de audição** *Proceedings of the 9th Brazilian Conference on Dynamics Control and their Applications. Anais...Serra Negra: DINCON'10, 2010*

FINK, D. S. et al. Subjective and objective voice outcomes after transoral laser microsurgery for early glottic cancer. **The Laryngoscope**, 24 nov. 2015.

FROHLICH, M.; MICHAELIS, D.; WERNER STRUBE, H. Acoustic “breathiness measures” in the description of pathologic voices. **ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings**, v. 2, p. 937–940, 1998.

GODINO-LLORENTE, J. I.; GOMEZ-VILDA, P. **Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors** *Biomedical Engineering, IEEE*

Transactions on, 2004.

GOSWANI, J. C.; CHAN, A. K. Digital Signal Processing Applications. In: **Fundamentals of Wavelets: theory, algorithms, and Applications**. Estados Unidos: Wiley, 1999. p. 226–8.

HAKKESTEEGT, M. M. et al. The Relationship Between Perceptual Evaluation and Objective Multiparametric Evaluation of Dysphonia Severity. **Journal of Voice**, v. 22, n. 2, p. 138–145, 2008.

HAMMARBERG, B. et al. Perceptual and acoustic correlates of abnormal voice qualities. **Acta otolaryngologica**, v. 90, n. 5-6, p. 441–51, jan. 1980.

HAMMARBERG, B. Voice research and clinical needs. **Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)**, v. 52, n. 1-3, p. 93–102, jan. 2000.

HARIHARAN, M.; YAACOB, S.; AWANG, S. A. Pathological infant cry analysis using wavelet packet transform and probabilistic neural network. **Expert Systems with Applications**, v. 38, n. 12, p. 15377–15382, 2011.

HAYKIN, S. S. Perceptrons de múltiplas camadas. In: **Redes Neurais - Princípios e práticas**. 2. ed. Porto Alegre: Bookman, 2001. p. 183–282.

HILLMAN, R. E.; MONTGOMERY, W. W.; ZEITELS, S. M. Current diagnostics and office practice: Appropriate use of objective measures of vocal function in the multidisciplinary management of voice disorders. **Current Opinion in Otolaryngology & Head and Neck Surgery**, v. 5, n. 3, 1997.

HIRANO, M. Psycho-Acoustic Evaluation of Voice. In: ARNOLD, G.; WINCKE, L F.; WYKE, B. (Eds.). **. Clinical Examination of Voice**. Austria: Springer-Verlag/Wien: [s.n.]. p. 81–4.

ISHI, C. T. et al. A Method for Automatic Detection of Vocal Fry. **IEEE Transactions on Audio, Speech, and Language Processing**, v. 16, n. 1, p. 47–56, jan. 2008.

KARNELL, M. P. et al. Reliability of Clinician-Based (GRBAS and CAPE-V) and Patient-Based (V-RQOL and IPVI) Documentation of Voice Disorders. **Journal of Voice**, v. 21, n. 5, p. 576–590, 2007.

KLATT, D. H.; KLATT, L. C. Analysis, synthesis, and perception of voice quality variations among female and male talkers. **The Journal of the Acoustical Society of America**, v. 87, n. 2, p. 820–57, fev. 1990.

KOVÁCS, Z. L. Redes Neurais em Processamento de Sinais. In: **Redes Neurais Artificiais: Fundamentos e Aplicações**. [s.l.] Copyright, 1996. p. 141–9.

KREIMAN, J. et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. **Journal of speech and hearing research**, v. 36, n. 1, p. 21–40, 1993.

KREIMAN, J.; GERRATT, B. R.; ITO, M. When and why listeners disagree in voice quality assessment tasks. **The Journal of the Acoustical Society of America**, v. 122, n. 4, p. 2354–2364, 2007.

LEÃO, S. H. D S. **Análise espectrográfica acústica de vozes rugosas, soprosas e tensas**. [s.l.] Universidade Federal de São Paulo, 2008.

LINDER, R. et al. Artificial Neural Network-based Classification to Screen for Dysphonia Using

Psychoacoustic Scaling of Acoustic Voice Features. **Journal of Voice**, v. 22, n. 2, p. 155–163, 2008.

