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

DIEGO RIBEIRO DE ALMEIDA

COMPARAÇÃO ENTRE MODELOS COM DIFERENTES ABORDAGENS PARA CLASSIFICAÇÃO DE SOTAQUES BRASILEIROS

CAMPINA GRANDE - PB

2022

DIEGO RIBEIRO DE ALMEIDA

COMPARAÇÃO ENTRE MODELOS COM DIFERENTES ABORDAGENS PARA CLASSIFICAÇÃO DE SOTAQUES BRASILEIROS

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.

Orientador: Professor Dr. Claudio Elízio Calazans Campelo

CAMPINA GRANDE - PB

2022

DIEGO RIBEIRO DE ALMEIDA

COMPARAÇÃO ENTRE MODELOS COM DIFERENTES ABORDAGENS PARA CLASSIFICAÇÃO DE SOTAQUES BRASILEIROS

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:

Professor Dr. Claudio Elízio Calazans Campelo

Orientador – UASC/CEEI/UFCG

Professor Dr. Carlos Eduardo Santos Pires

Examinador – UASC/CEEI/UFCG

Professor Dr. Francisco Vilar Brasileiro

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 02 de Setembro de 2022.

CAMPINA GRANDE - PB

RESUMO

O sotaque se apresenta como uma das variáveis mais desafiadoras para a eficácia de sistemas de Automatic Speech Recognition. Além disso, sua classificação automática possui diversas aplicações potenciais, como a seleção de modelos especializados para text-to-speech e speech-to-text. Neste trabalho, avaliamos dois modelos de classificação de sotaques a partir da base de dados Braccent, a fim de compará-los com os métodos GMM-UBM, GMM-SVM, iVector, CNN 1D, CNN 2D e CNN 1D + LSTM. Os resultados experimentais obtidos demonstram que as abordagens aqui avaliadas apresentam desempenhos consideravelmente abaixo dos reportados na literatura em métricas como acurácia, precisão, revocação, e F1-score, corroborando com a premissa de que sistemas de reconhecimento automático de sotaques no português brasileiro ainda são um desafio.

Comparação entre modelos com diferentes abordagens para Classificação de Sotaques Brasileiros

Trabalho de Conclusão de Curso

Diego Ribeiro de Almeida
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil
diego.almeida@ccc.ufcg.edu.br

Claudio Elízio Calazans Campelo
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil
campelo@computacao.ufcg.edu.br

ABSTRACT

O sotaque se apresenta como uma das variáveis mais desafiadoras para a eficácia de sistemas de *Automatic Speech Recognition*. Além disso, sua classificação automática possui diversas aplicações potenciais, como a seleção de modelos especializados para *text-to-speech* e *speech-to-text*. Neste trabalho, avaliamos dois modelos de classificação de sotaques a partir da base de dados Braccent, a fim de compará-los com os métodos GMM-UBM, GMM-SVM, *iVector*, CNN 1D, CNN 2D e CNN 1D + LSTM. Os resultados experimentais obtidos demonstram que as abordagens aqui avaliadas apresentam desempenhos consideravelmente abaixo dos reportados na literatura em métricas como acurácia, precisão, revocação, e F1-*score*, corroborando com a premissa de que sistemas de reconhecimento automático de sotaques no português brasileiro ainda são um desafio.

KEYWORDS

Accent Recognition, Brazilian Regional Accents, Audio Classification

1 INTRODUÇÃO

Com base nos inúmeros avanços na modelagem estatística de fala ocorridos nas últimas décadas, sistemas de Reconhecimento Automático de Fala (*Automatic Speech Recognition - ASR*) fazem cada vez mais parte do cotidiano. Evoluiu-se, desde uma simples máquina que responde a um pequeno conjunto de sons, até um sistema sofisticado que responde à linguagem natural falada fluentemente, levando em conta as variáveis estatísticas da língua em que a fala é produzida [6].

O processo em questão é baseado principalmente na identificação de padrões de fala. Dessa forma, para executá-lo, faz-se necessário definir um subconjunto finito de padrões - frases, palavras, fonemas - que seja unidade da fala. No entanto, isso é muito difícil de se executar através de meios automáticos: as unidades apresentam grande variabilidade, dada sua dependência tanto do contexto quanto do falante.

A construção da fala se dá a partir de inúmeras variáveis, tanto pessoais, como o timbre e a velocidade da fala, quanto regionais, como é o caso dos sotaques. Uma determinada variação linguística dentro de um idioma traz consigo variações no tom, bem como diferentes ênfases e extensões nas pronúncias das sílabas. Além disso, determinadas palavras deixam de ser utilizadas com o passar do tempo, bem como outras são criadas, implicando em diferenças no vocabulário.

Considerando estes aspectos, o sotaque se apresenta como uma das variáveis mais desafiadoras para a eficácia de sistemas de ASR [13]. Dessa forma, é comum utilizar-se de sistemas de classificação de sotaques como uma etapa anterior ao modelo de reconhecimento de fala.

Uma vez que cada língua possui suas características únicas de variação de sotaques, é comum que as pesquisas envolvendo esse tópico sejam feitas isoladamente, em idiomas específicos. Observa-se um grande número de trabalhos que estudam os sotaques na língua inglesa [1, 16, 18]. Além disso, existem estudos a respeito do sotaque também em francês [8], mandarim [17], árabe [4], bem como em português, onde se destaca a pesquisa de Batista et al. [9] e de Tostes et al. [15].

Batista et al. [9] propuseram alguns métodos para identificação de sotaques, sendo eles: mistura de gaussianas com modelo universal de fundo, o GMM-UBM (*Gaussian mixture model with universal background model*); *iVector*; e uma variante do GMM-UBM que usa os vetores GMM como entrada para um classificador SVM (GMM-SVM). Tostes et al. [15], por sua vez, também visando a classificação dos sinais de áudios de sotaques brasileiros, propuseram o uso de arquiteturas de redes neurais. Eles desenvolveram modelos utilizando Redes Neurais Convolucionais (*Convolutional Neural Network - CNN*), construídas usando como componentes camadas convolucionais 1D e 2D, gerando os modelos CNN-1D e CNN-2D. Além disso, em determinadas arquiteturas também foram usadas camadas recorrentes do tipo "memória de curto prazo longa" (*long short-term memory - LSTM*), produzindo o modelo CNN 1D + LSTM. Ambos autores utilizaram a base de dados Braccent, produzida na pesquisa de Batista et al. [9], que possui 1.757 áudios, reunindo sete sotaques diferentes: nortista, baiano, fluminense, mineiro, carioca, nordestino e sulista.

