https://eventos.utfpr.edu.br//sicite/sicite2020

Reconhecimento de edemas de Reinke por meio de redes neurais artificiais do tipo perceptron multicamadas

Reinke's edema voice recognition through multilayer perceptron artificial neural networks

Rogerio Pinhelli
rogerio.pinhelli@gmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Maria Eugenia Dajer
medajer@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil.

Danilo Hernane Spatti
spatti@icmc.usp.br
Universidade de São Paulo, São Carlos, SP, Brasil

RESUMO

A voz é a principal ferramenta de comunicação social. Os problemas de voz, especialmente em profissionais que a utilizam como ferramenta de trabalho, trazem prejuízos sociais e econômicos. Existem muitos desafios no diagnóstico de distúrbios vocais, principalmente devido às consequências negativas de métodos invasivos de exame, e os métodos de diagnóstico disponíveis amplamente possuem alto custo. Nesse sentido, este trabalho procura aplicar redes neurais artificiais do tipo *perceptron* multicamadas para reconhecer vozes saudáveis e vozes de pacientes com Edema de Reinke. Para isso, será feito um pré-processamento do sinal, removendo artefatos, e, posteriormente será utilizada a transformada *Wavelet Packet* para a extração de características vocais nos sinais. As famílias coif3 e db3 apresentaram a melhor taxa de acerto, 97,12%.

PALAVRAS-CHAVE: Edema de Reinke. Perceptron multicamadas. Transformada Wavelet Packet. Voz.

ABSTRACT

The voice is the main tool of social communication. Voice problems, especially in professionals who use it as a work tool, bring social and economic losses. There are many challenges in the diagnosis of vocal disorders, mainly due to the negative consequences of invasive examination methods, and the available diagnostic methods are expensive. In this sense, this work seeks to apply Artificial Neural Networks of the multilayer perceptron type to recognize healthy voices and the voices of patients with Reinke's Edema. For this, a pre-processing of the signal will be done, removing artifacts, and, later, the Wavelet Packet transform will be used to extract vocal characteristics in the signals. The coif3 and db3 families had the best hit rate, 97.12%.

KEYWORDS: Reinke's edema. Multilayer perceptron. Wavelet Packet transform. Voice.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

As alterações no padrão de vibração das pregas vocais na laringe são patologias chamadas disfonias, que podem causar problemas àqueles que a utilizam profissionalmente. De acordo com Behlau (2001), as disfonias podem ser caracterizadas em orgânicas, quando independem do uso da voz, e organofuncionais, quando se relacionam ao funcionamento da voz com lesões, e surgem da evolução de disfonias funcionais. Dentre as lesões organofuncionais mais comuns estão nódulos, pólipos e edemas de Reinke, que costumam ser diagnosticados por meio de exame laringoscópico (BARIZÃO, 2017).

Existem duas classificações para os procedimentos de avaliação vocal, que são os invasivos e os não invasivos. Os invasivos são feitos com ferramentas endoscópicas para analisar a estrutura e o padrão vibratório das pregas vocais e são os que causam maior desconforto ao paciente, além de ter custo elevado (GONZÁLES, 2008). Por outro lado, os procedimentos não invasivos podem fornecer informações quantitativas, que podem ser processadas e analisadas computacionalmente e criticamente por fonoaudiólogos e profissionais da área, servindo como ferramenta auxiliar no diagnóstico (TSUJI et al, 2014), (FERMINO, 2017).

Tendo isso em vista, esse trabalho tem por objetivo reconhecer edemas de Reinke por meio de redes neurais artificiais do tipo perceptron multicamadas, usando características extraídas com transformadas *wavelet packet* como parâmetro de entrada, visando desenvolver uma ferramenta para classificação de vozes.

METODOLOGIA

Para a realização do trabalho, foi necessário aplicar algumas técnicas de pré-processamento de sinais nos áudios de uma base de dados e separá-los em treinamento e teste, para depois realizar a classificação.

BASE DE DADOS

Nesse estudo, foram utilizados 36 arquivos de áudio que contém a vogal /a/ sustentada, sendo esses arquivos divididos em duas categorias: 20 áudios de vozes saudáveis e 16 áudios de vozes de pacientes com edema de Reinke. O material foi cedido pelo Grupo de Engenharia Médica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (GPEM/CNPq). As gravações foram feitas no Ambulatório de Voz do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo (HC-FMUSP).