LOPES, L. W. et al. Severity of voice disorders in children: Correlations between perceptual and acoustic data. **Journal of Voice**, v. 26, n. 6, 2012.

LUQUETTI, L. B.; LAGUARDIA, J. Confiabilidade dos dados de atendimento odontológico do Sistema de Gerenciamento de Unidade Ambulatorial Básica (Sigab) em Unidade Básica de Saúde do Município do Rio de Janeiro. **Epidemiologia e Serviços de Saúde**, v. 18, n. 3, p. 255–264, set. 2009.

MA, E. P. M.; YIU, E. M. L. Multiparametric Evaluation of Dysphonic Severity. **Journal of Voice**, v. 20, n. 3, p. 380–390, 2006.

MACKAY, D. J. C. **Information Theory , Inference, and Learning Algorithms**. Cambridge: Cambridge University Press, 2005.

MADAZIO, G. Diagrama de desvio fonatório na clínica vocal. 2009.

MADAZIO, G.; LEÃO, S.; BEHLAU, M. The phonatory deviation diagram: A novel objective measurement of vocal function. **Folia Phoniatria et Logopaedica**, v. 63, n. 6, p. 305–311, 2011.

MARTINEZ, C. E.; RUFINER, H. L. Acoustic analysis of speech for detection of laryngeal pathologies. **Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (Cat. No.00CH37143)**, v. 3, p. 2369–2372, 2000.

MARYN, Y. et al. Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels. **Journal of voice : official journal of the Voice Foundation**, v. 24, n. 5, p. 540–55, set. 2010.

MARYN, Y.; KIM, H.-T.; KIM, J. Auditory-Perceptual and Acoustic Methods in Measuring Dysphonia Severity of Korean Speech. **Journal of voice : official journal of the Voice Foundation**, 24 ago. 2015.

MARYN, Y.; ROY, N. Sustained vowels and continuous speech in the auditory-perceptual evaluation of dysphonia severity. **Jornal da Sociedade Brasileira de Fonoaudiologia**, v. 24, n. 2, p. 107–112, 2012.

MARYN, Y.; WEENINK, D. Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. **Journal of voice : official journal of the Voice Foundation**, v. 29, n. 1, p. 35–43, jan. 2015.

MONTAGNOLI, A. N. **Análise Residual do sinal de voz**. [s.l.] Universidade de São Paulo, 1998.

MOORE, B. C. J. **An Introduction to the Psychology of Hearing**. [s.l.] BRILL, 2012.

MORO-VELÁZQUEZ, L. et al. Modulation Spectra Morphological Parameters: A New Method to Assess Voice Pathologies according to the GRBAS Scale. **BioMed research international**, v. 2015, p. 259239, jan. 2015.

NEMR, K. et al. GRBAS and cape-V scales: High reliability and consensus when applied at different times. **Journal of Voice**, v. 26, n. 6, p. 17–22, 2012.

NEMR, K. et al. Correlation of the Dysphonia Severity Index (DSI), Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V), and Gender in Brazilians With and Without Voice Disorders. **Journal of voice : official journal of the Voice Foundation**, 25 nov. 2015.

OATES, J. Auditory-perceptual evaluation of disordered voice quality: **Folia Phoniatica et Logopaedica**, v. 61, n. 1, p. 49–56, 2009.

PEREIRA, J. C. et al. Residual signal auto-correlation to evaluate speech in Parkinson's disease patients. **Arquivos de Neuro-Psiquiatria**, v. 64, n. 4, p. 912–915, dez. 2006.

RABINOV, C. R. et al. Comparing reliability of perceptual ratings of roughness and acoustic measure of jitter. **Journal Of Speech And Hearing Research**, v. 38, n. 1, p. 26–32, fev. 1995.

RITCHINGS, R. T.; MCGILLION, M.; MOORE, C. J. Pathological voice quality assessment using artificial neural networks. **Medical engineering & physics**, v. 24, n. 7-8, p. 561–564, 2002.

RON KOHAVI. **A study of cross-validation and bootstrap for accuracy estimation and model selection** International joint Conference on artificial intelligence. **Anais...**Stanford, CA: 1995

ROSNER, B. **Fundamentals of Biostatistics**. Belmont, CA: Duxbury Press, 1995.