Observando o crescimento no número de pesquisas internacionais nessa área, e na intenção de ampliar os trabalhos existentes no Brasil, o presente trabalho visa adaptar, ao português brasileiro, tanto uma estratégia para produção de modelos que já se mostrou efetiva na classificação binária em outro idioma, quanto a aplicação de um método consolidado, a Aprendizagem por Transferência (*Transfer Learning - TL*), para ajuste de um modelo estado-da-arte em ASR. Para tanto, realiza-se um comparativo entre os resultados aqui obtidos e os trabalhos produzidos anteriormente no país, a partir da base Braccent.

A primeira estratégia citada envolve o uso de Regressão Logística, associado a um pré-processamento com Coeficientes Cepstrais em Frequência Mel (*Mel-Frequency Cepstral Coefficients - MFCCs*). Ela é

baseada no trabalho de Honnavalli et al. [7], que utilizou a mesma representação para a classificação binária de sotaques. A segunda, por sua vez, se dá a partir de um treinamento por meio da aplicação de *TL*, utilizando um modelo pré-treinado para *ASR*, o Wav2vec 2.0 [3], como classificador de áudio. Os experimentos utilizaram a base de dados Braccent e seus desempenhos foram avaliados a partir de métricas como a acurácia total, a média macro de precisão (PR), revocação (RE), e F1-*score* (F1), assim como nos trabalhos de Batista et al. [9] e Tostes et al. [15].

O restante deste artigo está organizado da seguinte forma. A Seção 2 apresenta conceitos imprescindíveis à realização da pesquisa. Em seguida, na Seção 3, é apresentada a metodologia utilizada nesta pesquisa. Na Seção 4 são apresentados e discutidos os resultados obtidos. Por fim, a Seção 5 apresenta conclusões para o artigo, e aponta para os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Com o objetivo de comparar diferentes abordagens para classificação de sotaques, foram utilizados neste trabalho tanto algoritmos tradicionais de aprendizagem de máquina, como é o caso da Regressão Logística, quanto modelos estado-da-arte para realização de *Transfer Learning*, como o Wav2vec 2.0. Como etapa de pré-processamento, efetuou-se uma extração de features a fim de gerar *MFCCs* como representações dos áudios. Os *MFCCs*, por sua vez, foram utilizados para treinamento do modelo de Regressão Logística. Nesta seção serão explicados conceitos importantes envolvendo estas abordagens, de forma a esclarecer os experimentos executados.

2.1 Regressão Logística Multiclasse

A Regressão Logística é um tipo de regressão que busca, a partir dos dados em questão, parâmetros para as variáveis independentes de uma função linear, de forma que sua combinação possua o menor erro. É comum que se utilize essa estratégia ao se trabalhar com situações onde a variável dependente é dicotômica. É o caso de Honnavalli et al. [7], que se utilizou dessa estratégia para treinar um modelo binário de classificação do sotaque em inglês entre falantes americanos e indianos.

Em linhas gerais, nesse algoritmo define-se uma função a partir de parâmetros obtidos. Ela será utilizada na função logística que, por sua vez, mapeiará seus valores originais para valores no intervalo entre 0 e 1. A função logística é definida pela fórmula:

$$f(x) = \frac{1}{1 + e^{-x}},$$

onde x é a função linear obtida inicialmente.

Dessa forma, classifica-se uma nova entrada a partir do resultado produzido pela função logística, que será comparado com um limiar. Valores acima do limiar são classificados como pertencentes à classe A, enquanto os abaixo dirão respeito à classe B, por exemplo.

Neste trabalho, o problema se refere à classificação de sete sotaques brasileiros, desta forma, faz-se necessário o uso de uma Regressão Logística Multiclasse (RLM). Para isso, algumas estratégias podem ser utilizadas. Dentre elas, o algoritmo de treinamento pode utilizar o esquema one-vs-rest (OvR), que foi a estratégia utilizada no experimento envolvendo regressão logística deste trabalho, adaptando a estratégia de Honnavalli et al. [7] para um contexto multiclasses. Poderia-se utilizar também a perda de entropia cruzada.

O esquema OvR divide uma classificação multiclasse em um problema de classificação binária por classe, assumindo que cada problema de classificação é independente.

2.2 Mel Frequency Cepstral coefficients

É comum, antes de se treinar um modelo computacional, submeter os dados a um pré-processamento. Dessa forma, as informações podem ser extraídas da melhor forma possível no treinamento, otimizando o desempenho do sistema. O *Mel Frequency Cepstral coefficients* (*MFCC*) é uma destas estratégias. Nele, sumariza-se a distribuição de frequência de um áudio em toda sua extensão, possibilitando uma análise tanto das características de frequência quanto de tempo. Essa representação de áudio é o que permite identificar recursos para serem utilizados em uma tarefa de classificação.

A técnica de extração de características *MFCC* basicamente inclui analisar o sinal a partir de uma janela deslizante, aplicar a Transformada Discreta de Fourier (DFT), obter o logaritmo da magnitude e, em seguida, deformar as frequências em uma escala Mel, seguida pela aplicação da inversa da Transformada Discreta do Cosseno (DCT) [12].

Essa é uma das representações de áudio utilizadas neste projeto. Sua aplicação foi feita como etapa de pré-processamento dos áudios com sotaque, gerando o conjunto de features utilizado na classificação com a Regressão Logística Multiclasse. Este processo foi adaptado a partir do trabalho de Honnavalli et al. [7], que efetuaram o mesmo pré-processamento, mas visando a classificação binária, no idioma inglês, entre os sotaques indiano e americano.

