REPRODUÇÃO DE ARTIGO COMO PROJETO FINAL PARA A DISCIPLINA DE FUNDAMENTOS DE PESQUISA PARA CIÊNCIA DA COMPUTAÇÃO 2 DO SEMESTRE 2023.1 DO MESTRADO EM COMPUTAÇÃO DA UFCG

Aluno: Diego Ribeiro de Almeida

Orientador: Cláudio Elízio Calazans Campelo

1. INTRODUÇÃO

O presente trabalho é uma reprodução do artigo intitulado "Supervised Machine Learning Model for Accent Recognition in English Speech using Sequential MFCC Features" de Dweepa Honnavalli e Shylaja SS, originalmente publicado no livro Advances in Intelligent Systems and Computing, em 2021. O artigo descreve os resultados de algumas implementações de modelos de Aprendizado de Máquina Supervisionado para Reconhecimento de Sotaque em Discurso em Inglês, utilizando Coeficientes Cepstrais em Frequência de Mel (Mel Frequency Cepstral Coefficients - MFCC) . O trabalho engloba falantes de inglês com sotaques indiano e americano, construindo características MFCC sequenciais a partir dos quadros da amostra de áudio a fim de diferenciar os áudios em função dos sotaques dos falantes.

Os sistemas de Reconhecimento Automático de Fala (Automatic Speech Recognition - ASR) apresentam um desempenho melhor no reconhecimento da fala americana em comparação com a fala com sotaque, principalmente devido à falta de dados de treinamento inclusivos. Esse problema surge devido à existência de diversos idiomas, dialetos e sotaques. Por exemplo, a língua inglesa possui cerca de 100 sotaques e dialetos, o que torna inviável coletar dados anotados suficientes para todos eles. Isso limita a capacidade dos ASRs. Embora a disponibilidade de dados tenha aumentado consideravelmente, tornando os ASRs mais inclusivos, ainda há espaço para melhorias contínuas.

A detecção de sotaques, impulsionada pelo aumento do uso de aprendizado de máquina em aplicações do mundo real, encontra diversas aplicações no mundo real, como no aprimoramento dos assistentes virtuais inteligentes, ampliando sua base de usuários e usabilidade. No artigo, propõe-se a classificação de sotaques utilizando a concatenação de características sequenciais MFCC, o que pode aprimorar a compreensão de fala com sotaque pelos sistemas de reconhecimento automático de fala.

Informações de sinais de áudio podem ser extraídas de várias maneiras - por amostragem, janelamento, expressando o sinal no domínio da frequência ou extraindo características perceptivas. Uma das características mais comumente extraídas é o MFCC. O MFCC é derivado do conceito de bancos de filtros espaçados de forma logarítmica, combinados com o conceito do sistema auditivo humano.

Para diferenciar entre o sotaque americano e indiano, as características MFCC são extraídas e os dados das características são alimentados em alguns modelos de aprendizado de máquina para obter os resultados de classificação. Os modelos utilizados foram os seguintes: K- Nearest Neighbours, Support Vector Machine, Gaussian Mixture Model, Neural Networks e Logistic Regression.

Os dados utilizados para a condução dos experimentos e avaliações são provenientes do VCTK-corpus, uma coleção de arquivos .wav com duração de 3 a 5 segundos nos quais os falantes com sotaque americano e sotaque indiano gravaram o mesmo conteúdo.

2. METODOLOGIA

Apenas dois dos experimentos realizados no artigo foram escolhidos para serem reproduzidos. A principal contribuição do trabalho é a aplicação da extração de características MFCC na tarefa de classificação de sotaques, etapa que existe em todos os experimentos, que são diferenciados apenas pelo modelo de classificação aplicado. Os experimentos reproduzidos aqui são dois dos melhores modelos apresentados, a rede neural e a regressão logística.

2.1. Metodologia do trabalho original

O conjunto de dados utilizado neste estudo é composto por arquivos .wav do VCTK-corpus, com duração de 3 a 5 segundos cada. Ele contém gravações de falantes com sotaque americano e sotaque indiano pronunciando o mesmo conteúdo. No total, são 2301 amostras de áudio, descritas na Tabela 1, que foram divididas em um conjunto de treinamento com 80% das amostras e um conjunto de teste com os 20% restantes.

Sotaque	Número total de amostras de áudio (original)
Indiano	1032
Americano	1269

Tabela 1: Quantidade de dados utilizada no artigo original.

