Reconhecimento de comandos de voz com e sem disartria usando extração de características da fala MFCC e algoritmos de aprendizagem de máquina

Jordana Seixas[1], Ailton Leite[1], Rodrigo de Paula[2], Sérgio Murilo Maciel Fernandes[1]

[1]Escola Politécnica de Pernambuco– Universidade de Pernambuco (UPE)
Recife – PE – Brazil

[2]Universidade Católica de Pernambuco – Unicap – Recife-PE – Brazil

{jls3,asl,smurilo}@ecomp.poli.br, rodrigo.paula@unicap.br

Abstract. *Dysarthric speech is among the problems in articulating and pronouncing words well due to damage to the neurological system responsible for speech. This study investigates whether machine learning classifiers recognize which words people with and without dysarthria speak by applying a speech feature extraction technique called MFCC (Mel Frequency Cepstral Coefficients). Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), and KNearest Neighbor (KNN) classifiers were tested. UASpeech dataset was used in the models, containing speakers with and without dysarthria. The results showed good performance with average accuracy for KNN (98.5%), ANN (95%), RF (91.8%), and SVM (89.5%).*

Resumo. *A fala disártrica está entre os problemas para articular e pronunciar bem as palavras devido aos danos no sistema neurológico responsável pela fala. Este estudo investiga se os classificadores de aprendizagem de máquina reconhecem quais palavras as pessoas com e sem disartria falam, aplicando uma técnica de extração de características da fala chamada MFCC (Mel Frequency Cepstral Coefficients). Os classificadores Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF) e KNearest Neighbor (KNN) foram testados. O conjunto de dados UASpeech foi usado nos modelos, contendo falantes com e sem disartria. Os resultados mostraram bom desempenho com acurácia média para KNN (98,5%), ANN (95%), RF (91,8%) e SVM (89,5%).*

1. Introdução

Em todo o mundo, os problemas mais comuns associados a uma população cada vez mais envelhecida são o aumento da incidência de distúrbios neurológicos, seja pela doença de Parkinson (DP), acidente vascular cerebral ou traumatismo cranioencefálico (TCE). As principais consequências dessas doenças são os distúrbios motores da fala, incluindo a disartria. A disartria é causada por problemas de controle neuromuscular, levando à diminuição da inteligibilidade da fala e comprometimento da comunicação [Yılmaz 2019]. A principal consequência da disartria é a degradação da inteligibilidade da fala causada pela má articulação das consoantes e, nos casos mais graves, pela distorção das vogais. Devido a essas variações críticas, os sistemas padrão de

reconhecimento de fala (ASR), quando projetados para falantes com disartria, não apresentaram um bom desempenho no reconhecimento da fala disártrica devido a articulações de fala deficientes [Yakoub 2020].

O objetivo deste trabalho é investigar se os classificadores de aprendizagem de máquina reconhecem as palavras faladas por pessoas com disartria e sem disartria com foco nos classificadores ANN, SVM, Random Forest e KNN, aplicando a técnica de extração de características da fala MFCC (*mel frequency cepstral coefficients*). A base de dados utilizada foi a UASpeech [Kim 2008] com falas disártricas (com níveis de inteligibilidade da mais baixa, baixa, média e alta inteligibilidade) e não disártricas (fala normal ou de controle).

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta trabalhos relacionados. A Seção 3 aborda a metodologia proposta. A Seção 4 discorre os resultados e análises. Na Seção 5 é apresentada a conclusão e trabalhos futuros.