2.3 *Transfer Learning*

Transfer Learning é um método de aprendizagem de máquina. Nele, utiliza-se de um modelo pré-treinado como ponto de partida no treinamento de um modelo para nova tarefa. Ao aplicar *TL* em uma pequena quantidade de dados, a partir de um modelo pré-treinado, pode-se obter um desempenho significativamente maior que o obtido caso o treinamento fosse realizado do zero, com a mesma quantidade de - ou até mais - dados . Dessa forma, para atividades pertinentes, pode-se começar com um modelo pré-treinado que já aprendeu recursos gerais do contexto abordado, como a classificação de objetos, por exemplo.

A classificação de áudio requer a compreensão da estrutura de frequência do sinal acústico. Nela, precisa-se construir um modelo que agrupe conhecimento das características de cada classe de áudio, para que durante a fase de avaliação seja possível que este entenda e classifique um determinado áudio em sua classe correspondente. Nesse sentido, um modelo pré-treinado com dados em um formato adequado, recebendo entradas de áudio e agrupando conhecimento a respeito da estrutura de frequência do sinal, pode representar um bom ponto de partida para a classificação de sotaques, demandando consideravelmente menos dados de treino [11, 14].

A segunda abordagem trazida para comparação neste trabalho faz uso dessa tecnologia, a fim de verificar o uso do Wav2vec 2.0,

um modelo estado-da-arte em ASR, na tarefa de classificação de sotaques brasileiros.

2.4 Wav2vec 2.0

No que diz respeito ao Reconhecimento Automático de Fala o Wav2vec 2.0 é um dos modelos mais atuais. Isso se dá graças ao seu caráter de treinamento: o treinamento auto-supervisionado, que é um conceito consideravelmente novo neste campo. Essa forma de treinamento permite-nos pré-treinar um modelo em dados não rotulados, que são geralmente mais acessíveis. Em seguida, o modelo pode ser ajustado para uma finalidade específica, fazendo uso de um conjunto de dados pertinente para o propósito. Como os trabalhos anteriores mostram, essa forma de treinamento é muito poderosa [5].

O modelo é originalmente treinado em duas fases. A primeira fase é em modo auto-supervisionado, onde o treinamento é efetuado a partir de dados não rotulados visando obter a melhor representação de fala possível, em um processo análogo à geração de embeddings de palavras, no qual se pretende gerar a melhor representação da linguagem natural possível. A principal diferença é que o Wav2vec 2.0 processa áudio ao invés de texto. A segunda fase do treinamento é o ajuste fino supervisionado, durante o qual os dados rotulados são usados para ensinar o modelo a prever palavras ou fonemas específicos, no contexto de ASR.

A primeira fase de treinamento é a principal vantagem deste modelo. Aprender bem uma representação de fala permite a obtenção de bons resultados em ASR a partir de uma pequena quantidade de dados rotulados. É neste contexto que o modelo pode ser utilizado para tarefas de classificação, e, neste caso específico, classificação de sotaques.

Considerando os conceitos apresentados de *Transfer Learning*, o segundo experimento visa classificar os sotaques brasileiros a partir de um modelo pré-treinado do Wav2vec 2.0.

3 METODOLOGIA

Esta seção apresenta a metodologia utilizada para o desenvolvimento dos experimentos realizados neste trabalho. As etapas realizadas foram as seguintes. Obtivemos uma base de dados já existente na área de sotaques brasileiros. Modelamos os dados, preparando-os com seus respectivos pré-processamentos. Executamos os treinamentos e as validações dos modelos treinados.

3.1 Base de Dados

Poucas bases de dados em português brasileiro, com representatividade quanto às variações dos sotaques regionais e suas características fonéticas, estão disponíveis para estudos. Batista et al. [9] construíram para sua pesquisa a base de dados Braccent, a partir de uma aplicação *Web* para captura online das amostras de fala, e esta foi a base utilizada para a realização dos experimentos neste trabalho.

O intuito de Batista et al. [9], ao criar a aplicação, foi alcançar uma quantidade considerável de locutores, de partes distintas do Brasil. Além disso, teve como objetivo retratar casos reais em que o locutor está em diversos ambientes com ruído ou com outras interferências que, de fato, comprometem o processo de reconhecimento. Segundo Batista et al. [9], essa variabilidade é positiva, pois permite que os modelos de reconhecimento de sotaques possam se tornar invariantes às características secundárias dos áudios, que não são relacionadas à fala.

Para a confecção da base Braccent, foram criadas 16 sequências de frases. Elas são foneticamente balanceadas em seu conjunto e não necessariamente apresentam uma coerência semântica. A base de dados contém 1743 amostras de áudio com duração de 8 a 14 segundos. As gravações são identificadas de acordo com o gênero e o sotaque regional falado na locução, sendo 714 amostras do sexo Feminino e 871 amostras do sexo Masculino, correspondentes a 142 locutores com diferentes faixas etárias e níveis de escolaridade, de acordo com a dissertação de Batista et al. [9].

Os falantes são provenientes de diferentes cidades e estados do Brasil, a fim de viabilizar uma análise dos sotaques. Eles pronunciaram, no geral, as 16 frases sequencialmente, num total de aproximadamente três minutos de fala. Em alguns casos, os locutores pronunciaram um subconjunto do total de frases, devido a problemas no acesso à ferramenta.

A base contém sete sotaques brasileiros: nortista, baiano, fluminense, mineiro, carioca, nordestino e sulista. A classificação do sotaque regional de cada áudio foi realizada manualmente pelos alunos do curso de Letras do ano de 2018 do Instituto de Estudos da Linguagem na Universidade Estadual de Campinas.

As informações detalhadas sobre a base de dados Braccent, incluindo divisões por sotaques, estão descritas na Tabela 1.

3.2 Configuração dos Treinamentos

Na implementação dos modelos de classificação de sotaques, foram utilizados 80% dos dados de áudio para o treino e 20% para testes. Os dados também foram separados, com base no gênero do falante,

Tabela 1: Descrição do Número de Gravações da Base de Dados Braccent.

Sotaques	Número de Gravações	Feminino	Masculino
Nortista	27	8	19
Baiano	183	103	80
Fluminense	114	63	51
Mineiro	148	63	85
Carioca	82	47	35
Nordestino	344	153	191
Sulista	845	435	410

em três grupos: Masculinos, Femininos e Geral - que agrupa os dois anteriores. Desejou-se com isso verificar o impacto dessa variável no desempenho da classificação.