Como observado na tabela, o sotaque indiano está sub-representado, dessa forma, após o split no conjunto de dados em treino e teste, na proporção 80:20, respectivamente, os dados de treinamento do sotaque indiano são submetidos a um oversampling.

Os autores calculam então 20 coeficientes cepstrais em frequência de Mel para cada arquivo de áudio no conjunto de dados. Essa extração de características é implementada utilizando a biblioteca “Librosa”, da linguagem Python. Essas características são utilizadas para treinar o modelo a distinguir um determinado sotaque dos demais.

A Figura 1 exemplifica os MFCCs extraídos de um falante com sotaque americano em comparação com um falante com sotaque indiano, ambos pronunciando a mesma frase “Please call Stella”, e mostra que há uma variação na fala. Essa variação auxilia o modelo na previsão da classe dos sotaques, quando aplicada aos modelos de aprendizado de máquina já citados.

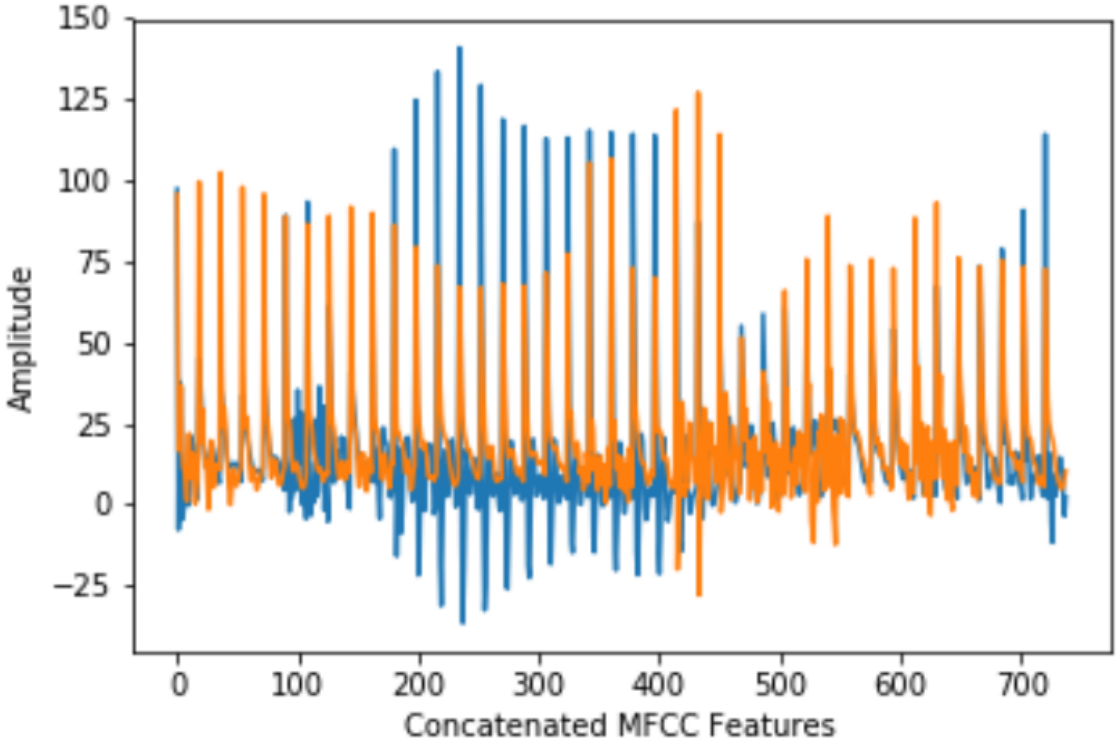

Figura 1: Features MFCC de duas amostras de áudio de sotaques diferentes

2.2. Metodologia da reprodução

O autor indica os falantes específicos para cada gênero e sotaque, a partir da base utilizada, que está disponível online, gratuitamente. O uso da base foi descrito de forma clara no artigo, de forma que sua obtenção e manipulação foram relativamente simples. Fez-se necessário baixar os dados, extraí-los e modificar a estrutura de arquivos de forma que fosse possível acessá-los pelos scripts dos experimentos. Foram utilizados exatamente os mesmos dados descritos no experimento original.

Além disso, as implementações também estão disponíveis em repositório online, público e gratuito. Após efetuar o clone do repositório, fez-se necessário criar um ambiente virtual, utilizando a biblioteca “virtualenv” da linguagem Python, de forma a evitar conflitos de versão.