2. Trabalhos Relacionados

Um estudo comparativo em [Joshy 2022], sobre a classificação dos níveis de gravidade da disartria usando diferentes técnicas de aprendizado profundo e características acústicas, utiliza a extração dos recursos básicos de fala, ou seja, MFCCs e Coeficientes Q cepstrais. Os MFCCs são empregados em inúmeras tarefas de classificação de áudio e sistemas de reconhecimento de fala com motivação perceptiva, além de seu uso generalizado no reconhecimento automático de timbres monofônicos ou polifônicos. As características do MFCC motivaram este trabalho de pesquisa, conforme apresentado em [Joshy 2022], o qual utiliza apenas os recursos básicos do MFCC para investigar o desempenho de vários modelos de aprendizado profundo para classificação de gravidade disártrica. Avaliar distúrbios de voz por meio do uso do MFCC para o reconhecimento da doença de Alzheimer foi realizado em [Boualoulou 2022], buscando distinguir duas categorias de pacientes. Para classificação utilizaram o classificador KNN para distinguir entre pessoas saudáveis e pessoas com doença de Parkinson.

O artigo [Kuresan 2021] analisa um conjunto de dados, contendo vários recursos vocais, usado como entrada para analisar o desempenho de vários algoritmos de aprendizado de máquina, dentre eles, KNN, SVM e ANN, para uma detecção precoce da Doença de Parkinson(DP) utilizando a técnica de extração MFCC. A melhor acurácia de classificação foi obtida pelo ANN em torno de 90%.

3. Metodologia Proposta

Esta seção apresenta a metodologia proposta para reconhecimento dos comandos de fala de pessoas com e sem disartria, conforme a Figura 1.

Figura 1 – Modelo proposto

As etapas do modelo proposto são: i) Base de dados: o conjunto de dados utilizado para este trabalho consiste em uma amostra de até 200 comandos de voz (ou palavras), totalizando 80725 exemplos (ou número de arquivos), entre fala disártrica e

não disártrica da base de dados UASpeech [Kim 2008]. UASpeech contém mais de 450 palavras composto por 13 falantes de controle (sem disartria) e 19 falantes com disartria (com quatro níveis de inteligibilidade, ou seja, *'very low', 'low', 'mid'* e *'high'*), com gêneros masculino e feminino; ii) Pré-processamento dos dados: os arquivos de áudio UASpeech são de canal único (ou seja, 1 canal de áudio), com uma taxa de amostragem de 16 kHz. Antes dos dados de áudio, no formato ".wav", serem inseridos nos modelos, todas as amostras foram redimensionadas para terem a mesma duração. Arquivos corrompidos foram eliminados; iii) Extração das características da fala (MFCC): a coordenação dos músculos vocais influencia a inteligibilidade da fala, os MFCCs podem capturar movimentos irregulares das pregas vocais ou falta de fechamento das pregas vocais devido a alterações de massa/tecido [Sahane 2021]; iv) Treinamento e teste: os dados foram divididos em conjuntos de treinamento e teste em uma proporção de 70:30. Esta divisão apresenta estas proporções devido ao pequeno número de amostras minoritárias em nosso conjunto de dados e possivelmente por implicar nas habilidades de aprendizado dos classificadores utilizados neste trabalho. É preferível um conjunto de treinamento o maior possível [Mooijman 2023]; v) Modelos de aprendizagem de máquina foram: SVM (amplamente utilizado para lidar com problemas de classificação e regressão. Tem o objetivo de encontrar a linha ideal ou limite de decisão para classificar o espaço dimensional em seções, de modo que os pontos de dados sucessivos possam ser classificados convenientemente [Sheth 2022]), *Random Forest* (um método de *ensemble learning* utilizado para problemas de classificação. RF apresenta uma combinação de várias árvores de decisão, cada uma contribuindo com um único voto. E a classificação ocorrendo por voto da maioria [Hashimoto 2021]), KNN (utilizado para resolver problemas de classificação e regressão. KNN avalia as distâncias entre uma consulta e cada exemplo nos dados, escolhe os K exemplos mais próximos da consulta e, em seguida, seleciona o rótulo com a maior frequência (no caso de classificação) ou calcula a média dos rótulos (no caso de regressão) [Sheth 2022]) e ANN (uma ferramenta importante para classificação. O sistema ASR baseado em ANN foi empregado com sucesso para fala normal. As ANNs foram aplicadas em um sistema de reconhecimento de fala disártrica, mas com sucesso limitado [Trentin 2001]); vi) Métricas de desempenhos: os classificadores de aprendizagem de máquina foram avaliados através das métricas de desempenho: acurácia (*accuracy*), precisão (*precision*) e sensitividade (ou *recall*). Estas métricas têm as seguintes fórmulas:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \quad Precision = \frac{TP}{TP+FP}, \quad Recall = \frac{TP}{TP+FN}.$$