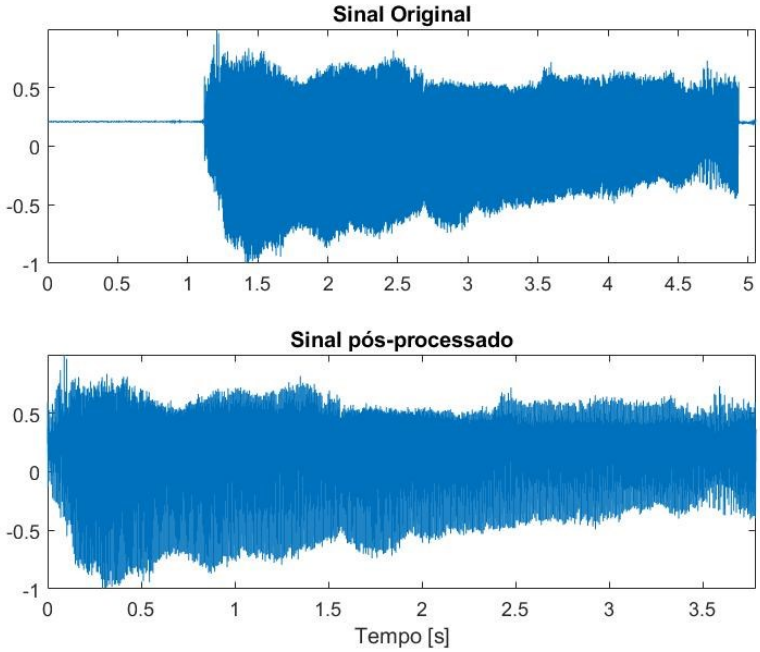
Os registros foram realizados por meio de um microfone *headset* unidirecional e a gravação e armazenamento foram feitos por meio do software *Audacity* em computador com Pentium II acoplado a uma placa de som.

PRÉ-PROCESSAMENTO DOS ÁUDIOS

Os arquivos de áudio das amostras tinham problemas de uniformidade na taxa de amostragem, ou seja, alguns áudios tinham taxa de amostragem de 44100 Hz e outros tinham taxa de amostragem de 22050 Hz. Para uniformizar a taxa de amostragem de todos os áudios, existem duas técnicas possíveis: *upsampling*, que consiste em gerar amostras artificiais por meio de interpolação para aumentar a taxa de amostragem, e o *downsampling*, que elimina algumas amostras. A técnica escolhida foi o *downsampling*, porque o processo de *upsampling* gera ruídos, e como os áudios de voz são bastante sensíveis, o procedimento de *downsampling* se torna mais viável (LIMA, 2018).

Outro procedimento importante foi ajustar o possível *DC offset* que alguns arquivos apresentavam, ou seja, as amostras estavam deslocadas do eixo original. Para solucionar esse problema, foi utilizada a função *detrend* do software *Matlab*, disponibilizado pela universidade na versão estudante. Por fim, foi necessário eliminar os trechos de áudio com silêncio, para que não sejam geradas janelas de amostra com sinais de silêncio, que podem prejudicar o treinamento da rede neural, e, para isso, foi aplicado um filtro de remoção de silêncio (LIMA, 2018). A etapa final do pré-processamento consistiu em analisar os áudios manualmente e remover artefatos que poderiam prejudicar o estudo, como inspiração de ar, vozes de outras pessoas, tosses etc. Na Figura 1, é possível observar o sinal original e o sinal após o pré-processamento.

Figura 1 – Sinal original e sinal após o pré-processamento.

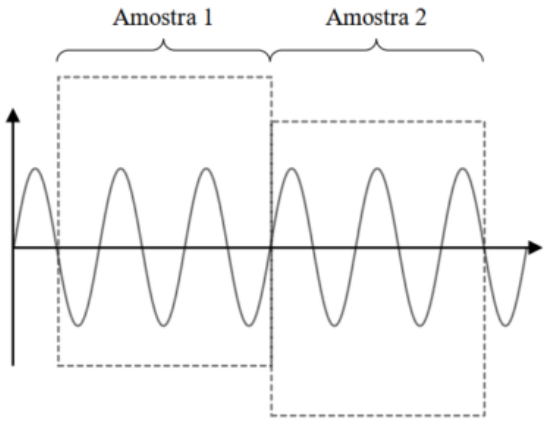


Fonte: Autoria própria (2020).

