Comparação das técnicas MFCC, PNCC e ZCPA na identificação de patologias relacionadas à voz, usando Redes Neurais Artificiais

Vinícius F. Cardoso, Edson Cataldo e Leonardo A. Forero M.

***Resumo*— Este estudo analisou a complexidade da voz humana e os desafios no pré-diagnóstico de condições vocais. Foram utilizadas técnicas como MFCC, PNCC e ZCPA, junto com IA avançada, para classificar a voz humana e suas patologias. Diversas arquiteturas de redes neurais — DNN, CNN, LSTM e BiLSTM — foram usadas para diferenciar vozes saudáveis e afetadas por patologias (nódulo, paralisia, edema de Reinke e cisto) sob diferentes condições. O método MFCC demonstrou alta eficácia com 99% de precisão. PNCC detecta muitos casos patológicos, mas com mais falsos negativos e positivos. ZCPA mostrou menor consistência, sugerindo refinamento e testes adicionais.**

***Palavras-Chave*— Inteligência artificial, Saúde vocal, Identificação de patologias vocais.**

***Abstract*— This study analysed the complexity of the human voice and the challenges in the pre-diagnosis of vocal conditions. Techniques such as MFCC, PNCC, and ZCPA, along with advanced AI, were used to classify the human voice and its pathologies. Various neural network architectures—DNN, CNN, LSTM, and BiLSTM—were employed to differentiate between healthy voices and those affected by pathologies (nodule, paralysis, Reinke's edema, and cyst) under different conditions. The MFCC method demonstrated high efficacy with 99% accuracy. PNCC detected many pathological cases but, had more false negatives and positives. ZCPA showed less consistency, suggesting the need for refinement and further testing.**

***Keywords*— Artificial intelligence, Vocal health, Identification of vocal pathologies.**

I. Introdução

A voz humana desempenha um papel fundamental na sociedade, sendo um instrumento vital de comunicação e expressão pessoal. Na comunicação diária, a voz não apenas transmite informações, mas também emoções, intenções e características pessoais, tornando-se uma ferramenta poderosa para a interação social [1]. No entanto, o uso extensivo da voz em certas profissões pode levar a problemas de saúde vocal, afetando não apenas a eficácia profissional, mas também a qualidade de vida [2]. Os avanços da inteligência artificial (IA) têm impactado diversos setores, incluindo a saúde. Pesquisas buscam melhorar diagnósticos, apoiando profissionais de saúde. O artigo investiga o uso de IA e redes neurais para classificar patologias vocais e vozes saudáveis. Características vocais são extraídas por meio dos métodos: *Mel-Frequency Cepstral Coefficients* (MFCC), *Power-Normalized Cepstral Coefficients* (PNCC) e *Zero-Crossings with Peak Amplitudes* (ZCPA); e diferentes redes neurais, como *Deep Neural Networks* (DNN), *Convolutional Neural Networks* (CNN), *Long Short-Term Memory* (LSTM) e *Bidirectional Long Short-Term Memory* (BiLSTM), são avaliadas.

A literatura sugere que, em sistemas de reconhecimento de voz que utilizam pronúncia de palavras e presença de ruídos, o método ZCPA frequentemente supera o MFCC [3][4][5]. Além disso, o método PNCC também é considerado mais robusto que o MFCC [6]. No entanto, em ambientes sem ruído, o MFCC tende a ser mais eficaz que o ZCPA. Quanto à comparação entre PNCC e MFCC nesses contextos, eles apresentam desempenho bastante similar. Em estudos voltados para a identificação de patologias nas pregas vocais, observa-se a eficácia do método MFCC [7][8][9][10].

A pesquisa emprega os métodos mencionados para identificar características de vozes que apresentam desordens provocadas por patologias, tais como nódulo, paralisia, edema de Reinke e cisto, além de vozes saudáveis. Essa análise foi conduzida utilizando fragmentos de vozes femininas e masculinas, com durações de 0,5 e 0,7 segundos, provenientes da emissão sustentada das vogais /a/ e /e/. Os testes foram realizados em vozes sem ruído e com adição de ruído branco.

