REPRODUÇÃO DE ARTIGO COMO PROJETO FINAL PARA A DISCIPLINA DE FUNDAMENTOS DE PESQUISA PARA CIÊNCIA DA COMPUTAÇÃO 2 DO SEMESTRE 2022.1 DO MESTRADO EM COMPUTAÇÃO DA UFCG

Professor: Fábio Jorge Almeida Morais
Aluno: Gabriel Almeida Azevedo
Orientadora: Joseana Macêdo Fechine Régis de Araújo
Matrícula: 0122015804-7M
Data: 02/07/2022

1. Introdução

A voz é um dos meios mais importantes de comunicação do ser humano, pois através dela pode-se transmitir uma mensagem [1]. A emissão da voz é um processo complexo que envolve a coordenação dos sistemas: respiratório, fonatório, ressonante, articulatório e nervoso. O sistema fonatório é o mais essencial deles, pois tem a função de gerar o som. Seu principal órgão é a laringe, que possui um mecanismo (comumente chamado de pregas ou cordas vocais) responsável por regular a passagem de ar vindo dos pulmões. Quando as pregas vocais se aproximam, o fluxo de ar as faz vibrar, gerando assim, o som [2].

Como toda parte do corpo humano, o sistema fonatório pode ser acometido por doenças que são comumente chamadas de patologias da voz. Para detectar e classificar os distúrbios da voz, os médicos elaboram seus diagnósticos baseados em exames instrumentais, como eletroglotografia e videolaringoscopia (buscando explorar as estruturas laríngeas), análise perceptiva do discurso do paciente, baseado na exploração de sintomas e na exploração do histórico de saúde [3]. Essa atividade por vezes é demorada e requer *expertise* do médico para propor boas hipóteses e testá-las. Além disso, alguns dos exames são invasivos, gerando desconforto ao paciente.

Nesse sentido, o estudo centrado na criação de sistemas que permitem detecção de patologias da voz, por meio da análise do som emitido pelo paciente, se apresenta como alternativa forte para apoio ao diagnóstico médico, por ser mostrar um método não invasivo, rápido e barato, atraindo assim a atenção de profissionais da saúde.

Dessa forma o presente trabalho trata sobre a reprodução de um dos experimentos do artigo *Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database* [4] a fim de entender como a área da ciência da computação pode auxiliar a área da saúde para acelerar e apoiar o diagnóstico de um paciente.

1.1. Breve resumo sobre o artigo

O objetivo principal do artigo é desenvolver um sistema de detecção de vozes com patologias, focando em métodos de aprendizagem profunda, mais precisamente redes neurais *Feedforward* (FNN) e convolucionais (CNN), utilizando coeficientes Mel-Cepstrais (MFCC), Delta MFCC, coeficientes cepstrais obtidos a partir da análise por predição linear (LPCC) e parâmetros estatísticos de alta ordem (HOS).

Os dados utilizados foram extraídos da base SVD [5] (Saarbruecken Voice Database). A duração das gravações varia de 1 a 4 segundos. O áudio tem formato de onda

(amostra de 16 bits), com 50 kHz. Ao todo, foram 518 registros de áudios femininos falando os fonemas /a/, /i/ e /u/ em tom normal, sendo esses compostos de 259 vozes saudáveis e 259 vozes com alguma patologia do trato vocal. A mesma quantidade de separação foi adotada para vozes masculinas. Experimentos foram executados variando os parâmetros de entrada da rede neural, o tipo da rede (FNN ou CNN), o gênero dos locutores e ainda os fonemas avaliados.

Os resultados obtidos demonstraram que, combinando os parâmetros avaliados e utilizando redes de aprendizagem profunda, é possível obter um classificador capaz de distinguir vozes normais de vozes patológicas, com 80% de acurácia. Ademais, os experimentos mostraram que a análise das vozes de forma isolada (masculinas ou femininas) proporcionou melhor resultado.