ROY, N. et al. Evidence Based Clinical Voice Assessment: A Systematic Review. **American Journal of Speech-Language Pathology**, v. 22, n. 2, p. 212, 1 maio 2013.

RUSZ, J. et al. Speech disorders reflect differing pathophysiology in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. **Journal of Neurology**, v. 262, n. 4, p. 992–1001, 17 fev. 2015.

SÁENZ-LECHÓN, N. et al. Towards objective evaluation of perceived roughness and breathiness: an approach based on mel-frequency cepstral analysis. **Logopedics, phoniatrics, vocology**, v. 36, n. 2, p. 52–59, 2011.

SALHI, L. Voice Disorders Identification Using multilayer neural network. **The International Arab Journal of Information Technology**, v. 7, n. 2, p. 177–85, 2010.

SALHI, L.; MOURAD, T.; CHERIF, A. Voice disorders classification using multilayer neural network. **2008 2nd International Conference on Signals, Circuits and Systems**, v. 7, n. 2, p. 177–185, 2008.

SANTOS, A. O. **Parâmetros acústicos e perceptivo-auditivos da voz de adultos e idosos**. [s.l.] Universidade de São Paulo, 2012.

SANTOS, I. et al. Parâmetros acústicos da voz relacionados à soproidade. **Rev Soc Bras Fonoaudiol.**, v. 10, n. 1, p. 53–9, 2005.

SCHÖNWEILER, R. et al. Novel approach to acoustical voice analysis using artificial neural networks. **JARO-Journal of the Association for Research in Otolaryngology-**, v. 1, n. 4, p. 270–282, 2000.

SHRIVASTAV, R. The Use of an Auditory Model in Predicting Perceptual Ratings of Breathy Voice Quality. **Journal of Voice**, v. 17, n. 4, p. 502–512, 2003.

STEINHAEUER, K. et al. The relationship among voice onset, voice quality, and fundamental frequency: a dynamical perspective. **Journal of voice : official journal of the Voice Foundation**, v. 18, n. 4, p. 432–42, dez. 2004.

STRÁNÍK, A.; ČMEJLA, R.; VOKŘÁL, J. Acoustic parameters for classification of breathiness in

continuous speech according to the GRBAS scale. **Journal of voice : official journal of the Voice Foundation**, v. 28, n. 5, p. 653.e9–653.e17, 9 set. 2014.

TEIXEIRA, L. C.; BEHLAU, M. Comparison Between Vocal Function Exercises and Voice Amplification. **Journal of voice : official journal of the Voice Foundation**, v. 29, n. 6, p. 718–726, 19 ago. 2015.

TITZE, I. R. **Workshop on Acoustic Voice Analysis**IowaNational Center for Voice and Speech, , 1995.

TITZE, I. R.; BAKEN, R. J.; HERZEL, H. Evidence of chaos in vocal fold vibration. In: TITZE, I. (Ed.). . **Vocal Fold Physiology: New Frontiers in Basic Science**. San Diego, CA: Singular Publishing Group, 1993. p. 143–188.

ULOZA, V. et al. Categorizing normal and pathological voices: Automated and perceptual categorization. **Journal of Voice**, v. 25, n. 6, p. 700–708, 2011.

VAZ FREITAS, S. et al. Integrating voice evaluation: correlation between acoustic and audio-perceptual measures. **Journal of voice : official journal of the Voice Foundation**, v. 29, n. 3, p. 390.e1–7, maio 2015.

WALKER, J. **Fundamentals of Physics**. 8. ed. Danvers: Wiley, 2008.

WANG, J.; JO, C. Vocal folds disorder detection using pattern recognition methods. **Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference**, v. 2007, p. 3253–6, jan. 2007.

WOLFE, V.; MARTIN, D. Acoustic correlates of dysphonia: Type and severity. **Journal of Communication Disorders**, v. 30, n. 5, p. 403–416, 1997.