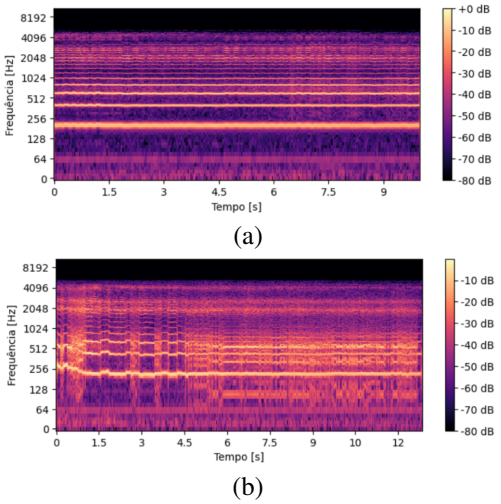
A. Patologias das cordas vocais

As patologias das cordas vocais referem-se a uma variedade de condições que afetam a saúde e o funcionamento das cordas vocais. Normalmente, as cordas vocais vibram de maneira uniforme e simétrica, criando ondas sonoras regulares que produzem uma voz clara e modulada. No entanto, os distúrbios das cordas vocais perturbam esse processo. As patologias podem limitar a capacidade das cordas vocais de vibrar em certas frequências, resultando em uma redução da amplitude tonal. Na Figura 1, apresentamos o espectrograma de uma voz saudável (a) e uma voz com paralisia (b).

II. Fundamentação Teórica

A extração de características da voz transforma sinais de áudio em representações numéricas. Essas características, originadas das propriedades acústicas da fala, são fundamentais em aplicações como reconhecimento de fala, identificação de locutor, detecção de emoções e alterações vocais. Nas

Vinícius Flores Cardoso, Universidade Federal Fluminense (UFF), PPGEET (Prog de pós-graduação em Eng. Elétrica e de Telecomunicações), Niterói-RJ, e-mail: vinicflores@gmail.com; Edson Luiz Cataldo Ferreira, Universidade Federal Fluminense (UFF), PPGEET, Niterói-RJ, e-mail: ecataldo@id.uff.br. Leonardo Alfredo Forero Mendoza, Universidade do Estado do Rio de Janeiro (UERJ), Depart. de Engenharia Elétrica, Rio de Janeiro-RJ, e-mail: leofome@hotmail.com



(a)

(b)

Fig. 1. Espectrograma de voz saudável (a) e com paralisia (b)

subseções A, B e C, são descritos os métodos de obtenção de características vocais MFCC, PNCC e ZCPA.

A. *Mel-Frequency Cepstral Coefficients - MFCC*

O método MFCC simula a forma como o sistema auditivo humano percebe diferentes frequências, por meio da escala Mel. A escala Mel é um componente fundamental no cálculo do MFCC e é usada para mapear as frequências lineares (geralmente em Hertz) em frequências Mel.

A fórmula para converter uma frequência f (em Hertz) para a frequência *m* (em Mel) e a sua inversa são, respectivamente:

$$m = M(f) = 1127 \times ln\left(1 + \frac{f}{700}\right) \tag{1}$$

$$f = M^{-1}(m) = 700\left(e^{\frac{m}{1127}} - 1\right) \tag{2}$$

Os filtros Mel Cepstrais são aplicados à transformada de Fourier para capturar as características perceptualmente relevantes da frequência. A criação dos filtros Mel é feita a partir dos pontos Mel usando funções triangulares [6], em que cada filtro é centrado em um ponto Mel e possui uma forma triangular. Os coeficientes dos filtros Mel são calculados com base nas equações:

$$H_m(k) = \begin{cases} 0, & \text{se } k < k_{m-1} \\ \frac{k - k_{m-1}}{k_m - k_{m-1}}, & \text{se } k_{m-1} \leq k < k_m \\ \frac{k_{m+1} - k}{k_{m+1} - k_m}, & \text{se } k_m \leq k < k_{m+1} \\ 0, & \text{se } k \geq k_{m+1} \end{cases} \tag{3}$$

em que k é a frequência em Mel; $k_{(m-1)}$ e $k_{(m+1)}$ são as frequências dos filtros vizinhos e $k(m)$ é a frequência central do filtro triangular.