Após esse pré-processamento, foi realizado o janelamento retangular dos áudios sem sobreposição, sendo responsável tanto por produzir um número maior de amostras para treinamento e validação da rede, além de aumentar a eficiência do processo de extração de características (FERMINO, 2017). Não foi feito janelamento com sobreposição para não aumentar a correlação entre as amostras. O janelamento retangular, mostrado na Figura 2, consiste em dividir os arquivos

de áudio em amostras menores. Para esse estudo, foram usadas janelas de 4096 pontos sem sobreposição, sendo geradas 813 amostras a partir dos 36 arquivos de áudio iniciais, para garantir um treinamento mais eficaz da rede neural artificial.

Figura 2 – Janelamento retangular sem sobreposição.



Fonte: Adaptado de Fermino (2017, p. 60).

EXTRAÇÃO DE CARACTERÍSTICAS

Após o pré-processamento, foram aplicadas algumas topologias de famílias *wavelet packet,* descritas na Tabela 1, todas com 4 níveis de decomposição, sendo geradas 16 folhas de decomposição por família, as quais 8 são folhas de aproximação e 8 são de detalhe. Em seguida, foi calculada a energia em cada uma dessas folhas, e o vetor resultante, de 16 posições de entrada para cada amostra de áudio, é o que foi usado como entrada no classificador.

Tabela 1 – Topologia das famílias *wavelets*

Famílias *wavelets*	Variações
Haar	-
Daubechies	2, 3, 4, 5
Coiflets	1, 2, 3, 4, 5
Symlets	2, 3, 4, 5

Fonte: Autoria própria (2020).

CLASSIFICAÇÃO

Para o processo de classificação das amostras em edemas de Reinke e saudáveis, foi utilizada a rede neural, que é um modelo computacional inspirado no sistema nervoso de seres vivos (SILVA; SPATTI; FLAUZINO, 2010, p.24), do tipo perceptron multicamadas, com algoritmo de treinamento *Levenberg-Marquardt*, que tem a estratégica de *backpropagation*, mas realiza o treinamento das redes de perceptron multicamadas muito mais rápido que o algoritmo de *backpropagation* convencional quando a rede tem menos de centenas de pesos (HAGAN; MENHAJ, 1994). O Quadro 1 mostra as configurações utilizadas para a rede neural artificial, assim como a arquitetura utilizada, a quantidade de amostras, quantas vezes a

topologia foi testada e, além disso, quantas amostras foram utilizadas para treinamento e quantas amostras foram utilizadas para teste.

Quadro 1 – Configurações utilizadas na rede neural artificial

Parâmetros da rede	Configuração utilizada
Arquitetura	Perceptron multicamadas
Algoritmo de treinamento	Levenberg-Marquardt
Erro final de treinamento	10^{-6}
Taxa de aprendizagem	10^{-2}
Épocas	2000
Número total de amostras	813
Treinamento/Validação	570 (70%) / 243 (30%)
Repetições por topologia	8
Camadas escondidas	3
Neurônios por camada escondida	[10 9 7]
Confiabilidade	98%

Fonte: Autoria própria (2020).

RESULTADOS

Após a realização de todas as etapas anteriores, foram gerados os arquivos de resultado, apresentados no Quadro 2, no qual é mostrado, além da taxa de acerto individual de cada família em cada repetição, a média de taxa de acerto por família. É possível notar a importância de se utilizar vários testes com a mesma topologia para atenuar a influência de *outliers*, como no teste 8 de Symlets 3, a qual teve uma taxa de acerto 22,58 pontos percentuais abaixo da média.

Quadro 2 – Configurações utilizadas na rede neural artificial

Fam.	1	2	3	4	5	6	7	8	Média
Haar	93,41	95,47	95,06	94,65	94,24	95,88	95,06	83,13	93,36
Db2	90,54	91,77	93,42	93,83	96,30	94,65	95,06	94,65	93,78
Db3	94,65	90,12	93,00	91,36	92,59	97,12	90,95	92,18	92,75
Db4	85,60	93,00	92,18	93,83	91,77	91,77	94,24	93,42	91,98
Db5	94,24	90,54	92,59	96,30	93,00	94,24	95,06	94,65	93,83

Fam.	1	2	3	4	5	6	7	8	Média
Coif1	93,83	92,18	88,89	93,42	95,47	95,06	92,59	93,83	93,16
Coif2	93,42	95,88	90,54	92,59	91,77	88,48	96,30	92,18	92,64
Coif3	97,12	82,72	93,42	89,30	96,30	93,83	93,00	95,06	92,59
Coif4	93,42	91,36	92,59	93,00	93,83	96,30	93,42	90,54	93,06
Coif5	90,12	93,42	94,65	90,54	92,18	93,42	92,59	94,65	92,70
Sym2	89,30	91,77	92,59	93,00	96,30	95,47	90,54	92,59	92,70
Sym3	92,18	90,95	91,77	92,18	92,59	89,30	89,71	65,43	88,01
Sym4	92,59	89,71	88,89	92,59	92,18	92,59	89,30	86,42	90,53
Sym5	89,71	93,00	93,42	90,54	94,65	93,00	93,00	91,77	92,39

Fonte: Autoria própria (2020).