WUYTS, F. L. et al. The Dysphonia Severity Index: an objective measure of vocal quality based on a multiparameter approach. **Journal of Speech, Language & Hearing Research**, v. 43, n. 3, p. 796–809, jun. 2000.

YAMASAKI, R. et al. **Correspondência entre Escala Analógico-Visual e a Escala Numérica na Avaliação Perceptivo-Auditiva de Vozes**16º Congresso Brasileiro de Fonoaudiologia. **Anais...**Campos do Jordão: Sociedade Brasileira de Fonoaudiologia, 2008

YU, P. et al. Objective voice analysis for dysphonic patients: A multiparametric protocol including acoustic and aerodynamic measurements. **Journal of Voice**, v. 15, n. 4, p. 529–542, 2001.

ZRAICK, R. I.; WENDEL, K.; SMITH-OLINDE, L. The effect of speaking task on perceptual judgment of the severity of dysphonic voice. **Journal of Voice**, v. 19, n. 4, p. 574–581, 2005.

APÊNDICES

APÊNDICE A

ESCALA VISUAL-ANALÓGICA: AVALIAÇÃO PERCEPTIVO-AUDITIVA

Voz nº: _____ **Avaliador (a):** _____ **Data:** _____

- 1) Os parâmetros da qualidade vocal deverão ser preenchidos conforme a avaliação da vogal sustentada com duração de 3 segundos.
- 2) A rugosidade deverá ser avaliada quanto ao grau e ao tipo, sendo discriminada entre voz rouca, áspera e/ou crepitante (campo entre parênteses).
- 3) Os campos em branco deverão ser preenchidos por características de desvio vocal adicionais apenas quando estas forem predominantes em relação à rugosidade ou soproidade.

		SCORE
RUGOSIDADE	_____	____/100
(_____)		
SOPROSIDADE	_____	____/100
_____	_____	____/100
_____	_____	____/100

CARACTERÍSTICA VOCAL PREDOMINANTE: _____

Comentários sobre ressonância: NORMAL OUTRA (descreva): _____

Características adicionais (por exemplo: diplofonia, som basal, falsete, astenia, afonia, instabilidade de frequência, tremor, qualidade molhada ou outras observações relevantes).

APÊNDICE B

ESCALA NUMÉRICA AVALIAÇÃO PERCEPTIVO-AUDITIVA

Voz nº: _____ Avaliador (a): _____ Data: _____

Instruções: Avalie o grau de presença da rugosidade e soprosidade em quatro pontos, sendo: 0 = ausência dos parâmetros avaliados, 1 = presença leve dos parâmetros avaliados, 2 = presença moderada e 3 = presença intensa. Além disso, abaixo do campo RUGOSIDADE aponte qual parâmetro é predominante entre rouquidão, aspereza ou crepitação. O campo em branco deve ser preenchido quando houver outra característica vocal predominante na amostra de voz avaliada.

Ao finalizar a avaliação indique qual parâmetro é predominante na voz avaliada (rugosidade, soprosidade ou outro) e descreva a ressonância.

RUGOSIDADE	0	1	2	3

(_____)

SOPROSIDADE	0	1	2	3

_____	0	1	2	3

CARACTERÍSTICA VOCAL PREDOMINANTE: _____

Ressonância: NORMAL OUTRA (descreva): _____

ANEXOS

ANEXO A

FACULDADE DE
ODONTOLOGIA DE BAURU-
USP

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: Validação de um sistema automático para a avaliação da qualidade da voz

Pesquisador: Paula Belini Baravieira

Área Temática:

Versão: 1

CAAE: 09635512.0.0000.5417

Instituição Proponente: Faculdade de Odontologia de Bauru-USP

DADOS DO PARECER

Número do Parecer: 161.349

Data da Relatoria: 28/11/2012

Apresentação do Projeto:

Este estudo será desenvolvido com gravações vocais contidas no banco de dados do Laboratório de Voz da Clínica de Fonoaudiologia da Faculdade de Odontologia de Bauru - Universidade de São Paulo (FOB/USP) as quais fazem parte dos procedimentos realizados rotineiramente com pacientes atendidos na Clínica de Voz I. Serão selecionadas 100 vozes, de indivíduos de ambos os sexos, com idade superior a 18 anos. As amostras vocais selecionadas serão subdivididas em três grupos de acordo com o grau e tipo de alteração vocal, sendo: Grupo sem desvio vocal: vozes com qualidade dentro da variabilidade esperada para a normalidade; Grupo com vozes predominantemente rugosas: vozes disfônicas com rugosidade como característica predominante da alteração vocal, englobando disfonias de grau leve a severo. Ressalta-se que se consideram vozes rugosas aquelas que apresentam impressão de irregularidade na fonte sonora (16). Grupo com vozes predominantemente soprosas: vozes disfônicas possuindo a soprosidade como característica predominante na alteração vocal, englobando disfonias de grau leve a severo. Define-se vozes soprosas aquelas que apresentam escape de ar audível na voz (16). A seleção dos indivíduos será por meio da análise dos dados contidos em prontuário, de acordo com os critérios de inclusão e exclusão da amostra, e por meio da avaliação perceptivo-auditiva da voz pela escala CAPE-V (Consensus Auditory-Perceptual Evaluation of Voice) (17).

Objetivo da Pesquisa:

Validar um sistema automático para objetivar a avaliação perceptivo-auditiva com relação à rugosidade e a soprosidade da voz

Endereço: DOUTOR OCTAVIO PINHEIRO BRISOLLA 75 QUADRA 9

Bairro: VILA NOVA CIDADE UNIVERSITARIA **CEP:** 17.012-901

UF: SP **Município:** BAURU

Telefone: (14)3235-8356

Fax: (14)3235-8356

E-mail: mferrari@fob.usp.br

FACULDADE DE
ODONTOLOGIA DE BAURU-
USP

Avaliação dos Riscos e Benefícios:

Não há.

Comentários e Considerações sobre a Pesquisa:

Trata-se de pesquisa que utilizará dados do banco de voz do Laboratório de Voz da Clínica de Fonoaudiologia da FOB/USP. Para tanto, 100 vozes de indivíduos adultos, de ambos os sexos serão analisadas por 3 avaliadores calibrados e divididos em 3 grupos: Grupo sem desvio vocal: vozes com qualidade dentro da variabilidade esperada para a normalidade; Grupo com vozes predominantemente rugosas: vozes disfônicas com rugosidade como característica predominante da alteração vocal, englobando disfonias de grau leve a severo e Grupo com vozes predominantemente soprosas: vozes disfônicas possuindo a soprosidade como característica predominante na alteração vocal, englobando disfonias de grau leve a severo.

Considerações sobre os Termos de apresentação obrigatória:

Toda a documentação exigida foi devidamente apresentada, incluindo os termos de permissão de consulta ao Banco de Voz da FOB-USP.

Recomendações:

Não há.

Conclusões ou Pendências e Lista de Inadequações:

Não se aplica

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

Esse projeto foi considerado APROVADO. O CEP-FOB/USP exige a apresentação de relatórios anuais (parciais e finais), conforme o cronograma apresentado. Qualquer alteração na metodologia e/ou título e a inclusão ou exclusão de autores deverá ser prontamente comunicada. Lembramos que na apresentação do relatório final, deverão ser incluídos todos os TCLEs e/ou termos de doação de dentes devidamente assinados e rubricados.

Endereço: DOUTOR OCTAVIO PINHEIRO BRISOLLA 75 QUADRA 9

Bairro: VILA NOVA CIDADE UNIVERSITARIA CEP: 17.012-901

UF: SP Município: BAURU

Telefone: (14)3235-8356

Fax: (14)3235-8356

E-mail: mferrari@fob.usp.br

FACULDADE DE
ODONTOLOGIA DE BAURU-
USP

BAURU, 03 de Dezembro de 2012

Assinador por:
Maria Teresa Atta
(Coordenador)

Endereço: DOUTOR OCTAVIO PINHEIRO BRISOLLA 75 QUADRA 9
Bairro: VILA NOVA CIDADE UNIVERSITARIA   **CEP:** 17.012-901
UF: SP   **Município:** BAURU
Telefone: (14)3235-8356   **Fax:** (14)3235-8356   **E-mail:** mferrari@fob.usp.br