Dessa forma, para cada um dos modelos, três diferentes execuções foram efetuadas. A primeira incluindo dados de áudio com vozes femininas, a segunda com dados de áudio com vozes masculinas, e uma terceira sem distinção de gênero.

As identidades dos falantes não foram consideradas na separação entre treino e teste. Isso, de certa forma, é um problema, uma vez que o aprendizado do modelo pode estar relacionado a fatores intríssecos às vozes dos indivíduos, e não necessariamente às caracteristicas dos sotaques. Ou seja, não se tem certeza se o modelo está classificando os falantes, ao invés de seus sotaques.

No entanto, considerando a necessidade de uma comparação adequada com os trabalhos anteriores, que executaram os experimentos dessa forma, este trabalho avalia os modelos a partir de áudios de falantes do mesmo conjunto de dados usado no treinamento.

O *fine-tuning* efetuado a partir do modelo Wav2vec 2.0 utilizou o modelo facebook/wav2vec2-base[1]. Esse modelo foi treinado a partir de dados do Librispeech[2], originalmente em inglês, e foi escolhido principalmente por necessitar de um baixo poder computacional, em comparação com as diversas opções disponíveis, que incluem fine-tunings e otimizações. O ajuste incluiu a execução de 5 épocas, com avaliações recorrentes ao fim de cada uma delas, gerando métricas como *Training Loss*, *Validation Loss* e *Validation Accuracy*, exibidas nas Figuras 1, 2 e 3.

Como hiperparâmetros para o treinamento utilizou-se o tamanho de lote de valor 4, no caso do Wav2vec 2.0. O modelo de Regressão Logística Multiclasse recebeu como parâmetros o tipo da classificação, cuja abordagem escolhida foi a *One-vs-Rest*, bem como o algoritmo de otimização no treinamento, cuja seleção foi o algoritmo *liblinear*, considerando seu bom desempenho em conjuntos de dados pequenos, conforme documentado no *scikit-learn*[3], biblioteca utilizada para carregamento da RLM.

3.3 Extração de features

3.3.1 *Representação a partir de MFCCs.*

Para a geração da representação de áudio utilizada no treinamento da Regressão Logística Multiclasse, efetuou-se a extração de features utilizando *MFCCs*. Para cada arquivo de áudio da base, foram calculados 20 *MFCCs*. Como foi apresentado, *MFCC* é a sumarização da distribuição de frequência de um áudio. Sua implementação foi feita a partir da biblioteca librosa[4] [10], de Python, que segue os passos listados abaixo.

(1) Enquadra-se cada sinal de áudio em janelas curtas de mesmo tamanho, com comprimento de quadro de 2048 amostras, e desliza-se a janela em um comprimento de salto de 512 amostras.
(2) Calcula-se a estimativa do periodograma do espectro de potência para cada janela.
(3) A fim de converter a frequência à escala Mel, aplica-se o banco de filtros Mel (*Mel filterbank*) aos espectros de potência e soma-se a energia em cada filtro, seguindo a seguinte equação:

$$M(f) = 1125 \cdot \ln(1 + (\frac{f}{700})),$$

(4) Obtém-se o logaritmo de todas as energias do *filterbank*. Calcula-se a Transformada Discreta do Cosseno a partir dos dados da etapa anterior.
(5) Selecionam-se os coeficientes *MFCC* necessários. Nesse caso, 20 coeficientes foram selecionados.

Os coeficientes de cada janela de um arquivo de áudio são então concatenados de forma que os recursos *MFCC* extraídos de uma amostra de áudio sejam emitidos na forma de uma matriz com 20 coeficientes para cada janela da amostra, ou seja,

$$MFCC = \begin{bmatrix} c_0 f_0 & c_0 f_1 & c_0 f_2 & \dots & c_0 f_m \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ c_{19} f_0 & c_{19} f_1 & c_{19} f_2 & \dots & c_{19} f_m \end{bmatrix},$$

onde $c_i f_0 \dots c_i f_m$, são os valores de coeficiente i para as janelas de 1 a m.

Essa matriz de 20 x m, que representa os coeficientes *MFCC* para uma amostra de áudio com m quadros, precisa ser transformada em um formato reconhecido pelo modelo de aprendizado de máquina. Os coeficientes *MFCC* são então concatenados sequencialmente, possibilitando que se contrastem os sotaques com base em seus conjuntos de recursos, e achatados em uma matriz unidimensional, conforme o vetor abaixo.

$$M = \begin{bmatrix} c_0 f_0 & \dots & c_0 f_m & \dots & \dots & \dots & c_{19} f_0 & \dots & c_{19} f_m \end{bmatrix}$$

Considerando a execução desse processo para todos os áudios da base, gerou-se as features necessárias ao treinamento do modelo de Regressão Logística Multiclasse, com o objetivo de distinguir um sotaque de outros.

3.3.2 *Representação a partir de Transformers.*

Antes de alimentar o modelo Wav2vec 2.0 com os clipes de áudio do conjunto de dados, também se fez necessário pré-processá-los. Neste caso, a abordagem foi distinta da anterior. Inicialmente, os dados foram carregados a partir da biblioteca librosa, mas sem a aplicação de *MFCCs*. O carregamento do áudio a partir desta biblioteca retorna uma série temporal (*time series*), que, de acordo com o glossário[5] da plataforma é definido como "um sinal de áudio, denotado por *y*, e representado como um *numpy.ndarray* unidimensional de valores de ponto flutuante. *y[t]* corresponde à amplitude da forma de onda na amostra *t*." Quando se fala de amplitude, se diz respeito, neste caso, à mudança na pressão ao redor do microfone ou dispositivo receptor que originalmente captou o áudio.

Uma vez processados os sinais de áudio, utilizou-se um *Feature Extractor* disponível na biblioteca Transformers, disponibilizada pelo *Hugging Face*[6], que normaliza as entradas, colocando-as em um formato adequado e esperado pelo modelo.