Observando o Quadro 2, é possível notar que as variações de famílias *wavelet* que apresentaram os melhores resultados, a partir da média dos 8 treinamentos de cada família, foram db5 (93,83%), db2 (93,78%), haar (93,36%) e coif4 (93,06%). No entanto, ao analisar cada um dos treinamentos de todas as famílias, nota-se que no treinamento 1, coif3 apresentou taxa de acerto de 97,12%, mesma taxa de acerto da família db3 no treinamento 6.

Percebe-se, portanto, que os melhores resultados, tanto se analisados individualmente quanto se analisados por meio da média dos treinamentos, são de famílias distintas, e mesmo dentro da mesma família, as variações das famílias podem apresentar grandes diferenças no resultado do processo.

CONCLUSÃO

Perante a problemática proposta, foi realizado o pré-processamento das amostras de áudio disponíveis, de forma a condicioná-las a serem utilizadas como entradas para a rede de perceptron multicamadas. Para isso, foi realizada a reamostragem, remoção de silêncio e correção de DC *offset*. Em seguida, foi feito o janelamento dos áudios e extraído a energia dos nós finais das árvores de algumas variações de famílias *wavelet packet*.

O resultado do estudo comprova que a extração de características de sinais de áudio de vozes disfônicas e saudáveis por meio de transformadas *wavelet packet* é bastante eficaz e adaptável, tendo em vista que topologias diferentes apresentam características distintas.

Diante desse estudo, propõe-se como trabalhos futuros a análise do desempenho da rede, caso sejam removidas as folhas iniciais da *wavelet*, ou ainda, combinações de folhas iniciais de uma família *wavelet* em nível de decomposição alto com folhas finais de nível de decomposição baixo.

AGRADECIMENTOS

Os autores agradecem à Universidade Tecnológica Federal do Paraná Campus Cornélio Procópio por proporcionar o espaço e oportunidade para essa iniciação científica.

Os autores agradecem o Grupo de Engenharia Médica do Conselho Nacional de Desenvolvimento Científico e Tecnológico (GPEM/CNPq) por compartilhar a base de dados.

REFERÊNCIAS

BARIZÃO, A. H. **Estimação do Grau de Parâmetros Subjetivos Vocais Aplicando Redes Neurais Artificiais.** 2017. Trabalho de Conclusão de Curso – Engenharia Elétrica – Universidade Tecnológica Federal do Paraná.

BEHLAU, M. **A Voz: O livro do especialista**. Vol. I. Rio de Janeiro, RJ: Revinter, 2001.

FERMINO, M. A. **Classificação de distúrbios vocais utilizando redes neurais artificiais.** 2017. Trabalho de Concusão de Curso – Engenharia Elétrica – Universidade Tecnológica Federal do Paraná.

GONZÁLES, I. V. **Videolaringoestroboscopia: uma técnica para visualizar las cuerdas vocales.** Estudios de fonética experimental, 2008. Disponível em: https://www.researchgate.net/publication/40901924_Videolaringoestroboscopia_una_tecnica_para_visualizar_las_cuerdas_vocales. Acesso em: 03 set. 2020.

HAGAN, M. T; MENHAJ, M. B. **Training feedforward networks with the Marquardt algorithm**. IEEE Trans Neural Netw, 1994. Disponível em: https://pubmed.ncbi.nlm.nih.gov/18267874/. Acesso em: 03 set. 2020.

LIMA, A. A. M. **Classificação de disfonias utilizando redes neurais artificiais e transformadas wavelet packet**. 2018. Trabalho de Conclusão de Curso – Engenharia de Controle e Automação – Universidade Tecnológica Federal do Paraná.

SILVA, I. N.; SPATTI, H. D.; FLAUZINO, R. A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas.** São Paulo: Artliber, 2010.

TSUJI, D. H. et al. **Improvement of Vocal Pathologies Using High-Speed Videolaryngoscopy.** International Archives of Otorhinolaryngology, Rio de Janeiro, v.18, n.3, p. 294-302, 2014. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1809-48642014000300294&lng=en&nrm=iso. Acesso em: 03 set. 2020.