2. Metodologia

Apenas um dos experimentos realizados no artigo foi escolhido para ser reproduzido. O experimento em questão consiste em utilizar áudios de falantes masculinos pronunciando o fonema /a/ e uma rede neural do tipo *Feedforward* (rede sem ciclos ou *loops*) cujos parâmetros de entrada são os coeficientes MFCC e Delta MFCC extraídos dos áudios. Este experimento atingiu acurácia de 80.13% com erro de ± 2.65%. Essa seção irá descrever como o experimento foi realizado no artigo original e como foi reproduzido.

2.1. Metodologia do trabalho original

2.1.1. Áudios e Pré Processamento

Embora o autor indique a quantidade de áudios que compõem sua base para cada gênero e tipo de doença, não sabe-se quais áudios da base foram escolhidos. A tabela 1 demonstra a separação utilizada.

Tabela 1: Quantidade de áudios utilizada no artigo original e suas classificações.

Diagnosis of Pathological Voices	Number of Samples	
	Female	Male
Hyperfunctional dysphonia	97	44
Functional dysphonia	51	33
Laryngitis	30	62
Vocal fold polyp	19	25
Leukoplakia	14	27
Vocal fold cancer	1	21
Vocal nodule	13	4
Reinke edema	27	7

Diagnosis of Pathological Voices	Number of Samples	
	Female	**Male**
Hypofunctional dysphonia	4	10
Granuloma	1	1
GERD	0	3
Contact ulcers	2	22
Subtotal	259	259
Health voices	259	259
Total	518	518

Fonte: Artigo reproduzido [4].

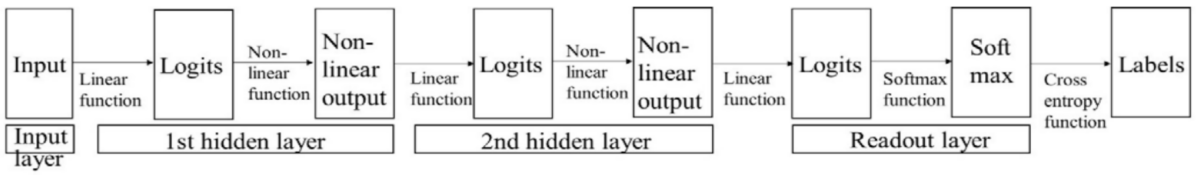
Para o experimento específico que foi reproduzido, o autor afirma ter utilizado áudios de vozes masculinas falando o fonema /a/ na seguinte divisão: 181 (70%) áudios para treinamento e 78 para teste (30%). Cada conjunto foi selecionado aleatoriamente para um esquema de validação cruzada de *5-Fold*. Todo o processo foi repetido 10 vezes, e os resultados foram calculados. Os resultados são apresentados como médias e desvios padrão.

Para a etapa de pré-processamento os coeficientes MFCC e Delta MFCC de 20 dimensões foram extraídos de um sinal de janela de 40 ms usando um deslocamento de quadro de 20 ms.

2.1.2. Rede Neural *Feed-Forward*

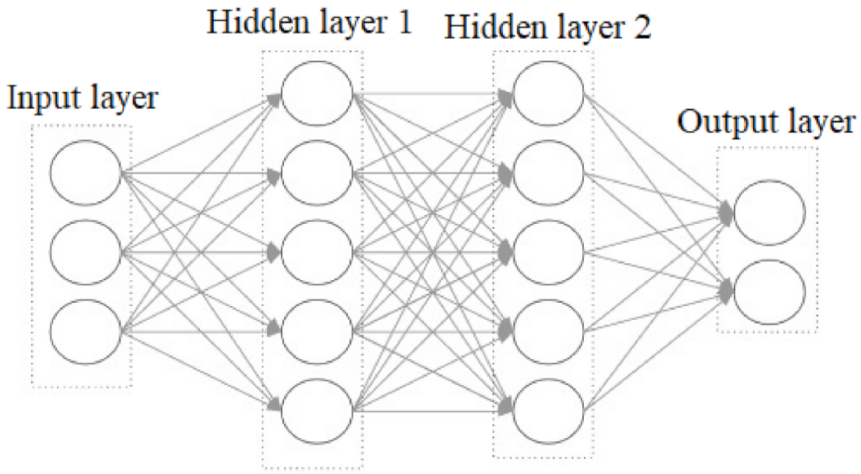
O autor utilizou uma rede neural do tipo *feedforward* totalmente conectada (cada nó de uma camada está conectado com todos os outros nós da próxima camada) com duas camadas ocultas. A rede neural foi treinada em 100 épocas e a saída da rede foi codificada para ser 0 caso a voz seja classificada como patológica e 1 caso seja classificada como voz saudável. As figuras 2 e 3 demonstram a rede criada e a tabela 2 demonstra os parâmetros utilizados para construir a FNN.

