UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS OLIVEIRA PEREIRA

INTELIGÊNCIA ARTIFICIAL APLICADA NO AUXÍLIO À DETECÇÃO DE PATOLOGIAS VOCAIS

CAMPINA GRANDE - PB

2022

MATHEUS OLIVEIRA PEREIRA

INTELIGÊNCIA ARTIFICIAL APLICADA NO AUXÍLIO À DETECÇÃO DE PATOLOGIAS VOCAIS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Professora Dra. Joseana Macêdo Fechine Régis de Araújo.

CAMPINA GRANDE - PB

2022

MATHEUS OLIVEIRA PEREIRA

INTELIGÊNCIA ARTIFICIAL APLICADA NO AUXÍLIO À DETECÇÃO DE PATOLOGIAS VOCAIS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professora Dra. Joseana Macêdo Fechine Régis de Araújo

Orientadora – UASC/CEEI/UFCG

Professor Dr. Herman Martins Gomes

Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 06 de Abril de 2022.

CAMPINA GRANDE - PB

ABSTRACT

Voice hearing is a form of voice analysis in which a physician diagnoses a patient's voice as pathological or not after hearing it. The problem with this method is its subjective nature, given that the result may vary depending on the examiner. Laboratory techniques may be used to achieve a more precise diagnosis, but given their invasive and expensive nature, they are frequently avoided by the patients. Due to that, more research is being made concerning acoustic analysis, as it is a non-invasive and automatic form of digital processing. This method consists in using techniques of digital signal processing and pattern matching as a means to determine whether a voice is pathological or not. In light of this, the present article aims to analyze the use of an artificial neural network (ANN) as a classifier, as well as the use of Mel-frequency cepstral coefficients (MFCC) to help in the voice pathology detection. The Saarbruecken Voice Database (SVD) was used for the training and validation of the ANN. Experimental results obtained with k-fold cross-validation show that the solution achieved accuracy above 85% in distinguishing healthy and pathological voices.

Inteligência Artificial Aplicada no Auxílio à Detecção de Patologias Vocais

Matheus Oliveira Pereira
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande,
Campina Grande, Paraíba, Brasil
matheus.pereira@ccc.ufcg.edu.br

Joseana Macêdo Fechine Régis de Araújo
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande,
Campina Grande, Paraíba, Brasil
joseana@computacao.ufcg.edu.br

RESUMO

A escuta da voz é uma forma de avaliação da saúde vocal, em que um profissional julga a voz do paciente como patológica ou não após ouvi-la. O problema desse método é o seu caráter subjetivo, devido à possibilidade do resultado variar conforme examinador. Para uma análise mais precisa, técnicas laboratoriais podem ser aplicadas; contudo, são frequentemente evitadas pelos pacientes devido ao caráter invasivo e oneroso. Assim, pesquisadores têm desenvolvido técnicas para auxiliar na discriminação de vozes patológicas usando análise acústica, por ser uma forma de processamento digital de sinais não invasiva e automática. Esse método consiste em utilizar técnicas de processamento digital de sinais e reconhecimento de padrões, para determinar se o sinal de voz é patológico ou não. Diante disso, este artigo objetiva analisar o uso de uma rede neural artificial (RNA) como classificador e características obtidas por meio de Coeficientes Mel Cepstrais (MFCC), que auxiliarão na detecção de patologias da voz. Para o treinamento e validação da RNA, foi utilizada a base de dados alemã Saarbruecken Voice Database (SVD). Os resultados demonstraram, com validação cruzada k-fold para treinamento e teste, que a solução atingiu níveis de acurácia acima de 85% na distinção entre vozes saudáveis e patológicas.

Palavras-Chave

Análise Acústica, Processamento Digital de Sinais, Coeficientes Mel Cepstrais, Rede Neural Artificial, Patologias Vocais.

1. INTRODUÇÃO

A voz é uma característica de muita importância para os seres humanos. Essa é uma das nossas principais formas de comunicação com o mundo externo, visto que a utilizamos para expressar pensamentos e sentimentos, além dela estar fortemente ligada à nossa necessidade de se agrupar e se conectar com outros indivíduos [16]. Por isso, as patologias vocais são um importante objeto de estudo na medicina e em outras áreas, visto que essas patologias afetam negativamente a capacidade humana de se expressar e se comunicar normalmente.

