https://eventos.utfpr.edu.br//sicite/sicite2019

Aplicação de rede neural artificial especialista em reconhecimento de transtornos vocais leves

Application of an artificial neural network specialized in recognition of mild vocal disorders

Mateus Morikawa
morikawamat@gmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Vinícius Suterio
vinicius.suterio@yahoo.com.br
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Aron Alexandre Martins Lima
aronmpa@gmail.com
Management Solutions, São Paulo, São Paulo, Brasil

María Eugenia Dajer
eugedajer@gmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

Recebido: 19 ago. 2019.
Aprovado: 01 out. 2019.

RESUMO

Distúrbios na laringe são caracterizados pela mudança do padrão vibratório das pregas vocais. Esse transtorno pode ter origem orgânica, descrito pela modificação anatômica das pregas, ou funcional, causado pelo abuso ou mal-uso da voz. Os métodos mais comuns para o diagnóstico são realizados por recursos invasivos de captura de imagens que causam desconforto para o paciente. Além disso, o grau leve da disfonia não impede o indivíduo de realizar o uso da voz, dificultando a identificação do problema e aumentando a possibilidade de complicações. Por tais motivos, objetivou-se realizar o estudo de uma alternativa não invasiva e de menor custo para a identificação de vozes disfônicas de grau leve. Para tanto, o presente trabalho visa a aplicação da transformada *Wavelet Packet* junto às Redes Neurais Artificiais do tipo *Perceptron* Multicamadas. Desta forma, foram extraídas as medidas de energia e entropia das famílias *Daubechies* 2 e *Symlet* 2. Observou-se que a *Symlet* se mostrou mais eficiente em sua generalização obtendo acurácias de 98,67% e 97,73% para medidas de energia e entropia, respectivamente. A família *Daubechies,* por sua vez, obteve-se acurácias de 94,71% e 78,56%.

PALAVRAS-CHAVE: Disfonia. Transformada Wavelet Packet. Perceptron Multicamadas.

ABSTRACT

LARYNGEAL DISORDER ARE CHARACTERIZED BY A CHANGE IN THE VIRATORY PATTERN OF THE VOCAL FOLDS. THIS DISORDER MAY HAVE NA ORGANIC ORIGIN, DESCRIBED BY ANATOMICAL FOLD MODIFICATION, OR FUNCTIONAL, CAUSED BY ABUSE OR VOICE MISUSE. THE MOST COMMON DIAGNOSTIC METHODS ARE PERFORMED BY INVASIVE IMAGING FEATURES THAT CAUSE PATIENT DISCOMFORT. IN ADDITION, THE MILD DEGREE OF DYSPHONIA DOS NOT PREVENT THE INDIVIDUAL FROM USING THE VOICE, WHICH MAKES IT DIFFICULT TO IDENTIFY THE PROBLEM AND INCREASES THE POSSIBILITY OF COMPLICATIONS. FOR THOSE REASONS, THE GOAL WAS TO STUDY A NONINVASIVE AND LOWER COST ALTERNATIVE FOR THE IDENTIFICATION OF MILD DYSPHONIC VOICES. TO MAKE IT POSSIBLE, THE PRESENT WORK AIMS TO APPLY THE WAVELET PACKET TRANSFORM TO THE ARTIFICIAL NEURAL NETWORKS FROM MULTILAYER PERCEPTRON TYPE. THUS, WERE EXTRACTED THE ENERGY AND ENTROPY MEASURES OF THE *DAUBECHIES* 2 AND *SYMLET* 2 FAMILIES. *SYMLET* WAS MORE EFFICIENT IN ITS GENERALIZATION OBTAINING 98.67% AND 97.73% OF ACCURACY BY ENERGY AND ENTROPY MEASURES, RESPECTIVELY. THE *DAUBECHIES* FAMILY, HOWEVER, OBTAINED 94.71% AND 78.56% OF ACCURACY.

KEYWORDS: Dysphonia. Wavelet Packet Transform. Perceptron Multilayer.

INTRODUÇÃO

A voz é um importante instrumento para determinados profissionais que a utilizam de forma funcional. Contudo, com o seu uso indevido, distúrbios vocais podem ser desenvolvidos. O grande problema é que o grau de disfonia leve não impede que o indivíduo atue seu cargo, além de se manifestar quase que imperceptivelmente (MEDEIROS *et al.,* 2016, p. 6).