[1] https://huggingface.co/facebook/wav2vec2-base
[2] http://www.openslr.org/12
[3] https://scikit-learn.org/stable/
[4] https://librosa.org/doc/main/index.html
[5] http://man.hubwiz.com/docset/LibROSA.docset/Contents/Resources/Documents/glossary.html
[6] https://huggingface.co/docs/transformers/index

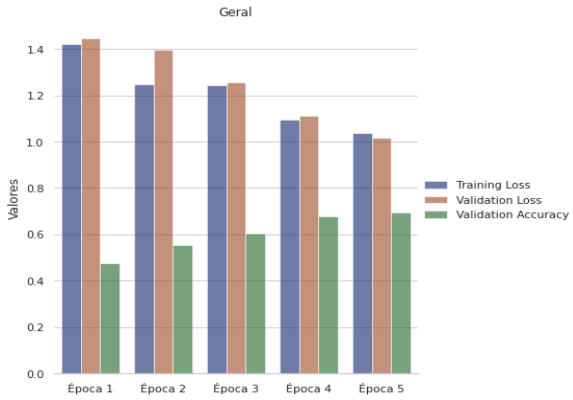


Figura 1: Métricas de treinamento do Wav2vec 2.0 - Cenário Geral (independente de gênero)

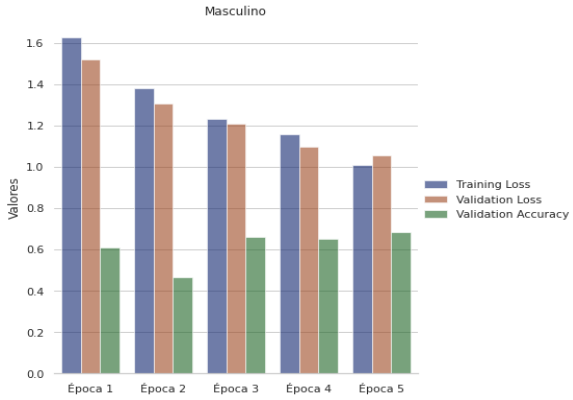


Figura 2: Métricas de treinamento do Wav2vec 2.0 - Cenário Masculino

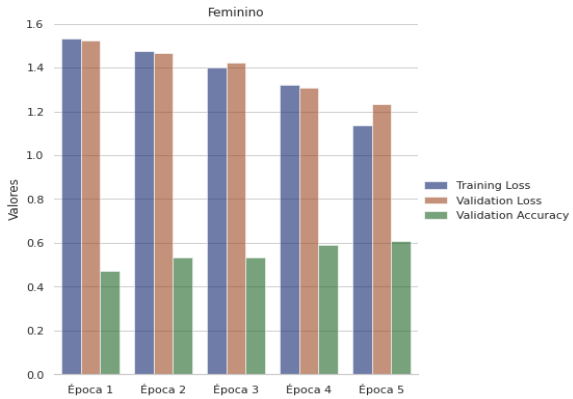


Figura 3: Métricas de treinamento do Wav2vec 2.0 - Cenário Feminino

Para isso, instanciou-se um extrator de features com parâmetros referentes ao truncamento de áudios com durações maiores que 20 segundos, bem como à conversão dos dados de áudio em listas de numpy arrays, formato com o qual o modelo foi treinado previamente. O pré-processamento adiciona também aos dados a informação referente à taxa de amostragem de cada áudio, que deve ser de 16.000Hz.

3.4 Avaliação Experimental

A fim de avaliar os modelos criados anteriormente, foram realizados testes de classificação dos áudios. Os testes foram efetuados com 20% dos dados da base, separados previamente para essa etapa.

Assim como os dados de treino, os dados de teste também foram submetidos às etapas de extração de features respectivas de cada abordagem, segundo as etapas descritas anteriormente. Os dados gerados a partir desse subconjunto de testes, por sua vez, foram submetidos aos modelos, visando a obtenção da classificação em uma entre as sete classes de sotaques às quais os modelos foram expostos.

Como a base de dados é anotada, foi possível então comparar as predições realizadas pelos modelos com a classe real à qual cada áudio pertence, calculando-se então quão bem performam os modelos.

3.5 Métricas de avaliação

Durante a etapa de treinamento dos modelos, os dados de treino são fornecidos para os algoritmos, de forma que eles aprendam os padrões inerentes ao conjunto. Uma vez que esse processo seja finalizado, pode-se fornecer uma nova entrada ao modelo, e ele conseguirá classificá-la entre as classes para as quais foi treinado.

Essa classificação, entretanto, nem sempre é precisa. É comum que o modelo, a depender de inúmeros fatores, classifique uma entrada para uma classe não correspondente à qual ela pertence de verdade. Dessa forma, faz-se necessário o uso de métricas bem definidas para a avaliação dos modelos treinados, frente aos dados que se propõe classificar.

Simplificando o contexto, para fins de exemplificação das métricas, verifica-se que os classificadores podem produzir resultados relativos a quatro categorias: verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN).

A partir dessas categorias, pode-se calcular algumas métricas representativas quanto ao desempenho do modelo. A acurácia é uma dessas métricas. Ela é uma das mais utilizadas para avaliar modelos de classificação. É importante notar, no entanto, que a acurácia pode possuir caráter enviesado em casos onde uma classe com um maior número de amostras eleva o valor da métrica sem que isso reflita fielmente a qualidade da classificação. Além disso, considerando o caráter do conjunto de dados, onde há um desbalanceamento considerável, optou-se por não avaliar os modelos seguindo estritamente essa variável, por mais que ela seja calculada e apresentada.

Outra métrica, que é uma alternativa para contornar o problema da acurácia com os dados desbalanceados, é a F1-*score* [2]. Ela é, na verdade, um balanceamento de duas outras medidas de avaliação, a precisão e a revocação. Em um cenário onde tanto a precisão quanto a revocação sejam iguais a 1, observa-se que o F1-*score* também será igual a 1, refletindo um grande número de acertos na classificação efetuada pelo modelo em questão.

4 RESULTADOS E DISCUSSÕES

A partir dos dados coletados e das etapas de pré-processamento e de treinamento apresentadas, foi possível obter os modelos resultantes dos algoritmos selecionados. Esta seção apresenta tanto resultados específicos dos modelos desenvolvidos, discutindo suas especificidades, quanto o comparativo com os trabalhos relacionados.