No contexto do diagnóstico de patologias da voz, existem diversas técnicas para a avaliação da saúde vocal de um paciente. Dentre as mais populares, encontra-se a escuta da voz, em que um profissional julga a voz como patológica ou não após ouvi-la. O problema desse método é o seu caráter subjetivo, haja vista a possibilidade do resultado variar conforme o examinador. Visando um diagnóstico mais preciso, médicos podem solicitar a aplicação de técnicas laboratoriais, a exemplo da videolaringoscopia (exame com um instrumento de fibra óptica para observação direta das dobras vocais) e da videofluoroscopia (técnica radiográfica, em que o paciente ingere uma determinada quantidade de uma substância rádio-opaca para avaliar a deglutição) [1]. Porém, exames laboratoriais são frequentemente evitados por pacientes, devido ao alto custo e ao caráter invasivo inerente aos mesmos [10].

Para reduzir a evasão de pacientes em relação a exames que fornecem diagnósticos mais precisos, outros métodos menos invasivos e menos custosos são avidamente estudados pela comunidade científica. Um deles, que é o objeto de estudo deste trabalho, é a análise acústica. Esse método consiste em obter características de um sinal de voz para que este seja processado por um classificador (como por exemplo, uma rede neural) que, por meio da detecção de padrões, irá determinar se o sinal de voz é patológico ou não. O método não é invasivo, visto que só necessita do fornecimento de um sinal de voz do paciente, que pode ser obtido por meio da gravação controlada da fala do paciente. A análise acústica é, também, menos custosa, uma vez que até mesmo computadores ou outros dispositivos (a exemplo de *smartphones*), com baixo poder de processamento, podem executar o algoritmo de classificação.

O aperfeiçoamento e a maior visibilidade, no que tange à análise acústica aplicada ao problema descrito, possuem um grande potencial para impactar no futuro dos diagnósticos das patologias vocais, deixando-os não invasivos, mais precisos, e menos onerosos.

2. TRABALHOS RELACIONADOS

A análise acústica aplicada à detecção de patologias na voz é um tema que tem ganhado cada vez mais atenção perante a comunidade acadêmica.

No âmbito do uso de *Support Vector Machines* para classificar os sinais de voz, tem-se trabalhos como o de Almeida [1], que, na extração de características relevantes do sinal, faz uso dos Coeficientes de Predição Linear, dos Coeficientes Cepstrais de Frequência Mel (ou MFCC, da sigla em inglês) e dos coeficientes obtidos através da Transformada Wavelet Packet.

No âmbito do uso de *Deep Neural Networks* para realizar a classificação dos sinais de voz, pode-se citar o trabalho

de Dias [5], que baseia a análise apenas no comportamento dinâmico dos sinais de voz, não realizando extração de características específicas do sinal. Neste trabalho, a necessidade da rede neural artificial ser do tipo profunda fica evidente, visto que a complexidade da classificação aumenta no cenário em que a extração de características relevantes não é realizada.

Também pode ser citado o trabalho de Marinus [10], que compara, no contexto da detecção de patologias vocais, o desempenho de classificadores dos tipos Redes Neurais Multilayer Perceptron, Modelos de Misturas de Gaussianas e Quantização Vetorial. Seus melhores resultados envolvem o uso de MFCC como característica relevante do sinal, e o uso de Rede Neural Multilayer Perceptron como classificador.

Conforme exposto, o uso de MFCC se mostra importante para soluções que, antes de iniciar a classificação, realizam a etapa de extração de características representativas do sinal de voz.

3. FUNDAMENTAÇÃO TEÓRICA

Nesta seção, serão apresentados conceitos necessários ao entendimento do trabalho desenvolvido.

3.1 Técnicas Tradicionais de Diagnóstico

As técnicas mais precisas e mais comumente empregadas, no diagnóstico de patologias da voz, são as laboratoriais. Compondo esse grupo, além das técnicas mencionadas em seção anterior, pode-se citar a videoestroboscopia (iluminação estróbica da laringe por meio de um endoscópio rígido, posicionado atrás da língua) e a eletromiografia (observação indireta do estado funcional da laringe). Embora bem estabelecidos para o diagnóstico de patologias na voz, esses procedimentos são dispendiosos, pois fazem uso de instrumentos sofisticados, tais como instrumentos endoscópicos, fontes de luz especial e câmeras de vídeo especializadas. Ademais, tais técnicas são, também, consideradas invasivas e de risco, por poderem causar mal estar aos pacientes e por necessitarem ser executadas em condições controladas, por profissional especializado [1, 3].

Diante desses fatores, os exames laboratoriais são frequentemente evitados pelos pacientes, o que dificulta o diagnóstico e o tratamento de patologias na voz, visto que a não realização de exames mais precisos aumenta os riscos de diagnósticos incorretos. Nesse sentido, diagnósticos incorretos pertencentes à classe dos falsos negativos são bastante problemáticos, visto que privam do paciente a oportunidade de iniciar o tratamento da patologia com mais antecedência; fator este que pode dar margem ao agravamento do quadro [6].