Com o intuito de realizar o diagnóstico, existem métodos invasivos de observação direta que se fundamentam na inserção de ferramentas de captura de imagens na laringe a fim de avaliar o comportamento das pregas vocais. Entretanto, essa alternativa acaba gerando um grande desconforto ao paciente e requer experiência por parte do avaliador (LLORENTE, VILDA e VELASCO, 2006).

Visto que os sinais biológicos não são estacionários, a aplicação da Transformada de *Fourier* não vem a ser uma boa alternativa para se realizar a análise. Contudo, a Transformada *Wavelet Packet* (TWP) vem sendo empregada como alternativa atuando como extrator de características dos sinais (LIMA, 2018). Além disso, essa ferramenta melhora a performance dos classificadores de padrões, como as Redes Neurais Artificiais (RNAs), ao decompor o sinal em diversas bandas de frequência (PARRAGA, 2002, p.1).

O presente trabalho tem como propósito o estudo de uma forma alternativa para a identificação de vozes disfônicas de grau leve, sendo um método não invasivo utilizando a Transformada *Wavelet Packet* junto às Redes Neurais Artificiais.

MATERIAIS E MÉTODOS

Para este trabalho, foi utilizado a ferramenta de desenvolvimento *MATLAB*, por conter as funções necessárias para o estudo, e o *software* de edição de áudio *Audacity*.

A base de dados utilizada é constituída por 74 arquivos de áudio classificados em 3 grupos: 25 vozes sem desvio vocal, 29 vozes com desvio vocal leve e 20 vozes com desvio vocal moderado, conforme os índices perceptivo-auditivo observados por Yamasaki *et al* (2016, p. 67-71) e apresentados na Tabela 1. Visto que o objetivo é identificar transtornos de grau leve, as vozes serão divididas dois grupos: pertencente, representado pelos indivíduos com grau leve e não pertencente, sendo composto pelos indivíduos saudáveis e de grau moderado.

Tabela 1 – Limiares de classificação do desvio vocal.

Normal	0 – 35,5
Leve	35,6 – 50,5
Moderado	50,6 – 90,5

Fonte: Autoria própria (2019).

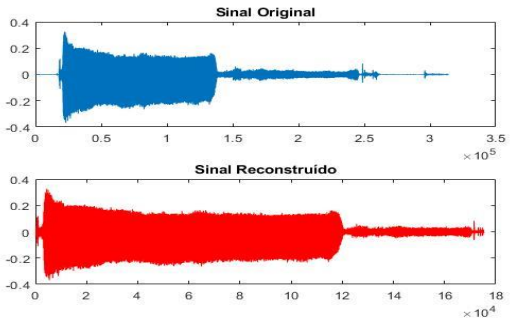
Como mencionado em Lima (2018, p. 40), um projeto de classificador de padrões se torna mais eficaz quando composto pelas seguintes etapas: pré-processamento, segmentação, extração de características, classificação e pós-processamento.

A fase de pré-processamento consistiu em remover os períodos de silêncio dos arquivos de áudio, bem como quaisquer tipos de som que não fossem do paciente, denominados artefatos. Antes, porém, foi preciso aplicar a função *detrend* do *MATLAB* para evitar que o fenômeno *DC offset* interferisse no reconhecimento de silêncio. Sendo assim, para a certificação dos períodos em que havia atividade vocal, foi realizada uma análise por *frames* de 25 ms de duração, sugerido por Paliwal, Lyons e Wójcicki (2010, p.1). Por conseguinte, obteve-se o maior valor de amplitude de cada *frame* para ser comparado com o limiar de 0,03 obtido empiricamente. Desta forma, os *frames* que possuíram a maior amplitude superior ao limiar foram considerados períodos com presença de voz. Logo, com o auxílio da função *reshape do MATLAB,* foi realizada a reconstrução do sinal com apenas esses trechos, removendo assim, os intervalos de silêncio como mostra a Figura (1). Visto que a aplicação de filtros pode acarretar a atenuação de características importantes, os artefatos tiveram que ser retirados manualmente com o uso do *Audacity.* O último estágio do pré-processamento se fundamentou em concatenar todos os arquivos de áudio, cada qual com a sua respectiva classe, e extrair seu valor rms (*root mean square*). Em seguida, o sinal concatenado foi normalizado pelo rms calculado como mostra a Equação (1).

$$x_{norm} = \frac{x_0}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}|x_n|^2}} \quad , \tag{1}$$

sendo x_{norm} o sinal normalizado, x_o o sinal concatenado não normalizado e N o número de amostras.

