Classificação de Sotaques Brasileiros usando Redes Neurais Profundas ⋆

Wagner A. Tostes [*,**] **Francisco A. Boldt** [*] **Karin S. Komati** [*]
Filipe Mutz [*]

[*] *Programa de Pós-Graduação em Computação Aplicada (PPComp)*
Instituto Federal do Espírito Santo (IFES), Campus Serra, ES
(e-mail: {franciscoa, kkomati, filipe.mutz}@ifes.edu.br)
[**] *VixTeam Consultoria e Sistemas SA, ES*
(e-mail: wagner.arca.tostes@gmail.com)

Abstract: The automatic classification of accents has several potential applications, for instance, the identification and authentication of users, forensic investigation tools and the selection of specialized models in text-to-speech and speech-to-text systems. In this work, we propose and evaluate several architectures of artificial neural networks for accent classification. The performance of these architectures in the Braccent dataset was compared with the methods GMM-UBM, GMM-SVM and iVector. Experimental results show that 4 out of 6 architectures achieve better values of accuracy, precision and recall than the previous methods. The best architecture reached 90% of accuracy, with precision, recall and F1-score of 0.92, 0.84 and 0.87, respectively.

Resumo: A classificação automática de sotaques possui diversas aplicações potenciais como a identificação e autenticação de usuários, ferramentas de investigação forense e a seleção de modelos especializados para *text-to-speech* e *speech-to-text*. Neste trabalho, propomos e avaliamos diversas arquiteturas de redes neurais artificiais para classificação de sotaques. A performance das arquiteturas na base de dados Braccent foi comparada com os métodos GMM-UBM, GMM-SVM e *iVector*. Resultados experimentais mostram que 4 das 6 arquiteturas alcançam valores melhores de acurácia, precisão e revocação que os métodos anteriores. A melhor arquitetura alcançou 90% de acerto, com precisão, revocação e F1-*score* de 0.92, 0.84 e 0.87, respectivamente.

Keywords: Accent Recognition; Convolutional Neural Networks; Recurrent Neural Networks;
Palavras-chaves: Reconhecimento de Sotaques; Redes Neurais Convolucionais; Redes Neurais Recorrentes;

1. INTRODUÇÃO

O processo de construção da fala (Brescancini, 2017) é influenciado por diversas características pessoais (e.g., timbre e velocidade da fala), de saúde do aparelho fonador (e.g., rouquidão e cansaço), estado emocional, traços demográficos (e.g., gênero e faixa etária), bases socioeducacionais, além do fator regional, o sotaque. O sotaque se refere à maneira distinta da fala de uma pessoa em uma língua. Diferentes sotaques podem ser identificados por variações no tom, ênfase e extensão da pronúncia de sílabas de uma palavra.

O sotaque é um dos principais fatores variáveis na fala humana, que representa um grande desafio para a robustez dos sistemas de reconhecimento automático de fala (Shi et al., 2021). Assim, é comum que seja feito uma classificação de sotaques, para ser usado em fase anterior ou em conjunto com o modelo de reconhecimento de fala. Além disso, esta classificação pode ser usada para reconhecimento automático de falantes em sistemas de identificação e autenticação, ou ainda em aplicações de investigação forense (Rose, 2002).

Devido às características únicas dos sotaques em diferentes línguas, os trabalhos na área em geral têm foco em línguas específicas. Por ser uma das línguas mais faladas no mundo (Shi et al., 2021), vários trabalhos estudam os sotaques na língua inglesa (Ahmed et al., 2019; Wang et al., 2020; Zhang et al., 2021). Encontram-se artigos sobre sotaque árabe (Biadsy et al., 2009), sotaque francês (Lazaridis et al., 2014), sotaque em mandarim (Weninger et al., 2019), dialetos da Nigéria Salau et al. (2020), dentre outros. Em português, destacam-se o trabalho realizado por Ynoguti (1999) e por Batista et al. (2019).

Uma grande contribuição do trabalho de Batista et al. (2019) foi a elaboração da base de dados Braccent que contém 1.757 áudios de sete sotaques diferentes: nortista, baiano, fluminense, mineiro, carioca, nordestino e sulista.