Na Figura 4, estão apresentadas as matrizes de confusão produzidas a partir da classificação, pelos modelos treinados, das amostras do conjunto de teste, segundo a etapa de validação apresentada. Nelas, os sotaques são definidos a partir de abreviações, sendo "BA" correspondente ao sotaque baiano, "CA" ao carioca, "FL" ao fluminense, "MI" ao mineiro, "ND" ao nordestino, "NO" ao nortista e "SU" ao sulista.

O valor em uma célula de linha "l" e coluna "p" representa o número de amostras que foram classificadas como um sotaque da classe "p", cuja classe verdadeira é a classe "l", ou seja, as matrizes de confusão apresentam, na diagonal da matriz, a quantidade de amostras classificadas corretamente. A cor da célula também representa um fator relevante na matriz, de forma que varia em uma escala referente à normalização do número de amostras para cada classe verdadeira.

Observa-se nas matrizes de confusão um viés de classificação paras as classes mais frequentes na base de dados. Esse comportamento, como dito anteriormente, faz com que a métrica de acurácia apresente valores superiores às outras métricas, uma vez que, com uma quantidade grande de dados na classe enviesada, o número bruto de acertos será igualmente grande. Esse comportamento também foi observado por Tostes et al. [15] em seus experimentos, embora que em menor dimensão.

Esse comportamento se observa de forma mais clara nas Figuras 4d, 4e e 4f, referentes aos treinamentos com o Wav2vec 2.0. Nas classificações por gênero, percebe-se que praticamente 100% das predições estão contidas nas classes dos sotaques sulista e nordestino. No entanto, ao se adicionar variabilidade no treino, como se observa no cenário geral, em 4d, onde ambos os gêneros estão presentes, percebe-se uma melhor adequação do modelo aos dados. Nota-se um acerto maior em sotaques como o mineiro e o baiano, comparado aos acertos nessas classes em 4e e 4f.

As observações feitas anteriormente, se comparadas com os resultados presentes nas Figuras 4a, 4b e 4c, podem indicar uma vantagem na etapa de pré-processamento que se utiliza da representação de áudio por meio de *MFCCs*, visto que há um viés menor na classificação gerada por essa abordagem, com predições mais dispersas ao longo da matriz. Outra hipótese, ainda, é a de que, por razão do *wav2vec2-base* se tratar de um modelo genérico, sem fine-tuning no idioma em questão, seu ajuste ao cenário pode não ter sido suficiente para uma boa classificação com tão poucos dados, ainda mais quando retirada a diferenciação proveniente do gênero.

Observa-se também uma pequena variação na quantidade de áudios por classe entre um modelo e outro, em uma mesma categoria. Essa variação é gerada graças à aleatoriedade presente na seleção da porcentagem de áudios para teste. Apesar disso, no entanto, percebe-se ainda uma diferença no comportamento dos modelos. Nota-se, na Figura 4d um acerto muito mais evidente que na Figura 4a, no que diz respeito às classes em que o modelo acerta. O modelo de RLM Geral se mostra com menos aptidão para classificar até na classe Sulista, errando cerca de 45% das vezes.

As métricas citadas anteriormente - Acurácia, Precisão, Revocação e F1-*score* - para ambos os modelos, em suas variações de treinamento a partir dos gêneros, são apresentadas na Tabela 2. Nela, constam os resultados tanto dos treinamentos sensíveis a gênero, quanto independentes de gênero. Caso desejado, as métricas por sotaque podem ser derivadas pelo leitor a partir dos valores presentes nas respectivas matrizes de confusão.

Tabela 2: Desempenho dos Sistemas Treinados e Testados na Base de dados BRAccent

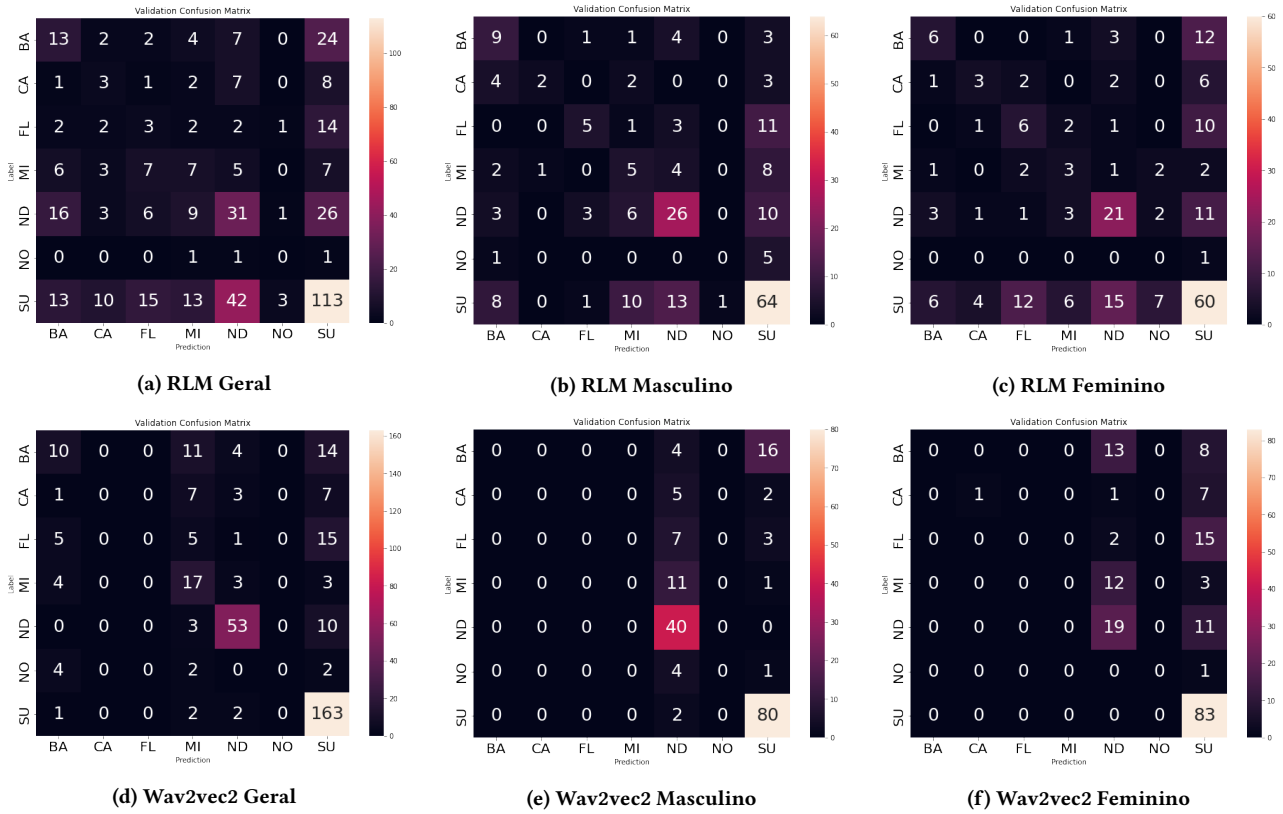
Modelo	Acurácia	Precisão	Revocação	F1-*score*
RLM Geral	0.39	**0.41**	0.34	0.34
RLM Masc	0.50	**0.41**	0.34	0.34
RLM Fem	0.45	0.32	0.30	0.31
Wav2vec2 Geral	**0.69**	0.33	**0.38**	**0.35**
Wav2vec2 Masc	0.68	0.19	0.28	0.22
Wav2vec2 Fem	0.59	0.29	0.25	0.21