O repositório base do artigo não disponibiliza um arquivo com as versões utilizadas, um “requirements.txt”, de forma que ficou incerto quais as versões adequadas para os experimentos. Utilizou-se Python 3.9, e as versões mais recentes das bibliotecas utilizadas para execução do código: sklearn, librosa, keras, tensorflow, numpy, matplotlib, entre outras.

Reproduziu-se a extração de recursos e o treinamento de dois modelos de classificação: A Rede Neural e a Regressão Logística. Para execução da reprodução utilizou-se uma máquina Linux, rodando o Ubuntu 20.04, com processador Intel Core i9-9900K, 32 GB de memória RAM, 1 TB (SSD) de armazenamento, e GPU RTX 2080 Ti, com 11 GB.

3. REPRODUÇÃO

Os experimentos duraram cerca de 2 minutos, cada, e consumiram recursos mínimos da máquina. O uso de GPU não foi necessário. Os resultados originais seguem os valores descritos na Figura 2, enquanto os resultados dos experimentos de reprodução seguem os valores descritos na Tabela 2.

Ao comparar os resultados obtidos para as métricas na reprodução com os obtidos pelo artigo original percebe-se que os modelos gerados na reprodução aparentam ter um comportamento similar, ou até ligeiramente superior.

Model	Precision	Recall	f-measure	Reject Rate	Accuracy
Neural Networks	0.96	0.94	0.95	0.97	0.95
KNN	0.9	0.92	0.91	0.9	0.91
Logistic Regression	0.94	0.96	0.95	0.95	0.95
SVM	1.0	0.02	0.04	1.0	0.54
GMM	0.43	1.0	0.60	0.0	0.43

Figura 2: Métricas dos modelos gerados no artigo original

Model	Precision	Recall	F-measure	Reject Rate	Accuracy
Neural Networks	0.9565	0.9659	0.9612	0.9648	0.9653
Logistic Regression	0.9578	0.938	0.9478	0.9701	0.9566

Tabela 2: Métricas dos modelos gerados nesta reprodução

3.1. Rede Neural

As Figuras 3 e 4 exibem os resultados do experimento original e da replicação, respectivamente, ao aplicar as features MFCC à uma Rede Neural. Percebe-se a diferença mínima nos acertos do modelo, que categoriza bem os áudios entre as diferentes classes.

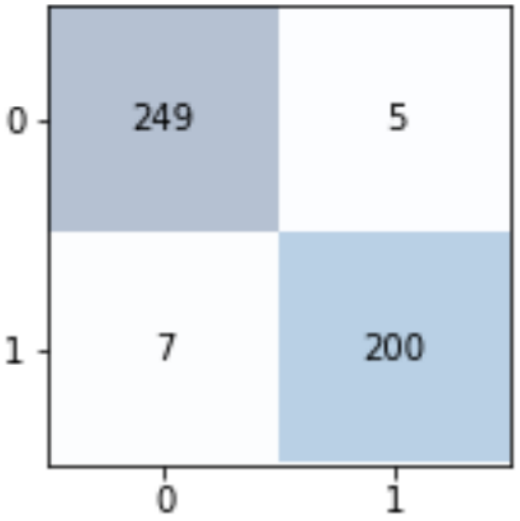

Figura 3: Matriz de Confusão da Rede Neural no artigo

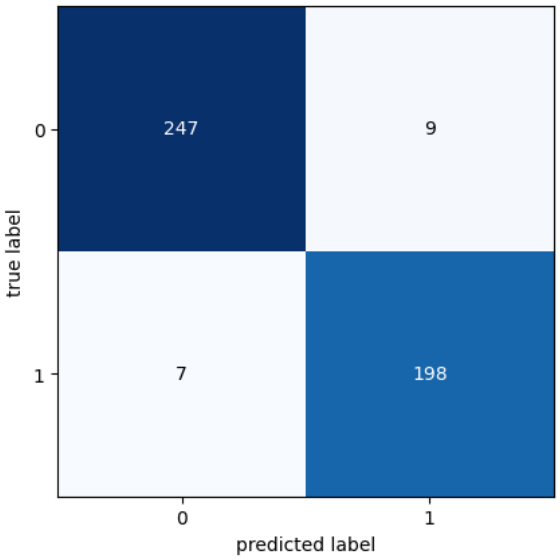

Figura 4: Matriz de Confusão da Rede Neural na reprodução

3.2. Regressão Logística

Um comportamento semelhante pode ser observado nas Figuras 5 e 6, que exibem, respectivamente, a matriz de confusão do experimento e da reprodução para o modelo de regressão logística.