Após aplicar o banco de filtros Mel, os coeficientes Mel são obtidos, representando a energia em cada faixa de frequência de maneira não linear. Posteriormente, os coeficientes Mel Cepstrais são submetidos ao logaritmo, seguido pela aplicação da Transformada Cosseno Discreta (DCT) para obter os coeficientes MFCC, e os dados são normalizados.

B. *Power-Normalized Cepstral Coefficients - PNCC*

Os PNCCs são uma extensão dos MFCCs e têm como objetivo principal capturar características perceptivas do áudio, semelhantes à forma como o ouvido humano processa o som, tendo a vantagem de ter mais robustez em ambientes ruidosos [11]. Os filtros gammatone são uma forma de modelar a resposta em frequência do ouvido humano, sendo baseados na escala de Bandas Retangulares Equivalentes (ERB) [12], que representam bem a resposta impulsional da membrana basilar. Eles foram projetados para simular a forma como a cóclea, a parte do ouvido que traduz o som em sinais nervosos, processa os sinais de áudio.

Em termos de processamento de sinais, um filtro gammatone é um filtro passa-banda, que pode ser caracterizado por sua função de transferência ou sua resposta ao impulso. Podendo ser representada como:

$$g(t) = a \cdot t^{n-1} \cdot e^{-2\pi bt} \cdot cos(2\pi f_c t + \phi) \qquad \text{para } t \geq 0 \tag{4}$$

em que a é a amplitude; t é o tempo; n é a ordem do filtro; b é a largura de banda retangular equivalente ERB; f_c é a frequência central do filtro; e ϕ é a fase inicial do filtro.

A escala ERB em função da frequência é:

$$ERB(f) = 24,7 + (4,32\frac{f}{1000} + 1)Hz \tag{5}$$

em que $ERB(f)$ é a largura de banda retangular equivalente em Hertz.

Após cada frame do sinal ser passado por meio de um banco de filtros gammatone, o resultado é uma representação do sinal de fala em termos de intensidade em várias bandas de frequência crítica. Em seguida é aplicada uma função de potenciação (operação não linear), facilitando a diferenciação entre componentes de sinal que são importantes para a percepção e aqueles que não são, como ruídos de fundo ou componentes irrelevantes. Por fim, é aplicada DCT e normalização dos dados.

C. *Zero-Crossings with Peak Amplitudes - ZCPA*

O método ZCPA é uma técnica de extração de características em processamento de sinais de voz, com uma abordagem robusta na análise de sinais acústicos, especialmente no reconhecimento de voz e de locutores em ambientes ruidosos. Essa técnica identifica e analisa dois componentes principais de um sinal de voz: a taxa de cruzamento por zero (*Zero Crossing Rate* - ZCR) e a amplitude de pico (*Peak Amplitude*).

O sinal passa por filtros, resultando em um histograma de frequência baseado nos pontos de cruzamento de zero. O acréscimo nesse histograma é determinado pelo logaritmo da maior intensidade registrada em cada intervalo. Finalmente, sobre esse histograma, aplica-se a DCT e normalização.

Inspirado pela forma como o ouvido humano processa sons, esse método simula a capacidade da cóclea, uma parte essencial do ouvido interno, de decompor sons complexos em componentes de frequência mais simples.

Na base do banco de filtros cocleares (representam o deslocamento mecânico da membrana basilar [13]), está a ideia de capturar as características essenciais dos sons da maneira como são percebidos pelo ouvido humano. Esse processo envolve a transformação do sinal de áudio em uma série de faixas de frequência, cada uma correspondendo a uma parte diferente da cóclea. As k bandas são dispostas segundo a escala *Bark* pela equação,

$$f_{Bark} = 13 \arctan\left(\frac{76f}{1000}\right) + 3,5 \arctan\left(\frac{f}{7500}\right)^2 \quad (6)$$

e f_{Bark} é a frequência perceptual em *Bark* e f é a frequência em Hertz.