Devido à grande evasão de pacientes em relação aos exames laboratoriais para diagnóstico de patologias na voz, muitos pesquisadores têm dedicado esforços no desenvolvimento de técnicas menos custosas e menos invasivas. Nesse sentido, uma técnica bastante promissora e que ganha cada vez mais espaço no meio científico é a de análise acústica.

3.2 Análise Acústica

A análise acústica da voz é uma técnica bastante utilizada na detecção e estudo de patologias da voz. Nesse contexto, a análise acústica relaciona-se ao uso de técnicas computacionais que visam medir propriedades do sinal acústico de uma voz gravada [2]. Dentre tais propriedades, pode-se citar o Jitter, o Shimmer e os Coeficientes Mel Cepstrais. Tais técnicas serão abordadas em seções posteriores.

Após a definição das propriedades a serem analisadas e sua extração ter sido feita a partir do sinal de voz, é dado início ao processo de classificação. Por meio da detecção de padrões na análise das características do sinal de voz, é possível detectar se a voz possui ou não alguma patologia [4]. Porém, a tarefa de análise e classificação se torna inviável de ser conduzida de forma manual, devido ao grande número de variáveis e dados a serem observados até que um padrão seja encontrado. Diante dessa situação, são introduzidos os classificadores inteligentes e automáticos, tal qual uma Rede Neural Artificial.

3.2.1 Pré-processamento do sinal de voz

O pré-processamento do sinal de voz é feito após a aquisição do sinal, e tem por objetivo fazer com que determinadas características do sinal de voz fiquem mais bem evidenciadas, para que o classificador consiga detectar padrões de forma mais eficiente.

No contexto da análise acústica, o pré-processamento do sinal de voz usualmente é composto por quatro principais etapas.

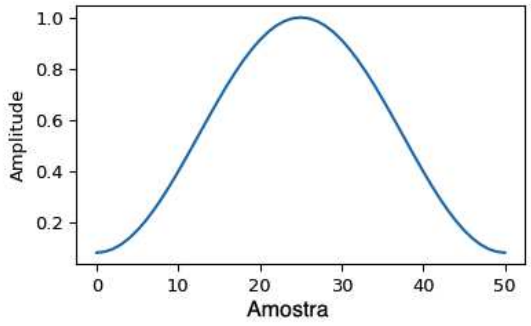
A primeira etapa consiste na eliminação de ruídos, que se dá por meio da aplicação, nos sinais de voz, de filtros especializados para esse fim. A depender da forma com que foi feita a aquisição dos dados, essa etapa é primordial para melhorar o processo de classificação. Em situações em que a aquisição dos dados foi feita de forma controlada, como por exemplo, com microfones de qualidade profissional e em ambiente com isolamento acústico, essa etapa tem menor importância, podendo ser opcional.

A segunda etapa é a de pré-ênfase. Nesse momento, o sinal de voz é passado por um filtro que reforça a magnitude de componentes de alta frequência que, em geral, são os mais importantes para a análise do sinal.

A segmentação do sinal de voz é a terceira etapa. Aqui, o sinal de voz é segmentado em *frames* (quadros) que serão processados separadamente, e cujos tamanhos normalmente variam entre 20 e 40 ms. Essa etapa é necessária devido à grande variabilidade das características do sinal de voz conforme sua duração aumenta.

Por fim, tem-se a quarta etapa, que é a de janelamento. Tal etapa faz-se necessária após a segmentação para que o efeito das descontinuidades do sinal seja minimizado. Essa etapa consiste na multiplicação do sinal por uma função janela, de forma a deixar mais evidenciadas as partes do sinal que são de maior interesse para a análise e minimizar os efeitos adversos da segmentação. Nesse sentido, uma opção comumente utilizada é a janela de Hamming, que amplifica os pontos conforme a sua proximidade da metade do sinal, resultando num formato de sino, conforme pode ser visto na Figura 1.

Figura 1. Visualização da Janela de Hamming.



Fonte: adaptado de [14].

3.2.2 Extração de Características

De forma a tornar o processo de classificação menos custoso, usualmente é extraído um vetor de características do sinal de voz, visando à redução do espaço de busca ou a melhor definição do que deve ser analisado, para que, então, seja feita a classificação. No contexto de detecção de patologias na voz, características comumente utilizadas [7] são o Jitter, o Shimmer, e os Coeficientes Mel Cepstrais (MFCC).