Figura 1 – Comparativo entre o sinal original e sem silêncio



Fonte: Autoria própria.

Na etapa da segmentação, objetivou-se a separação dos dados em conjunto de treinamento e teste. Para tanto, foi necessário mapear os sinais concatenados do estágio anterior com o desígnio de possuir novamente os áudios individuais. Posto isso, foi factível a segmentação dos arquivos em 80% treinamento e 20% teste por indivíduos de cada grupo aplicando janelamento de 4096 amostras discretizadas e *overlap* de 50%. O número de pessoas separadas para cada ocasião é representado pela Tabela 2, enquanto a quantidade de amostras geradas é mostrada na Tabela 3.

Tabela 2 – Segmentação dos arquivos em treinamento e teste.

	Pertencente	Não pertencente
Treinamento	23	36
Teste	6	9

Fonte: Autoria própria (2019).

Tabela 3 – Número de amostras geradas pelo processo de segmentação.

	Pertencente	Não pertencente
Treinamento	4402	7723
Teste	1156	1843

Fonte: Autoria própria (2019).

Visto que a quantia de amostras da classe não pertencente é superior, foi implementado uma rotina que escolhesse aleatoriamente 57% das amostras correspondentes ao grau moderado e saudável do conjunto de treinamento e 62% do conjunto de teste.

A ferramenta escolhida para extrair as características, foi a Transformada *Wavelet Packet* (TWP), pois essa, obtém informações tanto no domínio do tempo quanto na frequência. Optou-se também em utilizar as famílias *Daubechies* 2 e *Symlet* 2, por se mostrarem eficientes em Lima (2018), extraindo as medidas de energia e a entropia de *Shannon* dos coeficientes de aproximação e detalhe.

O processamento foi realizado com a rede *Perceptron* Multicamadas com o algoritmo de aprendizagem *Levenberg-Marquardt*, empregando a função tangente hiperbólica nas camadas intermediárias e taxa de aprendizagem de 0,2. A topologia utilizada é representada por duas camadas intermediárias, a qual possui 1 neurônio na primeira e 2 neurônios na segunda. Antes, entretanto, fez-se um escalonamento baseado no princípio de segmentos proporcionais de Tales nas amostras de entrada da PMC, como recomenda Silva, Spatti e Flauzino (2010, p. 159). Visto que a PMC tem o processo de aprendizagem supervisionado, fez-se necessário a indicação dos valores desejados das respostas. Dessa maneira, para que as saídas se adequassem melhor à região de operação da função tangente hiperbólica, definiu-se a saída para a classe pertencente o vetor [1 -1]. Para as amostras que não pertencem a classe desejada o vetor [-1 1]. Caso o resultado não se encaixasse em nenhuma das duas opções, foi designado o vetor [2 2] indicando incerteza.

Por fim, a fase de pós-processamento consistiu em ajustar os vetores de saída produzidas pela RNA. Sendo assim, foi estabelecido um grau de 98% de confiabilidade. Dessa forma, cada uma das duas posições do vetor de saída foi comparada com o limiar de ±0,98. Logo, caso o valor do termo fosse superior à 0,98, esse receberia 1. Se o valor do termo fosse menor do que -0,98, esse receberia -1. Para os valores entre -0,98 e 0,98, o termo receberia 2.

RESULTADOS E DISCUSSÕES

Para evitar que a aleatoriedade da inicialização dos pesos sinápticos interfira na resposta final, a rede foi treinada e testada 10 vezes. Objetivando realizar uma análise mais detalhada do classificador, foram montadas as matrizes de confusão

de cada família com a média dos 10 testes. Conforme os Quadros 1, 2, 3 e 4, é possível averiguar que o trabalho obteve acurácia de 98,67% e 97,73% para as medidas de energia e entropia da *Symlet* e, 94,71% e 78,56% para as mesmas medidas, porém, com a família *Daubechies.*

Quadro 1 – Matriz de confusão (Energia Symlet)

	Leve	Saudável e Moderado	Incerteza
Leve	1153,20 99,76%	1,70 0,15%	1,10 0,10%
Saudável e Moderado	13,00 1,14%	1114,30 97,57%	14,70 1,29%

Fonte: Autoria própria (2019).