Figura 2: FNN em detalhes.



Fonte: Artigo reproduzido [4].

Figura 3: FNN totalmente conectada.



Fonte: Artigo reproduzido [4].

Tabela 2: Parâmetros da FNN.

Parameter	Value
Loss function	Tanh
Optimization algorithm	SGD + Momentum
Regularization	L2
Mismatch propagation	BPTT
Minibatch size	100 samples
Learning rate	0.01 exponential attenuation
Loss function	Cross-entropy
Weights for samples	Yes

Fonte: Adaptado de [4].

2.2. Metodologia da reprodução

2.2.1. Áudios e Pré Processamento

Os áudios foram extraídos da mesma base SVD porém em proporções diferentes e contemplando menos patologias do que as utilizadas no artigo original. Ao todo foram 240 áudios de vozes masculinas pronunciando o fonema /a/ em tom normal. A tabela 3 descreve a distribuição dos áudios.

Tabela 3: Quantidade de áudios utilizada na reprodução e suas classificações.

Diagnosis of Pathological Voices	Number of Samples
	Male
Carcinoma	20
Reinke edema	7
Laryngitis	24
Leukoplakia	23
Vocal fold polyp	23
Bulba Paralyse	23
Subtotal	120
Health voices	120
Total	240

Fonte: Própria.

Para a etapa de pré-processamento os coeficientes MFCC e Delta MFCC de 20 dimensões foram extraídos de sinais de áudio com janela de 40 ms usando um deslocamento de quadro de 20 ms e pré-ênfase de 0.97.

As bibliotecas utilizadas foram librosa, para carregar os áudios e *python speech features* para extração dos MFCC e Delta MFCC.

2.2.2. Rede Neural *Feed-Forward*

A biblioteca Keras foi utilizada para construir o modelo sequencial com otimização SGD com *momentum* de 0.7 e *learning rate* de 0.01, *Binary Cross Entropy* como *loss function* e 4 camadas. As métricas avaliadas foram acurácia, precisão, *recall* e *F1-score.* A rede foi treinada em 100 épocas. Foi aplicado uma validação cruzada do tipo *10-Fold.* Os resultados são apresentados como médias e desvios padrão de todos os *fold*. A tabela 4 descreve cada uma das camadas da FNN construída.

Tabela 4: Descrição das camadas da FNN implementada.

Layer	Type	Neurons	Activation	Regularization
Input Layer	Input	40	-	-
Hidden Layer 1	Dense	40	Relu	L2
Hidden Layer 2	Dense	15	Relu	L2

Layer	Type	Neurons	Activation	Regularization
Readout Layer	Dense	2	Softmax	-

Fonte: Própria.

3. Resultados da Reprodução

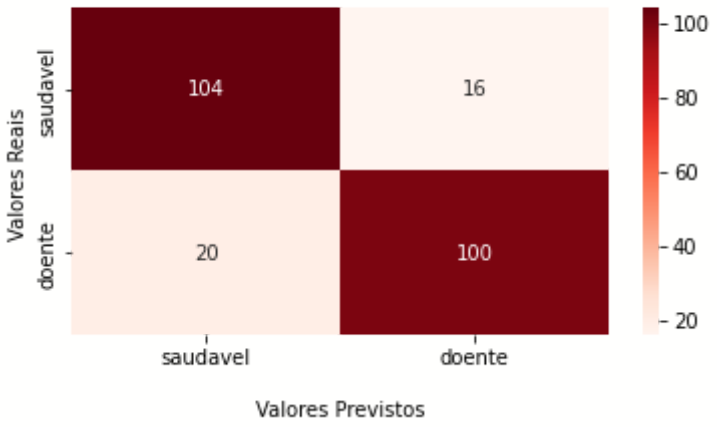
O experimento realizado com uma validação cruzada *10-Fold* atingiu acurácia de 85.00% ± 4.25%, precisão de 86.72% ± 3.82%, recall de 85.33% ± 4.74% e *F1-score* de 84.60% ± 4.63%.