O Jitter é o valor que expressa a variação entre os períodos glotais (ou ciclos periódicos glotais) das ondas de um sinal de fala sustentada. Tal variação acontece quando o paciente não consegue manter a frequência das vibrações das cordas vocais na produção de uma vogal de forma sustentada. Pacientes com alguma patologia na voz costumam apresentar valores mais elevados de Jitter [7, 15].

O Shimmer é a variação da amplitude entre os ciclos glotais do sinal de voz. Essa variação acontece quando o paciente não consegue manter a elocução (volume de produção) de maneira constante. Pacientes que possuem alguma doença vocal costumam apresentar valores de Shimmer mais elevados [7].

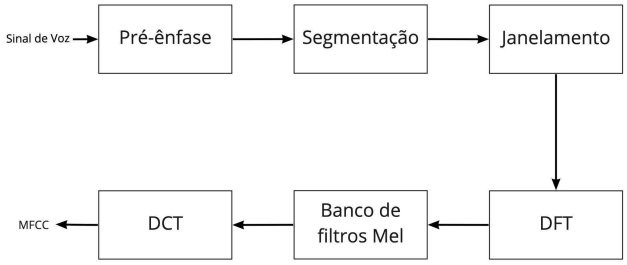
Os Coeficientes Mel Cepstrais são frequentemente usados para o reconhecimento de fala (o que está sendo dito) e reconhecimento de quem fala (locutor). Calculados com base no sistema auditivo humano e representados na escala Mel, os MFCC possuem grande potencial de representação do espectro de amplitudes do sinal de fala, com boas representações em casos de sinais sem a presença de ruídos [7]. O processo para obtenção dos MFCC é mostrado na Figura 2.

Como pode ser visto no diagrama da Figura 2, o cálculo para obtenção dos MFCC de um sinal de voz consiste em sete passos, descritos a seguir.

Passos 1 ao 3: Pré-ênfase, segmentação e janelamento

Conforme descrito na subseção anterior, esses passos compõem o pré-processamento do sinal de voz, processo que também se faz necessário no contexto de cálculo dos MFCC.

Figura 2. Etapas para cálculo dos MFCC.



Fonte: adaptado de [11].

Passo 4: Transformada Rápida de Fourier (discreta)

Tendo como entrada os *frames* resultantes da segmentação e janelamento, esta etapa consiste na conversão do domínio desses *frames*, do domínio do tempo para o domínio da frequência.

Passo 5: Processamento do banco de filtros Mel

O intervalo das frequências no espectro FFT é bastante amplo, e o sinal de voz não segue a escala linear. Um conjunto de filtros triangulares é utilizado para computar a soma ponderada dos filtros de componentes espectrais, de forma que a saída do processo se aproxima da escala Mel [11].

Passo 6: Transformada Discreta do Cosseno

Neste passo, é realizado o processo de conversão do espectro Mel para o domínio do tempo, por meio da Transformada Discreta do Cosseno. O resultado dessa conversão são os chamados Coeficientes Mel Cepstrais (MFCC).

Vale ressaltar, ainda, que no contexto de preparação do sinal de voz a ser classificado, há também a opção de não realizar a extração de características específicas. Tal abordagem implica em utilizar o próprio sinal de voz como entrada do classificador. Essa alternativa normalmente exige mais do classificador, sendo necessário que o mesmo possua níveis de detecção de padrões mais abrangentes, visto que o espaço de busca é maior quando comparado ao vetor de características específicas. Exige, ainda, uma base de dados grande (característica que não se aplica à SVD), uma vez que o aprendizado da rede precisará de mais dados para responder bem ao maior espaço de busca.

O método de classificação de sinal de voz sem extração de características específicas também foi explorado durante os experimentos referentes ao presente trabalho, e seus resultados serão apresentados em seção posterior.

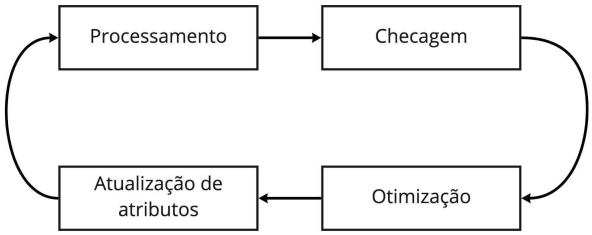
3.2.3 Classificação

Uma vez realizado o tratamento dos dados e a extração das características, os dados estarão prontos para serem inseridos na rede neural, para que esses sejam classificados.

Porém, antes que a RNA possa de fato iniciar o processo de classificação, é imprescindível que a rede passe por uma etapa de treinamento. Para o treinamento (ou aprendizagem), é ideal que a RNA processe um grande volume de dados cujas

classificações sejam previamente conhecidas, e que os dados sejam o mais próximo possível dos dados reais que serão classificados pela rede, processo que consiste em uma aprendizagem supervisionada. O processo de aprendizagem será melhor descrito com o auxílio do diagrama a seguir (Figura 3).

