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

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

Application of a specialist artificial neural network in the recognition of moderate vocal disorders

Eduardo Henrique da Silva
e_duhsilva@hotmail.com
Universidade Tecnológica Federal do Paraná, Cornélio Procópio, Paraná, Brasil

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

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

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

RESUMO

A voz é um elemento primordial para a realização de grande parte das atividades feitas pelos seres humanos, sejam elas do âmbito de lazer ou de trabalho. Distúrbios vocais ocorrem em um número elevado de pessoas, e podem ser causados por inúmeros motivos, sendo assim, tem-se como necessário um diagnóstico rápido e eficiente para o seu devido tratamento. O objetivo desse trabalho foi realizar o reconhecimento de transtornos vocais moderados por meio da aplicação de uma rede neural artificial especialista. Para a sua devida execução, foram necessárias determinadas etapas, sendo elas: o tratamento do banco de dados utilizado; o pré-processamento dos sinais de voz; a extração das características dos mesmos (energia e entropia) por meio da Transformada *Wavelet Packet*, e, por fim, a aplicação da rede neural artificial especialista, na qual são realizados seus devidos treinamentos e testes. Foi possível obter uma taxa de acerto total de 82,2% utilizando a energia extraída dos sinais de voz, ao mesmo tempo que, utilizando a entropia, foi obtida uma taxa de 99,5% de acerto.

PALAVRAS-CHAVE: Distúrbios da voz. Transformada *Wavelet Packet*. *Perceptron* Multicamadas.

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

ABSTRACT

The voice is a key element for the accomplishment of most of the activities done by human beings, whether recreation or work. Vocal disorders occur in a large number of people, and can be caused by a number of reasons, so a fast and efficient diagnosis is necessary for its proper treatment. The objective of this study was to recognize moderate vocal disorders by the application of a specialized artificial neural network. For its proper implementation, certain steps were required, namely: the database treatment used; preprocessing of voice signals; the extraction of their characteristics (energy and entropy) using the Wavelet Packet Transform, and, finally, the application of the specialist artificial neural network. It was possible to achieve a total hit rate of 82.2% using the energy extracted from the signals, while using entropy, a 99.5% accuracy rate was obtained.

KEYWORDS: Voice disorders. Wavelet Packet Transform. Perceptron Multilayer.

INTRODUÇÃO

Behlau (2001, p. 2) afirma que "a voz é uma característica própria do indivíduo, que pode até mesmo informar as condições de saúde, de idade, de sexo, do estado emocional e de traços da personalidade do indivíduo". Além de ser um fator determinante nas características de uma determinada pessoa, a voz pode ser considerada como o meio de comunicação mais utilizado em todo o mundo. Pode-se classificar a voz em duas classes, a eufônica e a disfônica. As vozes eufônicas são as vozes saudáveis que não apresentam nenhum tipo de distúrbio, enquanto as vozes disfônicas apresentam perturbações em sua emissão (PENTEADO; PEREIRA, 2006, p. 19-28).

Para auxiliar os profissionais da área da saúde, é possível utilizar ferramentas que ajudem no reconhecimento de disfonias em indivíduos, por meio de processos computacionais desenvolvidos em base a um banco de dados confiável. A voz pode ser representada como um sinal variante no tempo. Uma das etapas essenciais para a extração de características dos sinais de voz é a aplicação da Transformada *Wavelet Packet* (TWP). "Esta ferramenta permite analisar eventos irregularmente distribuídos e séries temporais que contenham potências não-estacionárias em diferentes frequências" (GUIMARÃES; FREIRE; TORRENCE, 2013, p. 1), sendo de grande utilidade para a análise de sinais de voz.

A representação da voz como um sinal é essencial para a realização do processamento digital da mesma, para esta ser utilizada posteriormente, em uma ferramenta computacional de reconhecimento de padrões, conhecida como rede neural artificial (RNA). As RNAs podem ser definidas como máquinas que simulam o funcionamento do cérebro humano, possuindo assim neurônios artificiais, que passam por um determinado processo de aprendizagem, sendo treinados e testados para realizar uma tarefa ou uma função específica. O processo de aprendizagem é realizado através do treinamento da RNA, na qual os pesos sinápticos, forças de conexão entre os neurônios, acumulam o conhecimento adquirido durante o processo (HAYKIN, 2001, p. 28).

Existem diversos tipos de arquiteturas de RNAs, neste trabalho foi utilizada a arquitetura *feedforward* de camadas múltiplas. Existem também diferentes tipos de redes neurais e processos de aprendizagem, no caso, foram utilizados uma RNA especialista e o treinamento supervisado, vistos como mais efetivos para este tipo de aplicação (SILVA; SPATTI; FLAUZINO, 2016, p. 45).