Onde: TP significa verdadeiro positivo, TN significa verdadeiro negativo, FP significa falso positivo e FN significa falso negativo. Neste trabalho, consideramos outra métrica chamada tempo de computação [Harris-Birtill 2021], que avalia o tempo gasto para cada modelo classificador. vii) Inferência: apresenta resultado do reconhecimento de quais palavras estão sendo faladas por pessoas com e sem disartria com percentual de acerto. Na seção de resultados será exibida com mais detalhes.

Utilizamos o ambiente virtual Google Colab com acelerador de *hardware* GPU, com 25,45 GB de RAM e capacidade de disco de até 166,77 GB.

4. Resultados e Análises

A Tabela 1 mostra o resultado da acurácia em reconhecer os comandos de voz com e sem disartria para cada classificador de acordo com o número de palavras, utilizando a técnica de extração de características MFCC.

Tabela 1 – Resultado da classificação usando MFCC

Número de palavras		4	25	50	100	200
Quantidade de exemplos		1818	10678	21360	40452	80725
Acurácia (%)	ANN	99,63	98,13	97,6	95,13	85,13
	SVM	98,95	92,29	89,05	86,26	80,85
	Random Forest	95,55	90,95	97,63	89,73	85,03
	KNN	98,95	97,95	98,37	98,73	98,66

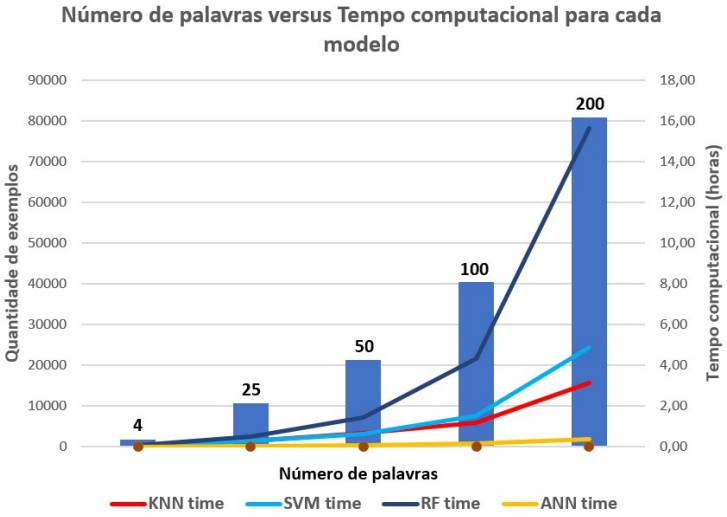


Figura 2 – Número de palavras versus Tempo

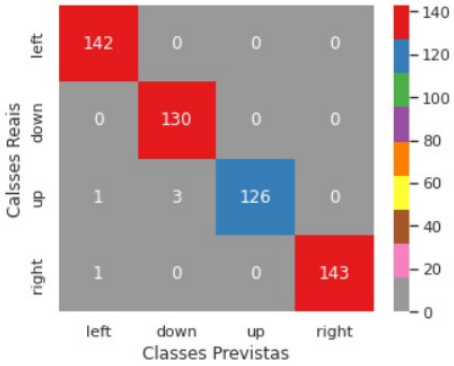


Figura 3 – Matriz de confusão do classificador ANN (para 4 palavras)