Figura 3. Loop para treinamento da RNA.



Fonte: autoria própria.

Inicialmente, a rede é inicializada com todos os seus pesos assumindo valores randômicos. A partir do processo de aprendizagem, os valores dos pesos são ajustados para melhor resolverem o problema em questão.

Assim, para o treinamento da rede, o primeiro passo consiste no processamento dos dados; isto é, a rede gera uma classificação dos dados baseada nos valores de peso que ela possui. Gerada a resposta da rede quanto à classificação, inicia-se o segundo passo do treinamento: a checagem. A rede compara suas respostas com as respostas corretas para que seja aplicado o terceiro passo, que é o de otimização. Nesse passo, são calculados os novos valores de peso da rede neural, visando melhorar a acurácia da classificação que foi feita durante o primeiro passo. Por fim, na última etapa do treinamento ocorre a atualização dos valores de peso da rede. Esses passos se repetem até que a rede seja considerada treinada.

Após o treinamento, a RNA está pronta para realizar as classificações. O processo de classificação consiste em realizar o processamento dos dados e, por meio da detecção de padrões, que é possibilitada devido ao aprendizado, a rede julga os sinais processados como sendo patológicos ou não. Nesse contexto, a rede pode ou não possuir o resultado esperado para as classificações que estão sendo realizadas, mas as atualizações de valores de peso da rede não mais são feitas, a não ser que uma nova fase de treinamento se inicie.

4. DESCRIÇÃO DA SOLUÇÃO

A solução proposta se baseia no desenvolvimento de uma Rede Neural Artificial (RNA), que será responsável por classificar sinais de voz como sendo patológicos ou não. Além disso, antes de serem enviados à RNA para a classificação, os sinais de voz passam por uma etapa de pré-processamento e extração de características, conforme descrito na Seção 3.2.

A RNA, o pré-processamento e a extração de características em questão foram desenvolvidos utilizando a linguagem de programação Python, com o auxílio das seguintes bibliotecas:

- *Keras* [8]: uma biblioteca *open-source* que provê uma interface em Python para desenvolver RNA. O Keras age como uma interface para a biblioteca *TensorFlow;*
- *Python speech features* [13]: provê funções comuns no desenvolvimento de software de reconhecimento automático da voz, tal qual o cálculo dos MFCC;
- *Numpy* [12]: uma biblioteca que adiciona suporte para o processamento de grandes volumes de dados, que podem vir em forma de *arrays* ou matrizes multidimensionais. Além disso, a *Numpy* também oferece uma grande coleção de funções matemáticas de alto nível para atuar nas estruturas de dados previamente mencionadas;
- *Librosa* [9]: um pacote que provê funções necessárias no desenvolvimento de software para análise de música e áudio em Python; e
- *Scipy* [14]: uma biblioteca *open-source* usada para fazer computações científicas e técnicas. A *Scipy* contém módulos para otimização, álgebra linear, FFT, processamento de sinais, dentre outros.

4.1 Pré-processamento e Extração de Características

Para o carregamento e leitura dos sinais de voz pelo sistema, foi utilizada a biblioteca *Librosa*. Uma vez lidos, os sinais de voz passam pelas fases de pré-processamento e extração dos coeficientes MFCC. Para esse fim, foi utilizada a biblioteca *python speech features*.

A biblioteca *python speech features* provê um método para extração dos MFCC a partir de um arquivo de áudio. Tal método permite a definição de diversos parâmetros pertinentes ao pré-processamento e cálculo dos MFCC. Os parâmetros utilizados na solução foram os seguintes:

- Taxa de Amostragem: 50.000 amostras/s;
- Superposição: 50%;
- Tamanho do vetor de características por *frame*: 13 coeficientes Mel-Cepstrais (MFCC);
- Pré-ênfase: 0,95;
- Janelamento: 32 ms; e
- Função Janela: Hamming.

Uma vez extraídas as características, os dados estariam prontos para serem processados pela rede neural. Porém, para fins de comparação, foi também testada a abordagem de não considerar a etapa de extração dos MFCC, isto é, enviando à rede neural os dados pré-processados, ao invés de seus MFCC.

4.2 Rede Neural Artificial

Para a implementação da Rede Neural Artificial (RNA), foi utilizada a biblioteca *Keras*.