MATERIAIS E MÉTODOS

A realização do trabalho só foi possível em razão da obtenção do banco de vozes, que foi cedido gentilmente pela Dra. Fabiana Zambon do Sindicato de Professores Privados de São Paulo SINPRO-SP. O banco de dados em questão continha um total de 75 áudios, com gravação da vogal /e/ sustentada, tendo uma diversidade de vozes eufônicas (saudáveis) e disfônicas de grau leve, moderado e intenso. Porém, os áudios não estavam classificados separadamente no banco de dados, portanto, foi necessário utilizar o método descrito por YAMASAKI et al (2016) para a devida classificação dos áudios em eufônicos, disfônicos de grau leve, de grau moderado e de grau intenso.

Após a classificação dos áudios, foi possível observar a quantidade de áudios pertencentes à cada grupo, como descrito no Quadro 1.

Quadro 1 – Quantidade de áudios pertencentes à cada grupo.

	Saudáveis	Grau leve	Grau moderado	Grau intenso
Quantidade de áudios	25	29	20	1

Fonte: Autoria Própria (2019).

Com os áudios devidamente separados, deu-se início a etapa de pré-processamento dos sinais de voz. Esta etapa, e todas as outras subsequentes, foram realizadas utilizando o *software* MATLAB, e, exclusivamente para esta etapa, foi utilizado o *software Audacity*.

A etapa de pré-processamento foi iniciada com a remoção do silêncio presente nos áudios, esta tarefa consistiu em retirar todos os trechos nos áudios nos quais ocorria silêncio, mantendo somente os trechos nos quais ocorria a vogal /e/ sustentada. Esse processo foi realizado a partir da construção de uma rotina no *software* MATLAB. Logo após, foi removido o sinal DC *offset*, sinal contínuo e de baixa frequência, que se encontra presente nos áudios. Esse sinal é proveniente dos componentes eletrônicos utilizados para a gravação das vozes, e é prejudicial para o bom funcionamento da RNA se não for removido. Para detectar e remover este tipo de sinal foi utilizada a função *detrend* presenta na biblioteca de funções do *software* MATLAB.

Para a remoção de possíveis artefatos presentes nos áudios, foi utilizado o *software Audacity*. Um artefato pode ser definido como um trecho do áudio que possua qualquer tipo de som que não seja a vogal /e/ sustentada, como tosses, risadas, entre outros. Esta etapa é de extrema importância, pois, a não remoção deste tipo de elementos podem introduzir dados errôneos na RNA que causem a diminuição de assertividade no reconhecimento de padrões e classificação dos sinais. Os artefatos foram removidos manualmente, utilizando o software citado anteriormente. Na Figura 1 é mostrado o exemplo de um áudio pré-processado.

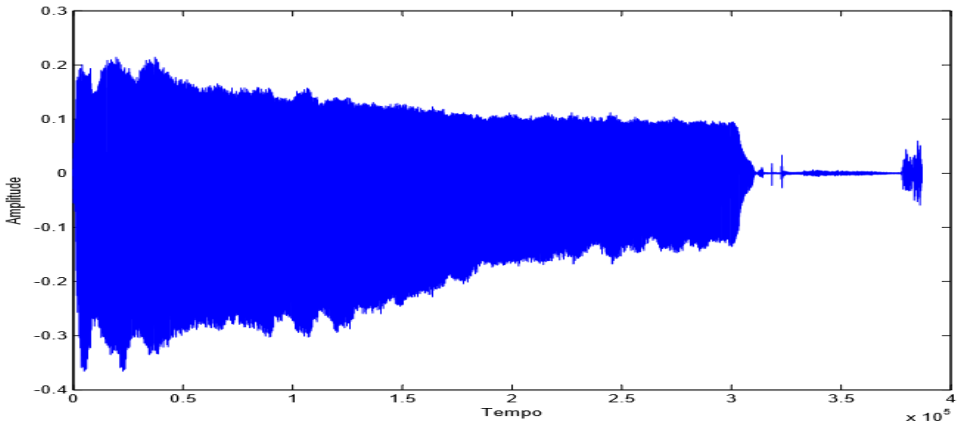

Figura 1 – Sinal de áudio resultante do pré-processamento, realizado com o software *Audacity.*
Fonte: Autoria Própria.

Antes do início do processo de extração de características, foi realizada uma separação dos áudios a serem utilizados. Para isso, foram incluídos todos os áudios do grupo de disfonias de grau moderado e apenas 11 áudios do grupo saudável. Os áudios pertencentes ao grupo disfônico de grau leve não foram utilizados, pois os mesmos não apresentaram uma boa adaptação à RNA, comprometendo os resultados. Assim como não foi utilizado o grupo de disfonias de grau intenso, devido à pequena quantidade de áudios pertencentes ao mesmo.