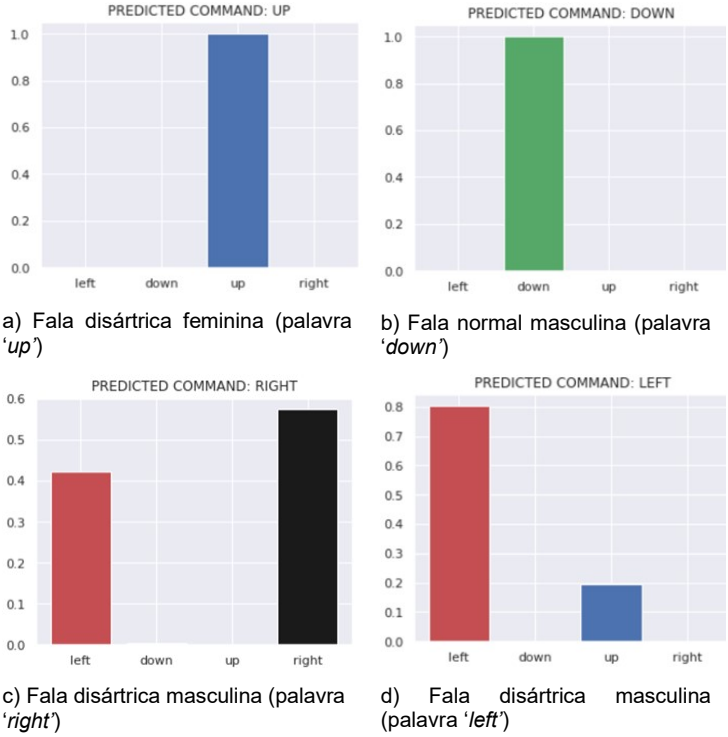
A Figura 2 ilustra o resultado do tempo (ou *time*) de computação (em horas) de cada classificador. O classificador ANN apresentou o melhor tempo computacional (inferior a 0,4 hora para 200 palavras), por outro lado, a acurácia diminui à medida que o número de palavras aumenta (de acordo com a Tabela 1). O classificador KNN teve o segundo melhor tempo de computação, aproximadamente 3 horas para 200 palavras. Enquanto que para os classificadores SVM e Random Forest a medida que o número de palavras aumenta, o tempo aumenta e a acurácia diminui (visto na Tabela 1).

Tabela 2 –Resultados das métricas para o classificador KNN

Número de palavras	Acurácia (%)	Precisão (%)	Sensitividade (ou *recall*) (%)
4	98,95	98,68	98,42
25	97,95	98,36	97,54
50	98,37	98,19	97,38
100	98,73	98,05	97,26
200	98,66	98,01	97,15

A Figura 3 exibe resultados satisfatórios do ANN (escolhido como exemplo para análise dos dados de teste) para 4 palavras (ou 30% de 1818 exemplos) utilizando a métrica matriz de confusão. Nesta matriz obervamos na diagonal principal o número de

acertos dos comandos de voz com e sem disartria. Se demonstrássemos exemplos com 25 ou mais palavras seria difícil visualizar o número de acertos das palavras na diagonal principal da matriz. A Tabela 2 destaca o classificador KNN que obteve a melhor acurácia média acima de 98%, com relação aos outros classificadores. E a métrica *recall* com média igual a 97,6% em proporção de predições corretas de uma classe e a métrica precisão com média igual a 98,3%, refletindo a confiabilidade do modelo na classificação de amostras como positivas.



a) Fala disártrica feminina (palavra '*up*')

b) Fala normal masculina (palavra '*down*')

c) Fala disártrica masculina (palavra '*right*')

d) Fala disártrica masculina (palavra '*left*')

Figura 4 – Resultado da inferência dos dados de teste do classificador ANN (para 4 palavras)

A Figura 4 exibe exemplos para 4 palavras, para melhor visualização do percentual de acertos, para a inferência do classificador ANN. A Figura 4(a) exibe 100% de acerto da palavra "*up*" com fala disártrica feminina. Enquanto que a Figura 4(b) ilustra 100% de acerto da palavra "*down*" com fala normal masculina. Já para a palavra "*right*" com fala disártrica masculina obteve 58% de acerto, ilustrado na Figura 4(c). E a palavra "*left*" com fala disártrica masculina apresenta 80% de acerto, exibida na Figura 4(d). O resultado foi satisfatório para o modelo ANN para 4 palavras (ou 1818 exemplos de falas disártrica e normal) com uma acurácia de 99,63% (exibido na Tabela 1).