Foram realizados testes empíricos quanto ao número de camadas ocultas e ao número de nós em cada camada, de forma a encontrar configurações mais promissoras da rede perante o problema de classificação em questão. Os melhores resultados

foram obtidos com o uso de apenas uma camada oculta com 50 nós, conforme será reportado na seção seguinte.

Para a função de ativação da camada oculta, foi utilizada a função unidade linear retificada, também conhecida como ReLU (*Rectified Linear Unit*). Quanto à função de ativação da camada de saída, foi utilizada a função sigmóide.

5. METODOLOGIA

Nesta seção, será descrita a metodologia envolvida no treinamento e validação da solução, bem como qual a base de dados utilizada para esses fins.

5.1 Base de Dados

A base de dados utilizada no presente trabalho foi a *Saarbruecken Voice Database* (SVD), uma base de dados desenvolvida na Alemanha e distribuída a partir de um repositório digital aberto. Apesar da SVD possuir sinais de vozes gravados e sinais de eletroglotografia capturados em laboratório, para os fins deste trabalho, apenas sinais de vozes gravados foram utilizados[1].

Em sua totalidade, a SVD possui 2.043 sinais de vozes, dentre as quais tem-se 687 saudáveis e 1.356 patológicos; 57% dos dados pertencem a indivíduos do sexo feminino. Os sinais patológicos são distribuídos em diversas categorias de desordens vocais.

Para o desenvolvimento deste trabalho, foram selecionados 700 sinais de voz cujas gravações contemplam indivíduos emitindo a vogal sustentada /a/. Na SVD existe um outro conjunto de sinais de voz que contemplam indivíduos recitando frases pré-definidas, tais como "*Guten Morgen, wie geht es Ihnen?*" ("Bom dia, Como está você?"); esse subconjunto, contudo, não foi utilizado no presente trabalho.

Dentre os 700 sinais selecionados, 300 desses são sinais de voz saudáveis; destes, 180 sinais pertencem a indivíduos do sexo feminino, e 120 pertencem a indivíduos do sexo masculino. Os 400 sinais de voz restantes são da categoria dos patológicos; 220 desses pertencendo a indivíduos do sexo feminino e 180 pertencentes a indivíduos do sexo masculino. As patologias vocais presentes no subconjunto dos dados selecionados para esse trabalho incluem: paralisia das pregas vocais, carcinoma na laringe, leucoplasia das pregas vocais, pólipos nas pregas vocais e edema de Reinke.

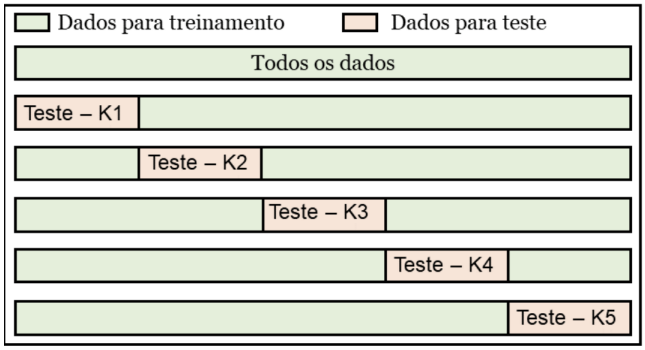
A taxa de amostragem original, de 50.000 amostras por segundo, foi mantida para os fins deste trabalho. Todas as gravações possuem um tempo de duração entre 1 e 3 segundos e possuem formato '.wav'.

5.2 Treinamento e Validação

Para o treinamento e validação da solução desenvolvida, foi utilizado o método de validação cruzada (*Cross-Validation) k-fold* [3]. Neste método, são geradas k partições aleatórias a partir dos dados disponíveis. Para cada uma das partições geradas, o classificador será validado fazendo uso da partição em questão; o restante dos dados é, então, utilizado para o treinamento do classificador. Esse processo é realizado k vezes e, ao final da execução, são calculados a média e o desvio padrão das acurácias obtidas em cada uma das iterações.

Na Figura 4 é ilustrada a forma como é realizado o particionamento dos dados para um valor de k = 5.

Figura 4. Exemplo de partição de dados *k-fold*.



Fonte: baseado em [5].

Em outras palavras, a rede é efetivamente treinada e testada k vezes. Em cada um desses testes, um conjunto diferente de dados é utilizado para o aprendizado da rede e o restante dos dados é usado na validação.

Para a aplicação da técnica de validação cruzada, o valor de k = 10 foi utilizado no presente trabalho, pois esse valor foi o mais comumente observado em trabalhos de aprendizagem de máquina encontrados na literatura [5].

6. RESULTADOS E DISCUSSÕES

Na Tabela 1 é apresentada a acurácia de algumas configurações que foram testadas para a Rede Neural Artificial (RNA), perante o método de validação cruzada *k-fold* com k = 10.