III. Metodologia e implementação

A. Base de dados

Os dados utilizados nas análises foram extraídos da base de vozes do professor Edson Cataldo e do site do *Saarbruecken Voice Database* (SVD), que chamaremos de base 1 e 2, respectivamente.

A base 1 é composta por 11 vozes saudáveis, 12 vozes com nódulo e 8 vozes com paralisia. A base 2 é constituída por 687 vozes saudáveis, 6 vozes com cisto, 68 vozes com edema de Reinke. Não houve distinção entre vozes masculinas e femininas; o banco de vozes (banco 1 ou 2) foi organizado de acordo com o tipo de patologia. As gravações incluem pronúncias sustentadas das vogais /a/ e /e/, com intensidade variando entre tom normal, alto e baixo em alguns casos.

As redes neurais precisam de um grande número de dados para serem capazes de fazer generalizações e, com isso, ter um bom desempenho no reconhecimento de objetos [14]. Por esse motivo, foram empregadas técnicas computacionais para aumentar os dados [15], como o corte dos áudios em 0,5 e 0,7 segundos (proporcionais aos tempos dos áudios originais), a adição de ruído branco e o deslocamento temporal da parte inicial do áudio (deslocamento dos elementos de um *array* x posições para frente, com valores de 1000, 3500 e 6000 para x).

Após a conclusão do aumento de dados, totalizamos 1359 vozes saudáveis, 1152 vozes com edema de Reinke, 1017 vozes com nódulo, 765 vozes com paralisia e 261 vozes com cisto.

B. Extração de características das vozes

Os processos de extração de características foram realizados utilizando *Python*. Para o método MFCC, foi empregada a biblioteca *Librosa* (versão 0.10.2.post1), com a função *librosa.feature.mfcc()* configurada para gerar 40 coeficientes. Para extrair características pelo método PNCC, foi utilizada a função *spafe.features.pncc()* da biblioteca *Spafe* (versão 0.3.3); 32 coeficientes foram obtidos para áudios de 0,5 s, e 45 coeficientes para áudios de 0,7 s, a quantidade de coeficientes está relacionada com o tempo dos áudios, uma característica específica da biblioteca *Spafe*. O método ZCPA foi implementado por [4], utilizando a biblioteca *Scipy* (versão 1.13.1) por meio da função *signal.filtfilt()*, resultando em 66 coeficientes.

C. Características da redes neurais

Foram alocados 70% dos dados para treinamento, 10% para validação e 20% para teste. A camada de saída de todas as redes tinha 5 neurônios e utilizava a função de ativação *softmax*. O treinamento dos quatro modelos foi realizado em 100 épocas, com lotes de 32 amostras cada.

A rede DNN foi construída com uma camada densa de 32 neurônios e função de ativação *ReLU*.

Na CNN, foram implementados dois blocos convolucionais. O primeiro tem uma camada convolucional unidimensional com 64 filtros, *kernel* de tamanho 10 e função de ativação *ReLU*, além de um *dropout* de 40% e uma camada de *max-pooling* de tamanho 4. O segundo bloco, também unidimensional, possui 128 filtros, *kernel* de tamanho 10, *padding* = *"same"*, garantindo que a saída tenha o mesmo tamanho que a entrada, função de ativação *ReLU*, seguido por um *dropout* de 40% e *max-pooling* de tamanho 4. Essa rede também inclui uma camada densa de 64 neurônios com 40%.

A LSTM conta com três camadas, iniciando com duas camadas LSTM de 100 unidades cada, seguidas de uma camada de saída, intercaladas por dois *dropouts* de 30%.

Por fim, a BiLSTM possui duas camadas LSTM bidirecionais de 100 unidades cada, com um *dropout* de 30% entre elas.

Podemos observar na Tabela I a relação entre o tempo total de treinamento e a época em que cada rede neural atingiu a melhor performance durante as 100 épocas realizadas.