Percebe-se, de forma geral, um desempenho consideravelmente semelhante entre os modelos Wav2vec 2.0 e RLM. O primeiro detém os maiores valores em Acurácia, Revocação e F1-*score*. Todos eles no modelo independente de gênero. O RLM se sobressai apenas na Precisão. Isso pode estar atrelado à complexidade consideravelmente maior associada ao primeiro, visto que, além de ser um modelo já pré-treinado, que apresenta uma suposta inteligência prévia em representação de áudio, ele carrega ainda um processo mais robusto de treinamento, que submete o modelo a várias épocas. A estratégia utilizando RLM, por sua vez, consiste de um treinamento simples, e deposita o sucesso da classificação principalmente na geração dos *MFCCs* no pré-processamento.

Partindo para o comparativo entre os modelos treinados e os revisados na literatura, pode-se observar os resultados presentes na Tabela 3. É importante notar que nessa comparação apenas os modelos independentes de gênero aqui desenvolvidos estão sendo considerados, uma vez que essa foi a abordagem utilizada pelos autores analisados. As três primeiras linhas, contendo as métricas dos modelos GMM-UBM, GMM-SVM e *iVector*, foram extraídas do trabalho de Batista et al. [9]. Os autores não disponibilizaram valores de precisão. As três linhas seguintes, por sua vez, foram extraídas do trabalho de Tostes et al. [15], e contém as métricas dos modelos CNN 1D, CNN 2D e CNN 1D + LSTM.

No trabalho de Batista et al. [9] a avaliação dos modelos foi efetuada a partir do método de *10-fold cross validation*. Considerando-se o custo associado à realização deste protocolo de avaliação, utilizamos uma divisão simples dos dados em treino e teste, assim como no trabalho de Tostes et al. [15]. As linhas subsequentes da tabela são referentes aos modelos treinados neste trabalho, com dados de falantes masculinos e femininos, juntos.

Os resultados apontam que nenhuma das abordagens avaliadas levam a resultados superiores aos alcançados em Batista et al. [9] ou Tostes et al. [15]. Percebe-se uma vantagem considerável do modelo CNN 1D + LSTM, que apresenta os maiores valores para acurácia, precisão e revocação. O modelo GMM-UBM, por sua vez, demonstra



(a) RLM Geral **(b) RLM Masculino** **(c) RLM Feminino**

(d) Wav2vec2 Geral **(e) Wav2vec2 Masculino** **(f) Wav2vec2 Feminino**

Figura 4: Matrizes de confusão referentes aos seis treinamentos executados na base Braccent dividida em função do gênero. Em 4a está a Regressão Logistica Multiclasses Geral (RLM Geral), que inclui ambos os gêneros, enquanto 4b e 4c apresentam os resultados dos treinamentos a partir dos dados Masculinos e Femininos, respectivamente. A divisão segue o mesmo esquema em 4d, 4e e 4f, que dizem respeito ao treinamento do modelo Wav2vec 2.0

o maior valor de F1-*score* dentre as abordagens comparadas. Esses resultados são explicados, principalmente, pela dimensão e o custo associado aos processos, nos experimentos aqui comparados. O treinamento da arquitetura CNN 1D + LSTM, por exemplo, foi feito por 250 épocas, em uma estrutura de rede neural complexa e rebuscada. Pode-se dizer o mesmo dos modelos apresentados por Batista et al. [9], que se configuram de forma consideravelmente mais elaborada, com uma demanda de poder computacional e de tempo muito superior às dos experimentos aqui reproduzidos.

É importante observar, no entanto, que assim como os modelos avaliados, os modelos da literatura apresentam características a serem otimizadas. Batista et al. [9] relatam em seu trabalho a necessidade de uma avaliação dos modelos de classificação de sotaques em outro conjunto de dados, que não aquele no qual o modelo foi treinado. Em seus experimentos, a performance dos modelos, com a modelagem de validação *cross-dataset*, caiu consideravelmente. Os valores de F1-*score*, por exemplo, caíram para um intervalo de 20 a 50%, nas diferentes arquiteturas analisadas.

Tostes et al. [15], por sua vez, também relatam a necessidade de futuras avaliações em um cenário *cross-dataset*. Além disso, reitera-se a existência do viés de classificação quanto às classes mais frequentes no conjunto de dados, característica também observada por eles em seus experimentos.

Tabela 3: Comparação de modelos na classificação de sotaques brasileiros.

Modelo	Acurácia	Precisão	Revocação	F1-*score*
GMM-UBM	0.78	-	0.73	**0.91**
GMM-SVM	0.70	-	0.61	0.82
iVector	0.61	-	0.51	0.71
CNN 1D	0.60	0.36	0.36	0.34
CNN 2D	0.73	0.64	0.54	0.57
CNN 1D + LSTM	**0.90**	**0.92**	**0.84**	0.87
RLM Geral	0.39	0.41	0.34	0.34
Wav2vec2 Geral	0.69	0.33	0.38	0.35

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, a base de dados Braccent foi usada para comparar algumas abordagens de classificação de sotaques. Dois modelos foram treinados, o primeiro envolvendo a representação do sinal de áudio com *MFCCs* e o treinamento de uma Regressão Logística Multiclasse, e o segundo se tratando de um ajuste de um modelo pré-treinado para *ASR*, o Wav2vec 2.0, para a tarefa de classificação.