Ao comparar os intervalos obtidos para a acurácia na reprodução ([80.75% : 89.25%]) com o obtido pelo artigo original ([77.48% : 82.78%]) percebe-se que o modelo gerado na reprodução apresenta indícios de ter superado o original. Vale salientar que os áudios usados não foram exatamente os mesmos e a quantidade também não foi a mesma (259 áudios no original e 240 na reprodução). Também não se sabe o coeficiente de *momentum* que o autor utilizou. Ademais, o autor menciona uma validação cruzada *5-Fold* com repetição de 10 vezes, como esta parte não ficou clara, na reprodução, foi utilizada uma validação cruzada *10-Fold*. Também não está claro como ele separou 70% dos dados para treino e 30% para teste se ele aplicou validação cruzada.

A figura 4 apresenta a matriz de confusão da FNN gerada para os dados de teste. A partir da matriz percebe-se que o modelo classifica corretamente uma voz 85% das vezes. Para o problema em questão o índice de falsos negativos é bastante relevante e deve ser reduzido ao máximo. O valor atingido pelo classificador para este índice foi de 8.3%, sendo ainda um valor significativo.

O notebook com o código utilizado na reprodução está disponível para leitura em: notebook.

Figura 4: Matriz de Confusão da FNN implementada.



Fonte: Própria.

4. Conclusões

A partir dos resultados obtidos conclui-se que o artigo original apresenta insumo suficiente, porém não completo, para uma reprodução e que, a metodologia adotada, MFCC Ge Delta MFCC com o classificador FNN, realmente é capaz de gerar resultados satisfatórios mesmo quando a quantidade de dados para treinamento não é elevada.

Essa reprodução foi utilizada como forma de validar a base de dados SVD para ser aplicada no trabalho de mestrado e, a fim de adquirir mais domínio e embasamento (teórico e prático) sobre redes neurais profundas, facilitando o estudo de mestrado a ser desenvolvido. Estes objetivos foram alcançados.

5. Referências

[1] Kadiri, Sudarsana Reddy and Paavo Alku. Analysis and Detection of Pathological Voice Using Glottal Source Features. IEEE Journal of Selected Topics in Signal Processing, 2020, 367-379. Disponível em https://www.semanticscholar.org/paper/Analysis-and-Detection-of-Pathological-Voice-Using-Kadiri-Alku/5448bb79a7f4fc8aad3f1674b1a0bd86a81eb174. Último acesso em 15 de abril de 2022.

[2] J.A. Gómez-García. Contributions to the design of automatic voice quality analysis systems using speech technologies. Thesis (Doctoral), E.T.S.I. Telecomunicación (UPM), 2018. Disponível em https://oa.upm.es/49565/. Último acesso em 15 de abril de 2022.

[3] J.A. Gómez-García, L. Moro-Velázquez, J.I. Godino-Llorente, On the design of automatic voice condition analysis systems. Part I: Review of concepts and an insight to the state of the art, 2019. Disponível em https://www.sciencedirect.com/science/article/abs/pii/S1746809418303239. Último acesso em 14 de abril de 2022.

[4] Lee, Ji-Yeoun. 2021. Experimental Evaluation of Deep Learning Methods for an Intelligent Pathological Voice Detection System Using the Saarbruecken Voice Database. Applied Sciences 11, no. 15: 7149. Disponível em https://doi.org/10.3390/app11157149. Último acesso em 15 de maio de 2022.

[5] M. Putzer, W. Barry, “Saarbrucken Voice Database 2.0”, Institute of Phonetics, Univ. of Saarland. Disponível em: http://www.stimmdatenbank.coli.uni-saarland.de/help_en.php4. Último acesso em 13 de junho de 2022.