TABELA I
TEMPO DE TREINAMENTO E ÉPOCA DE MELHOR PERFORMANCE

	MFCC			
Arquitetura	**DNN**	**CNN**	**LSTM**	**BiLSTM**
Tempo	28.06"	2' 23.04"	13' 26.25"	27' 24.36"
Época	82ª	69ª	73ª	86ª
	PNCC			
Arquitetura	**DNN**	**CNN**	**LSTM**	**BiLSTM**
Tempo	41.79"	56.62"	9' 21.89"	24' 32.36"
Época	88ª	100ª	89ª	72ª
	ZCPA			
Arquitetura	**DNN**	**CNN**	**LSTM**	**BiLSTM**
Tempo	41.77"	3' 23.08"	18' 25.89"	43' 40"
Época	99ª	66ª	12	72ª

IV. Discussão dos resultados

Neste trabalho, são apresentados os resultados da combinação dos métodos MFCC, PNCC e ZCPA com redes neurais DNN, CNN, LSTM e BiLSTM para a classificação de vozes [16][17]. Métricas como acurácia, precisão, *recall* e *F1-score* foram empregadas para avaliar detalhadamente o desempenho de cada modelo de rede neural.

A acurácia fornece uma visão geral da capacidade do modelo em fazer previsões corretas (tanto positivas quanto negativas), mas isso não revela o quadro completo. A precisão foca na exatidão das previsões positivas, calculando a proporção de verdadeiros positivos entre todas as previsões positivas, e é crucial quando o custo de falsos positivos é alto. O *recall*, por outro lado, avalia a capacidade do modelo de identificar todos os positivos reais, sendo vital em cenários onde não se pode deixar de detectar um caso real. O f1-Score

é uma métrica que harmoniza precisão e *recall*, fornecendo um indicador global de desempenho balanceado. Na definição do estudo, classe positiva indica a presença de patologia vocal, enquanto negativa representa voz saudável.

A. Resultados MFCC

A seguir, na Tabela II, são apresentados os resultados da utilização de redes neurais para a classificação de patologias vocais e vozes saudáveis por meio do método MFCC.

O modelo DNN de classificação vocal alcançou uma taxa de acurácia de 99%, evidenciando excelente desempenho. Destaca-se o diagnóstico preciso de nódulos e paralisia. Apesar de não atingir 100% para cistos e edema de Reinke, a alta acurácia indica a confiabilidade do método.

TABELA II
CLASSIFICAÇÃO DNN, CNN, LSTM E BILSTM - MÉTODO MFCC

	DNN			CNN		
Condição	**Precisão**	**Recall**	**F1**	**Precisão**	**Recall**	**F1**
Cisto	0.93	0.88	0.90	0.97	0.67	0.79
Nódulo	1.00	1.00	1.00	1.00	1.00	1.00
Paralisia	1.00	1.00	1.00	1.00	1.00	1.00
Reinke	0.96	0.98	0.97	0.92	0.98	0.95
Saudável	1.00	0.99	1.00	0.98	0.99	0.99
Acurácia			=0.99			=0.97
	LSTM			BiLSTM		
Condição	**Precisão**	**Recall**	**F1**	**Precisão**	**Recall**	**F1**
Cisto	0.93	0.93	0.93	0.95	0.91	0.93
Nódulo	1.00	1.00	1.00	1.00	1.00	1.00
Paralisia	1.00	1.00	1.00	1.00	1.00	1.00
Reinke	0.98	0.96	0.97	0.98	0.99	0.98
Saudável	0.99	1.00	0.99	1.00	1.00	1.00
Acurácia			=0.99			=0.99

O modelo CNN possui uma acurácia geral de 97%, ligeiramente abaixo do modelo DNN. Destaca-se a precisão na identificação de nódulos e paralisia. Embora a condição de cisto tenha alta precisão, seu *recall* é mais baixo em relação ao DNN, indicando uma tendência de não detectar todos os casos reais. O desempenho na identificação de edema de Reinke e vozes saudáveis permanece eficiente.