⋆ Agradecemos à FAPES e a CAPES pelo apoio financeiro dado por meio do PDPG (Parcerias Estratégicas nos Estados da CAPES) (PROCESSO: 2021-2S6CD, TO/nº FAPES: 132/2021). Também agradecemos ao Propós (Programa Institucional de Apoio à Pós-graduação Stricto Sensu) do IFES pela apoio financeiro. Filipe Mutz agradece ao Instituto Federal do Espírito Santo (IFES) por incentivar sua pesquisa via o Programa Pesquisador de Produtividade (PPP) - portaria n. 1072 de 21 de maio de 2020.

O trabalho propôs 3 métodos para identificação de sotaques: (i) mistura de gaussianas com modelo universal de fundo, o GMM-UBM (*Gaussian mixture model with universal background model*); (ii) *iVector* e (iii) uma variante do GMM-UBM que usa os vetores GMM como entrada para um classificador SVM (GMM-SVM). O GMM-UBM alcançou a melhor performance na base de dados Braccent com taxa de reconhecimento de 73% e f1-*score* de 0,91.

Tal como no trabalho de Batista et al. (2019), o problema investigado neste trabalho é a classificação dos sinais de áudio nos sete diferentes sotaques brasileiros. Diferente do trabalho anterior, contudo, a proposta é usar arquiteturas de redes neurais para a tarefa, uma vez que este tipo de modelo têm apresentado boa performance em artigos recentes de classificação de sotaques (Badhon et al., 2021; Weninger et al., 2019; Wang et al., 2020; Salau et al., 2020; Ahmed et al., 2019; Wu et al., 2018).

Foram analisadas 6 arquiteturas diferentes: (a) redes convolucionais 1D (Goodfellow et al., 2016) sobre o sinal cru de áudio, (b) redes convolucionais 2D (Goodfellow et al., 2016) sobre o espectrograma do áudio, (c) redes híbridas construídas usando camadas convolucionais e camadas recorrentes dos tipos *long short-term memory* (LSTM) (Hochreiter and Schmidhuber, 1997) sobre o espectrograma do áudio, (d) redes híbridas construídas usando camadas convolucionais e camadas LSTM em sua versão bidirecional, a *bidirectional long short-term memory* (BiLSTM) (Graves and Schmidhuber, 2005) sobre o espectrograma do áudio, (e, f) são variações das arquiteturas (c) e (d) e se diferenciam por aplicarem duas camadas convolucionais em sequência. Os experimentos com os modelos neurais usaram a base de dados Braccent e a performance das redes neurais foi avaliada usando as métricas acurácia total (AC) e a média macro de precisão (PR), revocação (RE), e F1-*score* (F1), tal como no trabalho de Batista et al. (2018).

O restante do trabalho está organizado da seguinte forma: na Seção 2 são apresentados trabalhos relacionados com a tarefa de classificação de sotaques; a Seção 3 descreve a base de dados; na Seção 4 detalham-se as arquiteturas das redes neurais; na Seção 5 os experimentos, resultados e discussão e; na Seção 6 as conclusões do trabalho.

2. TRABALHOS RELACIONADOS

Wu et al. (2018) propuseram uma arquitetura de redes neurais chamada FreqCNN, construída para o processamento geral de áudio, sem se ater à uma tarefa específica. O sinal de voz é representado como espectrograma e, posteriormente, dividido ao longo do domínio da frequência para formar o espectrograma com distribuição de frequência.

O modelo foi avaliado três cenários: classificação de sotaques usando a "UT-Podcast corpus", identificação de falantes usando a "CHAINS speech corpus" e reconhecimento de emoções na fala usando a base de dados eNTERFACE. NO 'UT-Podcast corpus", os sotaques ingleses são da Austrália, dos Estados Unidos e do Reino Unido, com 1.101 amostras para treinamento e 661 amostras para testes. A revocação média na classificação de sotaques foi de 79,32%, melhores que as abordagens por i-Vector, CNN, AlexNet, VGG-11 e ResNet-18.

Weninger et al. (2019) usaram BiLSTMs e iVectors para a classificação de 15 sotaques de Mandarim. Foi alcançada uma acurácia de 26.09% por amostra de fala e 34.1% por falante. Agrupando os 15 sotaques em 3 grupos relativos a regiões geograficamente próximas, o modelo alcançou revocação média de 66,4%.