Quadro 2 – Matriz de confusão (Entropia Symlet)

	Leve	Saudável e Moderado	Incerteza
Leve	1150,90 99,56%	3,60 0,31%	1,50 0,13%
Saudável e Moderado	25,00 2,19%	1099,60 96,29%	17,40 1,52%

Fonte: Autoria própria (2019).

Quadro 3 – Matriz de confusão (Energia Daubechies)

	Leve	Saudável e Moderado	Incerteza
Leve	1053,90 91,17%	42,50 3,68%	50,60 5,16%
Saudável e Moderado	5,70 0,50%	1122,00 98,29%	13,80 1,21%

Fonte: Autoria própria (2019).

Quadro 4 – Matriz de confusão (Entropia Daubechies)

	Leve	Saudável e Moderado	Incerteza
Leve	814,50 70,46%	22,80 1,97%	328,70 28,43%
Saudável e Moderado	3,90 0,34%	990,70 86,75%	147,40 12,91%

Fonte: Autoria própria (2019).

Os resultados apresentados nas matrizes de confusão, sugerem que a *Symlet* obteve um desempenho superior à *Daubechies*, como se pode aferir a partir das medidas de incertezas, erros e acertos na identificação da classe desejada.

Apesar de trabalhos anteriores terem mostrados que as famílias *Daubechies* 2 e *Symlet* 2 foram eficientes para a análise de sinais vocais, para esse estudo o desempenho da Daubechies 2 por medida de entropia não teve um bom resultado. No entanto, por intermédio do Quadro 4, percebe-se que embora o rendimento da *Daubechies* 2 tenha decrescido, grande parte da causa é devido ao incremento na taxa de incerteza. Uma vez que este estudo se trata de uma aplicação de RNA para o auxílio de profissionais em diagnósticos, se torna mais viável a rede neural ter indecisão ao invés de classificar amostras incorretamente.

Também, observa-se que apenas 3 neurônios nas camadas intermediárias já foram o suficiente para desempenhar uma boa generalização, não requisitando assim, uma grande performance computacional.

CONCLUSÃO

Esta pesquisa se fundamentou no treinamento de uma rede neural especialista em reconhecer transtornos vocais de grau leve. Para tanto, o trabalho se baseou em uma cronologia de processos que se inicia desde o tratamento de dados, e se encerra no classificador. Partindo desse método, foi possível a geração dos resultados que mostram a efetividade de se utilizar essas duas famílias de *Wavelets* para a análise de sinais vocais, sendo essa circunstância também verificada em Lima (2018, p. 61). Por fim, através da Tabela 1, constata-se que os limiares de transição de um nível de desvio vocal para o outro são muito próximos. Devido a isso, conclui-se que a RNA se mostrou robusta o suficiente para gerar uma elevada taxa de acerto em sua classificação, em que, na maioria das vezes, superou os 94% de acurácia com 98% de confiabilidade.

REFERÊNCIAS

MEDEIROS, J. da S. A. et al. **Sintomas vocais relatados por professoras com disfonia e fatores associados**. Audiology Communication Research, v. 21, São Paulo, 2016. Disponível em: http://www.scielo.br/pdf/acr/v21/2317-6431-acr-2317-6431-2015-1553.pdf. Acesso em: 19 ago. 2019.

LLORENTE, J. I. G.; VILDA, P. G.; VELASCO, M. B. **Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters**. IEEE Transaction on Biomedical Engineering, v. 53, n. 10, set. 2006. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1703745. Acesso em: 19 jul. 2019.

LIMA, A. A. M. **Classificação de Disfonias Utilizando Redes Neurais Artificiais e Transformadas Wavelet Packet**. Trabalho de Conclusão de Curso em Engenharia de Controle e Automação, Universidade Tecnológica Federal do Paraná, Cornélio Procópio, PR, 2018.

PARRAGA, A. **Aplicação da Transformada Wavelet Packet na Análise e Classificação de Sinais de Vozes Patológicas**. Dissertação de Mestrado em Engenharia Elétrica, Escola de Engenharia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, 2002.

YAMASAKI, R. et al**. Auditory-perceptual Evaluation of Normal and Dysphonic Voices Using the Voice Deviation Scale**. Journal of Voice, v. 31, n. 1, p. 67-71, 2016.

PALIWAL, K. K.; LYONS, J. G.; WÓJCICKI. K. K. **Preference for 20-40 ms window duration in speech analysis**. 4th International Conference on Signal Processing and Communication Systems, Austrália, 2010. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5709770. Acesso em: 20 jul. 2019.

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