Os modelos LSTM e BiLSTM se mantêm com uma alta precisão para nódulo, paralisia, edema de Reinke e vozes saudáveis e obteve um aumento considerável no *recall*, mostrando a melhora na detecção de cisto. A performance geral do modelo é acentuada pelo alto valor da acurácia.

Por fim, o modelo de classificação mantém uma eficácia notável com uma acurácia de 99%, evidenciando uma precisão e *recall* perfeitos na identificação de nódulos, paralisia e vozes saudáveis. O desempenho na detecção de cistos e edema de Reinke também foi elevado.

B. Resultados PNCC

Os resultados dos modelos de classificação vocal para o método PNCC apresentados na Tabela III apresentam uma acurácia de 47% a 65%, demonstrando a necessidade de aprimoramento no método.

C. Resultados ZCPA

A Tabela IV exibe os resultados dos modelos de classificação vocal usando o método ZCPA com acurácia variando entre 30% e 50%. Isso sugere a necessidade de melhorias no método.

TABELA III
CLASSIFICAÇÃO DNN, CNN, LSTM E BILSTM - MÉTODO PNCC

	DNN			CNN		
Condição	**Precisão**	**Recall**	**F1**	**Precisão**	**Recall**	**F1**
Cisto	0.00	0.00	0.00	1.00	0.09	0.16
Nódulo	0.50	0.70	0.58	0.60	0.96	0.74
Paralisia	0.44	0.08	0.13	0.88	0.09	0.17
Reinke	0.48	0.58	0.52	0.61	0.66	0.63
Saudável	0.44	0.52	0.48	0.65	0.77	0.70
Acurácia			=0.47			=0.63
	LSTM			BiLSTM		
Condição	**Precisão**	**Recall**	**F1**	**Precisão**	**Recall**	**F1**
Cisto	0.44	0.25	0.31	0.47	0.30	0.37
Nódulo	0.74	0.88	0.80	0.76	0.77	0.76
Paralisia	0.80	0.44	0.57	0.65	0.62	0.64
Reinke	0.55	0.64	0.59	0.57	0.71	0.64
Saudável	0.63	0.67	0.65	0.69	0.62	0.65
Acurácia			=0.64			=0.65

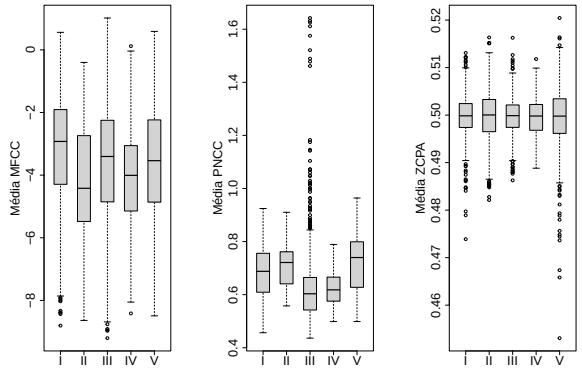
TABELA IV
CLASSIFICAÇÃO DNN, CNN, LSTM E BILSTM - MÉTODO ZCPA

	DNN			CNN		
Condição	**Precisão**	**Recall**	**F1**	**Precisão**	**Recall**	**F1**
Cisto	0.00	0.00	0.00	0.67	0.04	0.07
Nódulo	0.54	0.86	0.66	0.57	0.87	0.69
Paralisia	0.00	0.00	0.00	0.54	0.20	0.30
Reinke	0.37	0.27	0.31	0.40	0.35	0.37
Saudável	0.46	0.70	0.55	0.49	0.63	0.55
Acurácia			=0.47			=0.50
	LSTM			BiLSTM		
Condição	**Precisão**	**Recall**	**F1**	**Precisão**	**Recall**	**F1**
Cisto	0.00	0.00	0.00	0.15	0.05	0.08
Nódulo	0.00	0.00	0.00	0.55	0.71	0.62
Paralisia	0.00	0.00	0.00	0.47	0.30	0.36
Reinke	0.00	0.00	0.00	0.41	0.46	0.43
Saudável	0.30	1.00	0.46	0.51	0.52	0.52
Acurácia			=0.30			=0.48