Tabela 1. Acurácia para diferentes configurações da RNA na classificação de duas classes (Patológica *versus* Saudável).

	Número de camadas ocultas		
Nós em cada camada oculta	1	4	6
10	82,00%	82,43%	82,57%
50	**85,71%**	83,57%	83,29%
200	83,57%	84,43%	83,29%

Fonte: autoria própria.

No contexto de análise da acurácia do classificador, tem-se um verdadeiro positivo quando um sinal de voz possui alguma patologia e a RNA classifica o sinal como patológico; um falso positivo ocorre quando o sinal de voz é saudável e a rede o classifica como patológico. Por outro lado, um verdadeiro negativo acontece quando a RNA classifica corretamente um sinal de voz saudável e o falso negativo se dá quando a rede neural falha em classificar o sinal de voz saudável.

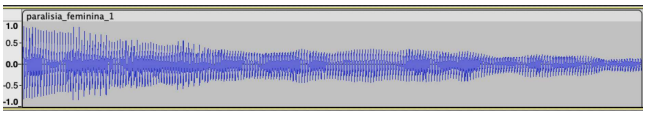
Conforme reportado na Tabela 1, a melhor configuração encontrada para o classificador do problema em questão foi com

[1] https://github.com/Matheus7OP/voice-pathology-detection

apenas uma camada oculta contendo 50 nós. A configuração em questão proporcionou os seguintes resultados:

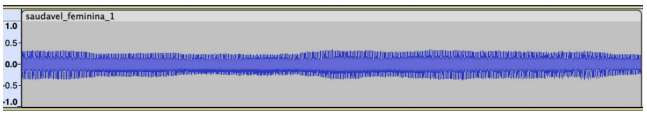
- Média das taxas de acerto dentre os casos positivos (em que há patologia): 90,36%;
- Média das taxas de acerto dentre os casos negativos (em que não há patologia): 80,49%;
- Média das taxas de acurácia dentre todos os casos: 85,71%; e
- Desvio padrão das taxas de acurácia: 5,26%.

Figura 5. Exemplo de voz patológica da base de dados SVD, visualizada no software *Audacity*.



Fonte: autoria própria.

Figura 6. Exemplo de voz saudável da base de dados SVD, visualizada no software *Audacity*.



Fonte: autoria própria.

Outro resultado a ser destacado diz respeito à abordagem que consiste em utilizar os próprios sinais de voz pré-processados no momento de inserir os dados a serem classificados na rede ao invés de inserir os seus MFCC. Utilizando essa abordagem, o melhor resultado de acurácia obtido durante os testes foi de 57,14%. Esse resultado foi obtido com uma RNA de 4 camadas ocultas, cada uma possuindo 50 nós.

A disparidade nos resultados entre as abordagens mencionadas se dá, em especial, pelo aumento da complexidade do problema ao ser removida a etapa de extração de características. Quando essa etapa é realizada, o que ocorre na prática é que o espaço de busca é efetivamente reduzido, facilitando a detecção de padrões por parte do classificador.

Uma possível forma de melhorar os resultados da abordagem que não faz a extração de características consiste no aumento do número de camadas ocultas e de nós em cada camada oculta, aumentando a complexidade e a profundidade da rede. Dessa forma, a rede estaria mais apta a encontrar padrões mais complexos, cujo espaço de busca é maior, quando comparado à detecção de padrões por meio dos MFCC.

Por fim, uma forma de melhorar expressivamente os resultados de ambas as abordagens consistiria no uso de uma base de dados maior. Dado o contexto do problema, há dificuldade em se encontrar grandes bases de dados públicas com boa qualidade; o subconjunto da base utilizada, apesar de possuir 700 amostras, ainda se mostra insuficiente para o aproveitamento de todo o potencial de aprendizado de uma RNA profunda. Bases de dados com pelo menos dezenas de milhares de amostras seriam mais adequadas para um melhor treinamento da rede neural, propiciando melhores resultados durante a classificação.

7. CONCLUSÃO

Neste trabalho, foi apresentada uma opção de classificador cuja acurácia supera 85% para a detecção de patologias vocais.

O aperfeiçoamento e a maior visibilidade da análise acústica aplicada à detecção de patologias da voz possui grande potencial para impactar positivamente o futuro dos diagnósticos das patologias vocais. Isso se deve em especial à análise acústica não ser invasiva, possuir grande capacidade de representação de determinadas patologias e ser menos onerosa do que os exames laboratoriais usuais. Tais fatores podem beneficiar grandemente futuros pacientes que necessitem de exames dessa natureza.