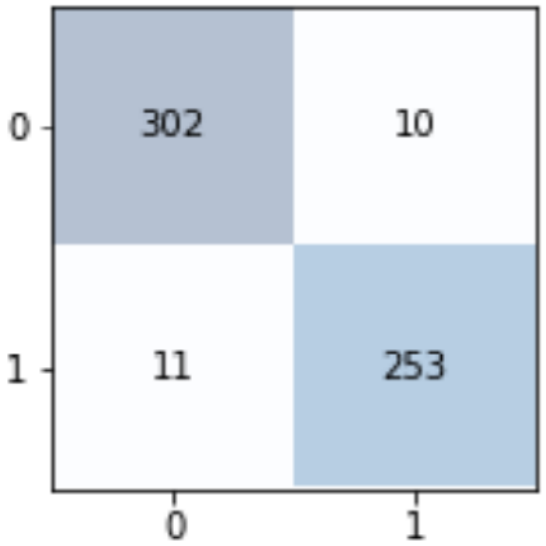

Figura 5: Matriz de Confusão da Regressão Logística no artigo

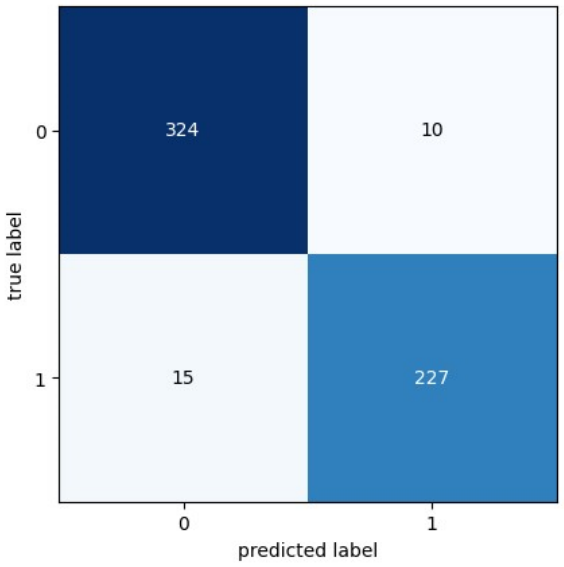

Figura 6: Matriz de Confusão da Regressão Logística na reprodução

3.3. Tempo

Uma outra variável observada pelos autores é o tempo que cada experimento leva para processar a inferência em todos os dados. No artigo original a Regressão Logística se destacava por seu tempo reduzido, em comparação com a Rede Neural. Esse comportamento se mantém também na reprodução, conforme se observa na Tabela 3, onde observa-se uma diminuição no tempo de execução da inferência com o modelo de Regressão Logística, e o aumento no tempo de execução no experimento com a Rede

Neural, em comparação com os resultados descritos no artigo. As variações, por serem em magnitudes baixas, podem se dar devido a fatores externos ao experimento, como a demanda geral do sistema, bem como pelo fato de que a reprodução foi feita em uma máquina distinta do artigo, que não informou as configurações utilizadas.

Modelo	Tempo (original)	Tempo (reprodução)
Rede Neural	18.33s	33.97s
Regressão Logística	0.58s	0.09s

4. CONCLUSÃO

Observa-se que o modelo é tanto reproduzível quanto consistente, uma vez que os dados obtidos durante a execução da reprodução se aproximam consideravelmente dos encontrados pelos autores. As métricas de qualidade e de tempo de execução permaneceram praticamente as mesmas, ou seguiram a mesma configuração. Uma possível melhoria para o trabalho seria a adição de outros falantes, bem como sua separação nos conjuntos de treino e teste. É importante tornar o conjunto de dados o mais diverso possível para garantir que o modelo esteja capturando características de sotaques propriamente ditas, e não apenas categorizando outras características específicas das vozes dos falantes selecionados. Por essa razão, não incluir os mesmos falantes no treino e no teste se configura como um teste apropriado para o contexto. Outra possível otimização seria testar os modelos obtidos em uma outra base de dados, de forma a perceber a capacidade de generalização dos classificadores.

Pode-se concluir que o artigo original apresenta informações suficientes para possibilitar uma reprodução, embora haja carência de algumas informações adicionais, como as versões das bibliotecas. Além disso, os classificadores reportados no artigo, e reproduzidos neste trabalho, aparentam classificar bem os áudios do sotaque indiano e americano no idioma inglês, para o conjunto de testes em questão.