V. ANÁLISE DAS CARACTERÍSTICAS EXTRAÍDAS DAS VOZES

Para auxiliar a compreensão dos desempenhos dos métodos utilizados, apresentamos a distribuição das médias e amplitudes das características extraídas das vozes por meio de *boxplots*, em que o eixo vertical representa os valores das características, enquanto o eixo horizontal indica o tipo de voz: I (saudável), II (nódulo), III (edema de Reinke), IV (cisto) e V (paralisia). Comparando os *boxplots* das médias apresentados na Figura 2, observa-se que o segundo quartil (mediana) do método ZCPA é praticamente o mesmo valor para os cinco tipos de vozes e se tornam mais diferentes nos métodos PNCC e MFCC. Além disso, a dispersão dos dados entre o primeiro e segundo quartis, e entre o segundo e terceiro quartis, é maior no método MFCC e menor nos métodos PNCC e ZCPA. Essa maior dispersão nos dados do MFCC sugere que suas características são mais distintas entre si, facilitando a classificação.

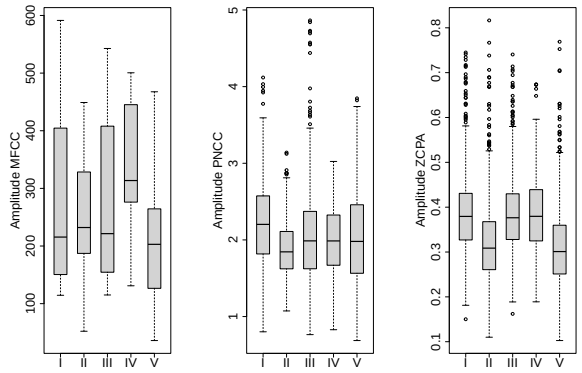
Em relação aos *boxplots* das amplitudes, Figura 3, no método MFCC, o segundo quartil revela valores mais divergentes em comparação com os métodos PNCC e ZCPA.



(a) (b) (c)

Fig. 2. Boxplot, médias das características MFCC (a), PNCC (b) e ZCPA (c).

A variabilidade dos dados entre o primeiro e segundo quartis, bem como entre o segundo e terceiro quartis, é mais evidente no método MFCC em comparação com os métodos PNCC e ZCPA. Essa observação da dispersão dos dados entre os quartis ressalta a disparidade entre as características, indicando uma diferenciação mais clara entre elas. Isso sugere que os dados são mais facilmente distinguíveis pelas redes neurais durante o processo de classificação das vozes.



(a) (b) (c)

Fig. 3. Boxplot, amplitudes das características MFCC (a), PNCC (b) e ZCPA (c).

VI. Conclusões

O método MFCC combinado com as redes neurais LSTM e BiLSTM demonstrou ser altamente eficaz na classificação de patologias vocais e vozes saudáveis, alcançando métricas elevadas de precisão, *recall* e *F1-score*, além de uma acurácia geral de 99%, um dos melhores resultados observados na literatura. Isso indica que os modelos são extremamente confiáveis e eficazes para aplicações práticas em reconhecimento e classificação de patologias vocais.

O método PNCC combinado com a rede neural BiLSTM alcançou uma acurácia de 65%, indicando que as métricas de precisão, *recall* e *F1-score* revelam dificuldades do modelo em classificar corretamente as condições propostas. O desempenho relativamente baixo em algumas dessas métricas sugere melhorias.

Finalmente, o método ZCPA mostrou-se o menos eficaz entre os três, tendo uma performance máxima de 50%, mostrando a incapacidade de detectar condições patológicas e vozes saudáveis.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

[1] M. Behlau e P. Pontes e F. Moreti, “Higiene vocal: cuidando da voz,” *Thieme Revinter Publicações LTDA*, Rio de Janeiro: Revinter, 2018.

[2] S. S. L. da Silva, “Principais patologias laríngeas em professores,” *Distúrbios da Comunicação*, v. 30, n. 4, pp. 767–775, 2018.