Os resultados apresentados neste trabalho servem como mais uma evidência de que o método de análise acústica se mostra eficiente e promissor, merecendo ser mais pesquisado e mais aplicado no que diz respeito ao auxílio no diagnóstico de patologias vocais.

7.1 Limitações

A base de dados utilizada para o treinamento e validação da solução proposta possui tamanho bem menor que o ideal. Idealmente, a base de dados utilizada teria ao menos algumas dezenas de milhares de amostras, para que a validação do classificador pudesse ser mais robusta.

7.2 Trabalhos futuros

Futuramente, pode-se aumentar o tamanho da base de dados a ser utilizada no treinamento e na validação do classificador em questão. Tal fator proporcionará processos de aprendizado e de validação mais robustos para a rede neural artificial, o que é desejado.

8. AGRADECIMENTOS

Agradeço aos meus amigos e familiares, por me acompanharem durante todos os altos e baixos na jornada da graduação.

Agradeço ao professor Rohit Gheyi, por toda a motivação e apoio que me foram dados desde antes da minha entrada na UFCG. Minha experiência na graduação foi melhorada exponencialmente graças ao senhor.

Agradeço à professora Joseana Macêdo Fechine Régis de Araújo, por toda a paciência e comprometimento durante a orientação do presente trabalho.

9. REFERÊNCIAS

[1] Almeida, N. C. de, "Sistema Inteligente para Diagnóstico de Patologias na Laringe utilizando Máquinas de Vetor de Suporte". Dissertação - Universidade Federal do Rio Grande do Norte, 2010.

[2] Alves, N. F. R., "Diagnóstico Inteligente de Patologias da Laringe". Dissertação - Instituto Politécnico de Bragança, 2016.

[3] C. Manfredi, "Adaptive noise energy estimation in pathological speech signals," in IEEE Transactions on

Biomedical Engineering, vol. 47, no. 11, pp. 1538-1543, Nov. 2000, doi: 10.1109/10.880107.

[4] COSTA, W. C. de A. et al. "Análise dinâmica do sistema de produção vocal na presença de patologias da laringe". Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, v. 1, n. 1, p. 1–6, 2013

[5] Dias, L. C., "Detecção de patologias laríngeas por meio da análise de sinais de voz utilizando *Deep Neural Networks*". Dissertação - Instituto Federal de Educação da Paraíba, 2020.

[6] Fortes, F. S. G., Imamura, R., Tsuji, D. H., Sennes, L. U., "Perfil dos profissionais da voz com queixas vocais atendidos em um centro terciário de saúde". Revista Brasileira de Otorrinolaringologia, 2007.

[7] Guedes, V. de O., "Deep Learning aplicado à classificação de patologias da voz". Dissertação - Instituto Politécnico de Bragança, 2019.

[8] Keras: the Python deep learning API. Disponível em: <https://keras.io/>. Acesso em: 10 de fev. de 2022.

[9] Librosa documentation. Disponível em <https://librosa.org/doc/latest/index.html>. Acesso em: 10 de fev. de 2022.

[10] Marinus, J. V. de M. L., "Estudo de Técnicas para Classificação de Vozes Afetadas por Patologias". Dissertação - Universidade Federal de Campina Grande, 2010.

[11] Muda, L., Begam, M., Elamvazuthi, I., "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques". Journal of Computing, ISSN 2151-9617, 2010.

[12] Numpy: The fundamental package for scientific computing with Python. Disponível em <https://numpy.org/>. Acesso em: 10 de fev. 2022.

[13] Python speech features documentation. Disponível em <https://python-speech-features.readthedocs.io/en/latest/>. Acesso em: 10 de fev. de 2022.

[14] Scipy: Fundamental algorithms for scientific computing in Python. Disponível em <https://scipy.org/>. Acesso em: 10 de fev. de 2022.

[15] Teixeira, J. P., Fernandes, P. O., "Acoustic Analysis of Vocal Dysphonia". Procedia Computer Science, ISSN 1877-0509, 2015.

[16] Tiwari, M., Tiwari, M., "Voice - How humans communicate?". Journal of Natural Science, Biology and Medicine 3, 3–11 10.4103/0976-9668.95933, (2012).

Sobre os autores:

Matheus Oliveira Pereira. Graduando em Ciência da Computação e entusiasta de competições de programação. Em 2018, obteve o primeiro lugar na Olimpíada Brasileira de Informática. Recentemente, participou do estágio de inverno da VTEX e atualmente é engenheiro de software na Griaule Biometrics, em Campinas/SP.

Joseana Macêdo Fechine Régis de Araújo. Professora orientadora do TCC.