5. Conclusões e Trabalhos Futuros

Ao implementar a análise de componentes principais, KNN, SVM, ANN e Random Forest tiveram um bom desempenho para a fusão da técnica de extração de características da fala MFCC para um tamanho de teste de 30%. A principal contribuição científica deste trabalho é ajudar pessoas com fala disártrica serem compreendidas. À medida que aumentamos o número de palavras, os resultados experimentais indicaram que o KNN apresentou melhor desempenho (acurácia média acima de 98%) comparado com os demais classificadores. Em se tratando de aplicações em tempo real para o reconhecimento das palavras faladas com e sem disartria, o classificador ANN seria um candidato satisfatório para dar o suporte com acurácia media de 95%. Nas pesquisas futuras serão testados outros métodos para gerar dados sintéticos mais representativos em relação aos dados reais, por exemplo, as redes GAN (*Generative Adversarial Network*) e trabalhar com outros classificadores, como o classificador CNN (*Convolutional Neural Network*), convertendo voz em imagem, e reconhecer frases com duas ou mais palavras combinando algoritmos de aprendizagem de máquina com os melhores parâmetros.

Referências

Boualoulou, N., Nsiri, B., Drissi, T.B. and Zayrit, S., 2022. Speech analysis for the detection of Parkinson's disease by combined use of empirical mode decomposition, Mel frequency cepstral coefficients, and the K-nearest neighbor classifier. In ITM Web of Conferences (Vol. 43, p. 01019). EDP Sciences.

Harris-Birtill, D. and Harris-Birtill, R., 2021. Understanding computation time: a critical discussion of time as a computational performance metric. In Time in Variance (pp. 220-248). Brill.

Hashimoto, BF, Shigueoka, LS, Costa, VP e Gomi, ES, 2021, junho. Treinamento de Classificadores de Aprendizagem de Máquina para o Diagnóstico do Glaucoma com o Uso de Dados de Perimetria Automatizada Padrão (SAP). In Anais Estendidos do XXI Simpósio Brasileiro de Computação Aplicada à Saúde (pp. 151-156). SBC.

Joshy, A. A., & Rajan, R. (2022). Automated Dysarthria Severity Classification: A Study on Acoustic Features and Deep Learning Techniques. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 30, 1147-1157.

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., & Frame, S. (2008). Dysarthric speech database for universal access research. In Ninth Annual Conference of the International Speech Communication Association.

Kuresan, H., Samiappan, D., Jeevan, A. and Gupta, S., 2021. A Performance Study of ML Models and Neural Networks for Detection of Parkinson Disease using Dysarthria Symptoms. European Journal of Molecular & Clinical Medicine, 8(03), p.2021.

Mooijman, P., Catal, C., Tekinerdogan, B., Lommen, A. and Blokland, M., 2023. The effects of data balancing approaches: A case study. Applied Soft Computing, 132, p.109853.

Sahane, P., Pangaonkar, S., & Khandekar, S. (2021, September). Dysarthric Speech Recognition using Multi-Taper Mel Frequency Cepstrum Coefficients. In 2021 International Conference on Computing, Communication and Green Engineering (CCGE) (pp. 1-4). IEEE.

Sheth, V., Tripathi, U. and Sharma, A., 2022. A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. Procedia Computer Science, 215, pp.422-431.

Trentin, E. and Gori, M., 2001. A survey of hybrid ANN/HMM models for automatic speech recognition. Neurocomputing, 37(1-4), pp.91-126.

Yakoub, M.S., Selouani, S.A., Zaidi, B.F. and Bouchair, A., 2020. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. EURASIP Journal on Audio, Speech, and Music Processing, 2020(1), pp.1-7.

Yılmaz, E., Mitra, V., Sivaraman, G. and Franco, H., 2019. Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech. Computer Speech & Language, 58, pp.319-334.