Ahmed et al. (2019) propuseram a VFNet (*Variable Filter Net*), uma arquitetura baseada em rede neural convolucional (CNN) que captura uma hierarquia de características. O sinal de áudio bruto é convertido em um espectrograma pela aplicação da *Short-Time Fourier transform* (STFT), pela qual o sinal do domínio do tempo foi convertido para o domínio da frequência. Para melhor convergência e generalização das redes, o espectrograma é dividido em colunas de tamanho 120, e todas elas são rotuladas com os alvos correspondentes. Em seguida, eles são amostrados aleatoriamente do conjunto de dados e, portanto, criam um espectrograma segmentado.

A rede foi treinada em amostras retiradas do "Speech Accent Archive", um repositório de arquivos de áudio que consiste em uma frase falada por mais de 2.000 falantes em mais de 100 sotaques. O conjunto de dados fornece sinais de voz e seus rótulos de acento correspondentes. O corpus contém a fala sem ruídos de 109 falantes com diferentes sotaques que leem a mesma frase na língua inglesa. Foram usados falas de 74 mulheres com o mesmo sotaque para treinar nossas redes, o restante foi usado para testes. A acurácia chegou em 70,33%, maior que a AlexNet e a Resnet.

Salau et al. (2020) construíram um modelo com 6 camadas LSTM seguidas por uma camada convolucional 1D para classificação e três dialetos da Nigéria: Hausa, Igbo e Yoruba. Foi alcançada uma acurácia média de 94,9%.

Wang et al. (2020) apresentaram a SAR-Net, uma arquitetura de aprendizado profundo que adota um mecanismo de aprendizagem multitarefa e consiste principalmente em três módulos: um *encoder* baseado em CNNs e redes recorrentes, um classificador central de sotaques e um classificador auxiliar de reconhecimento de fala. Os áudios foram da base de dados do "Accented English Speech Recognition Challenge 2020" (AESRC2020), em que pessoas de 8 diferentes nacionalidades falam inglês: chinês, indiano, japonês, coreano, americano, britânico, português e russo. Comparando a função de perda *circle-loss*, o SAR-Net é o melhor comparado às propostas da competição.

3. BASES DE DADOS

A base de dados Braccent construída por Batista et al. (2019) foi utilizada para realização dos experimentos. Ela consiste de 1.757 amostras de áudio com duração de 8 a 14 segundos gravada por locutores com sete sotaques diferentes. A distribuição de amostras por sotaque é apresentada na Figura 1. Além dos sotaques, as gravações são também identificadas de acordo com o gênero (714 amostras do sexo feminino e 871 do sexo masculino). Cada áudio possui ainda transcrições fonéticas que podem ser utilizadas para treinamento de sistemas *speech-to-text*.

As gravações foram realizadas pelos próprios locutores, em diversos ambientes e usando diferentes tipos de microfones. Segundo Batista et al. (2019), essa variabilidade é

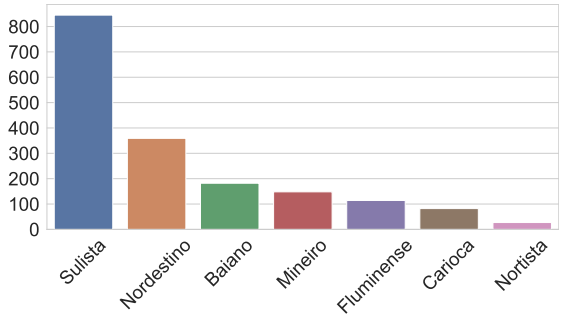


Figura 1. Número de amostras por sotaque na base de dados Braccent.

positiva pois permite que os modelos de reconhecimento de sotaques possam se tornar invariantes à características secundárias que não são relacionadas à fala.

Para a realização deste trabalho, a base de dados foi particionada de forma estratificada em conjuntos de treino (70%), validação (15%) e teste (15%). Como as amostras de áudio da Braccent possuem tamanhos diferentes, eles foram truncados em 195.000 amostras, o que equivale à aproximadamente 12s de áudio. As amplitudes das ondas de som foram normalizadas para o intervalo $[-1, 1]$.

Algumas arquiteturas de redes neurais avaliadas receberam como entrada o espectrograma do áudio. Para calcular o espectrograma foi usada a *Short-Time Fourier Transform* (Yu and Deng, 2016). Os parâmetros utilizados na transformada foram *frame length* igual a 3000 e *frame step* igual a 2000, o que resulta em espectrogramas com 2049 frequências e 97 intervalos de tempo. Apenas os valores absolutos dos espectrogramas foram considerados. Nenhum pré-processamento adicional foi aplicado nem aos áudios nem aos espectrogramas.