Foi iniciado, então, o processo de extração de características, o qual teve como primeira tarefa a concatenação de todos os áudios, para que fosse realizada a normalização por meio da função *mapminmax*, presente na biblioteca do *software* MATLAB. Com os áudios separados novamente, foi realizado um *overlap* (sobreposição) de 50% nos áudios, totalizando 7100 amostras e separando-as em uma porção para treinamento e outra para teste da RNA. Foram separadas 5680 amostras para treinamento e 1420 para teste. Com as amostras já separadas para treinamento e teste, e com sobreposição de 50% aplicada, foi aplicada a TWP da Família *Daubechies* 2, com 5 níveis de decomposição e janelamento de 4096 amostras, escolhas feitas a partir da classificação feita por Lima (2018, p. 14). A TWP foi aplicada para extrair a energia e a entropia das amostras, a partir das funções *wenergy* e *wentropy*, presentes na biblioteca do *software* MATLAB.

Em seguida, as matrizes de energia e entropia das amostras de treinamento e teste obtidas pela TWP foram normalizadas levando em conta a função de ativação a ser utilizada na RNA, que no caso foi a logística (SILVA; SPATTI; FLAUZINO, 2016, p. 38). Sendo assim, foi possível iniciar o processo de treinamento e teste da RNA especialista. A saída da RNA foi construída com a resposta à disfonia de grau moderado representada pelo vetor [1 0], enquanto a resposta ao grupo saudável pelo vetor [0 1]. As características fixadas da RNA especialista estão descritas no Quadro 2.

Quadro 2 – Características da RNA especialista utilizada.

Taxa de Aprendizagem	0,2
Épocas	200
Algoritmo de Aprendizagem	Levenberg-Marquardt
Função de Ativação (intermediária)	Logística
Função de Ativação (saída)	Rampa Linear

Fonte: Autoria Própria

RESULTADOS E DISCUSSÃO

Foram utilizadas 720 amostras do grupo disfônico de grau moderado e 700 amostras do grupo eufônico (saudável), totalizando 1420 amostras de teste. Após a realização de 20 testes com diferentes topologias da RNA, utilizando um grau de confiabilidade de 98%, foi possível identificar o tipo de topologia mais efetivo, a utilização de apenas uma camada intermediária, contendo 2 neurônios artificiais. A partir da topologia citada anteriormente, foram coletados os resultados da RNA após realizar o treinamento e teste por 10 vezes, utilizando tanto os valores de energia, mostrados na Figura 2, como os de entropia, mostrados na Figura 3. Pode-se observar pela Figura 3, que foi obtida uma taxa de acerto total maior quando utilizada a entropia, extraída dos sinais de áudio pela TWP.

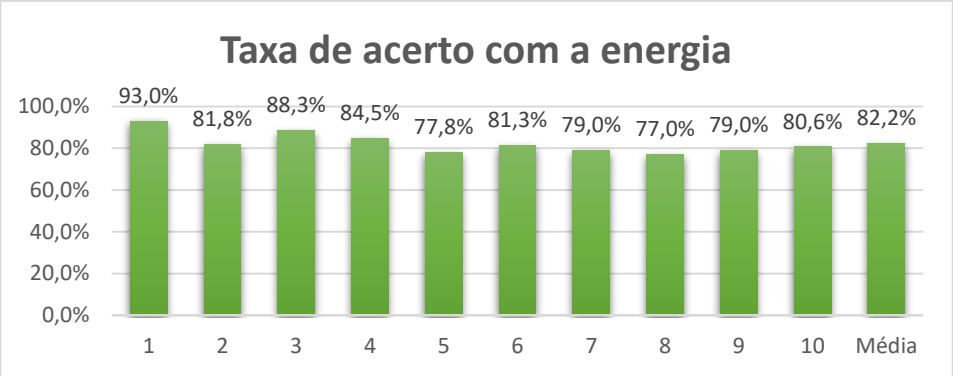

Figura 2 – Taxa de acerto obtida utilizando a energia.
Fonte: Autoria Própria (2019).