Ambos foram comparados com os métodos GMM-UBM, GMM-SVM e *iVector*, bem como com os métodos CNN 1D, CNN 2D e CNN 1D + LSTM, presentes na literatura. Além disso, verificou-se também o impacto da divisão dos dados por gênero no desempenho dos modelos.

Observou-se que as abordagens avaliadas não obtiveram desempenhos tão bons quanto os existentes na literatura, em um cenário de avaliação utilizando a mesma base de dados. Notou-se também um melhor desempenho dos modelos treinados sem a distinção de gênero, fator associado à melhor adequação do modelo aos dados mais diversificados, envolvendo tanto falantes femininos quanto masculinos.

Em trabalhos futuros, pretende-se avaliar os modelos a partir de outra base de dados, a fim de verificar sua capacidade de generalização. Como verificado por Batista et al. [9] em seus experimentos, o desempenho dos modelos nem sempre generaliza para outras bases, possivelmente pelo fato de que os modelos podem aprender padrões nos áudios que não estejam necessariamente relacionados às características dos sotaques. Almeja-se também estender os experimentos utilizando-se de *Transfer Learning*, selecionando modelos mais robustos e específicos, já pré-treinados e ajustados em uma quantidade maior de dados, idealmente no português.

Por fim, os resultados mostraram que a detecção automática de sotaques regionais brasileiros ainda é um desafio. O reconhecimento automático de sotaques, portanto, ainda é um campo potencialmente prolífico para trabalhos futuros.

6 AGRADECIMENTOS

Inicialmente, agradeço à Nathalia Alves Rocha Batista, ao seu orientador Prof. Dr. Lee Luan Ling e coorientador Prof. Dr. Tiago Fernandes Tavares por cederem acesso à base de dados Braccent para uso neste trabalho.

Agradeço ao meu orientador, professor Cláudio Campelo, pelo suporte, orientação e paciência durante o desenvolvimento dessa pesquisa. Estendo o agradecimento aos demais professores do curso de Ciência da Computação, que contribuíram ativamente na minha formação, e à Universidade Federal de Campina Grande, pela oferta de um curso tão brilhante.

Sou muito grato também às amizades cultivadas ainda no início da graduação, e fortalecidas a cada projeto em grupo. Brener, Iago, Paulo, Raiany, sem vocês não teria como, obrigado!

Finalmente, obrigado família, que mesmo de longe acredita e torce por mim, me dando apoio sempre que necessário. Às minhas irmãs, Danielle Ribeiro e Jakeline Ribeiro, devo isso a vocês. Seu suporte e incentivo, principalmente nos últimos anos, foi substancial, muito obrigado!

REFERENCES

[1] Asad Ahmed, Pratham Tangri, Anirban Panda, Dhruv Ramani, and Samarjit Karmakar. 2019. VFNet: A Convolutional Architecture for Accent Classification. 1–4. https://doi.org/10.1109/INDICON47234.2019.9030363

[2] Aida Ali, Siti Mariyam Shamsuddin, and Anca Ralescu. 2015. Classification with class imbalance problem: A review. 7 (01 2015), 176–204.

[3] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *CoRR* abs/2006.11477 (2020). arXiv:2006.11477 https://arxiv.org/abs/2006.11477

[4] Fadi Biadsy, Julia Hirschberg, and Nizar Habash. 2009. Spoken Arabic Dialect Identification Using Phonotactic Modeling. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*. Association for Computational Linguistics, Athens, Greece, 53–61. https://aclanthology.org/W09-0807

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018). https://doi.org/10.48550/ARXIV.1810.04805

[6] Wiqas Ghai and Navdeep Singh. 2012. Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications* 41, 8 (2012).

[7] Dweepa Honnavalli and Shylaja S S. 2021. *Supervised Machine Learning Model for Accent Recognition in English Speech Using Sequential MFCC Features*. 55–66. https://doi.org/10.1007/978-981-15-3514-7_5

[8] Alexandros Lazaridis, Elie Khoury, Jean-Philippe Goldman, Mathieu Avanzi, Sébastien Marcel, and Philip Garner. 2014. Swiss French Regional Accent Identification. https://doi.org/10.21437/Odyssey.2014-17

[9] Lee Ling, Tiago Fernandes Tavares, Plinio Barbosa, and Nathalia Batista. 2018. Detecção Automática de Sotaques Regionais Brasileiros: A Importância da Validação Cross-datasets. https://doi.org/10.14209/sbrt.2018.335

[10] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

[11] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[12] K. Rao and Anil Vuppala. 2014. *Speech Processing in Mobile Environments*. Appendix A: MFCC Features pages. https://doi.org/10.1007/978-3-319-03116-3

[13] Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie. 2021. The Accented English Speech Recognition Challenge 2020: Open Datasets, Tracks, Baselines, Results and Methods. (2021).

[14] Lisa Torrey and Jude Shavlik. 2009. Transfer Learning. In *Handbook of Research on Machine Learning Applications*. IGI Global, 242–264.

[15] Wagner Tostes, Francisco Boldt, Karin Komati, and Filipe Mutz. 2021. Classificação de Sotaques Brasileiros usando Redes Neurais Profundas. https://doi.org/10.20906/sbai.v1i1.2768

[16] Wei Wang, Chao Zhang, and Xiaopei Wu. 2020. SAR-Net: A End-to-End Deep Speech Accent Recognition Network.

[17] Felix Weninger, Yang Sun, Junho Park, Daniel Willett, and Puming Zhan. 2019. Deep Learning Based Mandarin Accent Identification for Accent Robust ASR. 510–514. https://doi.org/10.21437/Interspeech.2019-2737

[18] Zhan Zhang, Xi Chen, Yuehai Wang, and Jianyi Yang. 2021. Accent Recognition with Hybrid Phonetic Features. https://doi.org/10.48550/ARXIV.2105.01920