[3] D. G. da Silva e C. D. R. Cuadros e A. Alcaim, “Reconhecimento Robusto de Locutor Baseado nos Atributos ZCPAC,” *Simpósio Brasileiro de Telecomunicações*, Recife, PE, 2007.

[4] K. R. F. da Silva, “Reconhecimento de locutor em ambientes ruidosos: uma comparação entre os métodos de extração de características MFCC e ZCPA,” *Trabalho de Conclusão de Curso (Graduação) - Engenharia de Telecomunicações, Universidade Federal Fluminense.* Niterói, RJ, p. 98, 2023.

[5] V. G. R. da Silva, “Análise do sinal de fala para reconhecimento de emoções utilizando representação semântica,” *Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal de Sergipe.* São Cristóvão, SE, p. 9, Dez. 2022.

[6] J. K. Siqueira, “Reconhecimento de Voz Contínua com Atributos MFCC, SSCH e PNCC, Wavelet Denoising e Redes Neurais,” *Dissertação (Mestrado em Engenharia Elétrica), Pontifícia Universidade Católica do Rio de Janeiro.* RJ, p. 261, Set. 2011.

[7] G. L. Ribeiro, R. Gomes, S. C. Costa e W. C. A. Costa, “Análise melcepstral na discriminação de patologias laríngeas,” *XXIV congresso brasileiro de engenharia biomédica - CBEB*, Uberlândia-MG, 2014.

[8] V. J. D. Vieira, S. C. Costa, W. C. de A. Costa, S. E. N. Correia e J. M. F. R. de Araújo, “Avaliação de Desempenho na Classificação de Patologias Laríngeas por Análise LPC de Sinais de Voz e Redes Neurais MLP,” *Anais do XIII Congresso Brasileiro de Inteligência Computacional - SBIC*, pp. 1–6, 2013.

[9] A. Ali and S. Ganar, “Intelligent pathological voice detection,” *Int. J. Innov. Res. Technol*, v. 5, n. 5, pp. 92–95, Oct. 2018.

[10] H. Cordeiro, J. Fonseca, I. Guimarães and C. Meneses, “Voice pathologies identification speech signals, features and classifiers evaluation”, *2015 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp. 81–86, IEEE, 2015.

[11] M. T. S. Al-Kaltakchi and H. A. Al-R. Taha and M. A. Shehab and M. A. M. Abdullah, “Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database,” *Indonesian Journal of Electrical Engineering and Computer Science*, v. 18, pp. 782–789. Maio 2020.

[12] C. Kim and R. M. Stern, “Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition,” *Ieee transactions on audio, speech, and language processing*, v. 24, pp. 1315–1329, July 2016.

[13] C. R. Almeida, “Extratores de características acústicas inspirados no sistema periférico auditivo,” *Dissertação (Mestrado em Engenharia Elétrica), Universidade Federal de Sergipe.* São Cristóvão, SE, p. 56, Set. 2014.

[14] E. Cataldo and C. Soizeb, “A stochastic model of voice generation and the corresponding solution for the inverse problem using Artificial Neural Network for case with pathology in the vocal folds,” *Biomedical Signal Processing and Control,* v. 1, p. 10, 2021.

[15] L. Ferreira-Paiva and E. Alfaro-Espinoza and V. M. Almeida and at al, “A survey of data augmentation for audio classification,” *Congresso Brasileiro de Automática-CBA.*, Vol. 3. No. 1. 2022.

[16] L. C. Dias, “Detecção de patologias laríngeas por meio da análise de sinais de voz utilizando Deep Neural Networks,” *Dissertação (Mestrado em Engenharia Elétrica), Instituto Federal da Paraíba.* João Pessoa, PB, p. 82, Jul. 2020.

[17] J. N. Salvado Júnior, “Identificação de patologias em pregas vocais por meio de classificador bag of features e redes neurais convolucionais,” *Dissertação (Mestrado em Engenharia Elétrica), Faculdade de Engenharia de Bauru.* Bauru, SP, p. 70, 2022.