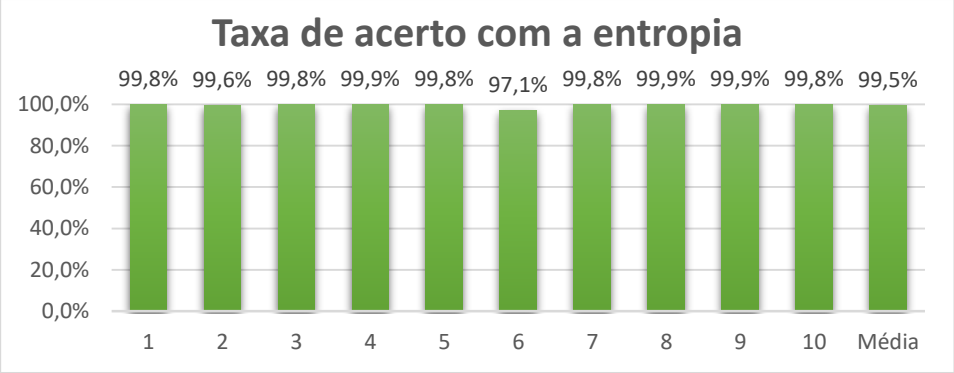

Figura 3 – Taxa de acerto obtida utilizando a entropia.
Fonte: Autoria Própria (2019).

As Tabelas 1 e 2 mostram as matrizes confusão obtidas a partir do valor médio de acerto, da entropia e da energia, respectivamente.

Tabela 1 – Matriz confusão da média dos resultados obtidos com a energia.

	Grau Moderado	Saudáveis	Incerteza
Grau Moderado	98,90%	0,04%	1,06%
Saudáveis	31,38%	65,22%	3,40%

Fonte: Autoria Própria (2019).

Tabela 2 – Matriz confusão da média dos resultados obtidos com a entropia.

	Grau Moderado	Saudáveis	Incerteza
Grau Moderado	99,58%	0,10%	0,32%
Saudáveis	0%	99,54%	0,46%

Fonte: Autoria Própria (2019).

É possível analisar, a partir da matriz confusão dos valores obtidos com a energia, que a RNA obteve um rendimento satisfatório na identificação das

amostras pertencentes ao grupo disfônico de grau moderado, porém, teve rendimento um pouco menor na identificação das amostras do grupo saudável, obtendo um erro maior na classificação desse grupo. Este fato pode ser justificado por uma adaptação não tão boa ao grupo em questão, que pode ser explicado pela menor quantidade de áudios do grupo em questão presente no banco de dados, para treinamento da RNA. Ao mesmo tempo que, pela matriz de confusão dos valores obtidos com a entropia, percebe-se um rendimento satisfatório tanto na identificação das amostras do grupo disfônico de grau moderado, como das amostras do grupo saudável, mostrando assim uma melhor adaptação da RNA, e comprovando um melhor rendimento desta, em geral, quando é utilizada a entropia.

CONCLUSÕES

Com os resultados da RNA especialista, conclui-se que o método de pré-processamento dos sinais de voz em conjunto à extração das características desses sinais a partir da TWP, mostraram-se muito efetivos para essa aplicação. Nota-se também a eficiência da RNA como uma ferramenta de reconhecimento de transtornos vocais de grau moderado, pois os resultados obtidos após o treinamento e o teste foram satisfatórios, tendo um grau de confiabilidade alto para a aplicação. Confiabilidade, necessária para ferramentas de auxílio diagnóstico.

Portanto, após todos os procedimentos realizados, desde a verificação e tratamento do banco de dados utilizado, até o treinamento e teste da RNA especialista, conclui-se que os objetivos relacionados à execução do trabalho foram cumpridos satisfatoriamente.

REFERÊNCIAS

BEHLAU, M; AZEVEDO, R; PONTES, P. Conceito de voz normal e classificação das disfonias. **Voz: o livro do especialista.** Rio de Janeiro: Revinter, v.1, 2001. p. 1.

GUIMARÃES, C, A, S; FREIRE, P, K, M; TORRENCE, C**. A Transformada *Wavelet* e sua Aplicação na Análise de Séries Hidrológicas.** Revista Brasileira de Recursos Hídricos, v.18, n.3, 2003, p. 271-280.

HAYKIN, Simon. **Redes Neurais:** Princípios e Prática**.** 2ª. ed. Hamilton,Ontario, Canadá. Bookman, 2001, p. 28.

YAMASAKI, R; MADAZIO, G; LEÃO, S, H, S; PADOVANI, M; AZEVEDO, R; BEHLAU, M. **Auditory-perceptual Evaluation of Normal and Dysphonic Voices Using the Voice Deviation Scale**. Journal of Voice. Auckland, v.31, n.1, 2016, p.67-71.

SILVA, I. SPATTI, D. H. FLAUZINO, R. A. **Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas.** São Paulo: Artliber Editora Ltda., 2016, p. 45.

PENTEADO, R, Z; PEREIRA, I, M, T, B**. *Avaliação do impacto da voz na qualidade de vida de professores*.** Revista Soc. Bras. Fonoaudiol, 2(2), 2006, p.19-28.

LIMA, A, A, M. **Identificação de Disfonias Utilizando Redes Neurais Artificiais**. 2018, p. 14.