4. ARQUITETURAS DE REDES NEURAIS

As arquiteturas de redes neurais foram construídas usando como componentes camadas convolucionais 1D (Conv1D) e 2D (Conv2D), camadas totalmente conectadas (*fully connected* - FC), operadores de *max-pooling* (MP) e operadores de *global max-pooling* (GMP) (Goodfellow et al., 2016). Além destes, também foram usadas camadas recorrentes dos tipos *long short-term memory* (LSTM) (Assis et al., 2019) e *bidirectional long short-term memory* (BiLSTM) (Graves and Schmidhuber, 2005). A última camada de todos os modelos é totalmente conectada e possui um neurônio para cada sotaque. A função de ativação *softmax* é usada para obter a probabilidade da amostra de entrada ser da classe representada por cada neurônio de saída.

As arquiteturas e suas configurações são apresentados na Figura 2. Cada coluna representa um modelo cujo nome e o número total de parâmetros treináveis é apresentado abaixo da coluna. Os blocos representam camadas das redes neurais e as setas indicam que a saída de uma camada é usada como entrada para a próxima. Os números ao lado das setas representam as dimensões dos tensores transportados entre camadas. Cores representam os tipos de camadas: camadas convolucionais são amarelas, totalmente conectadas ocultas são laranja, camadas sem parâmetros treináveis são representadas em azul, camadas totalmente conectadas de saída em roxo e camadas recorrentes em verde. Em camadas convolucionais, o número após o símbolo de arroba representa o número de *kernels*, os valores após a letra K representam as dimensões dos *kernels* e os valores após a letra S representam os *strides*. Nas camadas totalmente conectadas e recorrentes, o valor após o sinal de arroba representa o número de neurônios e o tamanho do estado, respectivamente. As não linearidades aplicadas às saídas das camadas são apresentadas ao final de cada bloco.

Os tamanhos dos *kernels* foram escolhidos de acordo com a proporção aproximada das dimensões do tensor utilizado como entrada para as camadas convolucionais. Quando as entradas possuíam largura maior que altura, por exemplo, os *kernels* também possuem esta característica. Nas duas últimas redes, foram usadas duas camadas convolucionais em sequência no início para permitir a extração de características de mais alto nível e para dar ao modelo a oportunidade de realizar transformações mais complexas sobre os dados de entrada antes de usar as *features* como entrada para a (Bi)LSTM. Esta estratégia é similar àquela utilizada com redes convolucionais 2D operando sobre imagens.

Os demais hiperparâmetros relacionados às arquiteturas como tipo e ordem de camadas, funções de ativação e número de *kernels*/neurônios foram selecionados manualmente em experimentos preliminares usando os conjuntos de treino e validação. Nestes experimentos, os modelos foram treinados por um número pequeno de épocas e a evolução da acurácia de treino e validação foi analisada para verificar que os modelos apresentavam capacidade, generalização e velocidade de convergência razoáveis.

O primeiro modelo na Figura 2 é uma CNN 1D que recebe como entrada o sinal de áudio cru representado por um vetor com 195.000 posições. Os demais modelos operam sobre o espectrograma do áudio com 2049 frequências e 97 intervalos de tempo. O segundo modelo é uma CNN 2D similar em estrutura às redes usadas para reconhecimento de imagens (LeCun et al., 2015).

Enquanto estas duas primeiras arquiteturas seguem o paradigma comumente usado na literatura, os próximos modelos possuem configurações novas e não usuais nas camadas iniciais. Estas mudanças foram decisivas para os resultados positivos alcançados neste trabalho.

A Figura 3 ilustra a arquitetura dos modelos CNN 1D + LSTM/BiLSTM. A primeira camada realiza a operação de convolução 1D sobre faixas completas de frequências do espetrograma. Cada *kernel* desta camada consiste de uma coluna com tantas linhas quanto o número de frequências. O resultado desta operação é um mapa de características com um vetor de características para cada intervalo de tempo do espetrograma. Cada valor do vetor de características representa a ativação gerada por um *kernel* para uma dada faixa de frequências.

Os vetores de características são usados como entrada para uma camada recorrente (LSTM ou BiLSTM) que é responsável por integrar as informações de diferentes instantes de tempo. O estado da camada recorrente após observar todos os vetores de características é usado como entrada para uma sequência de camadas totalmente conectadas que ao final produzem uma distribuição de probabilidades

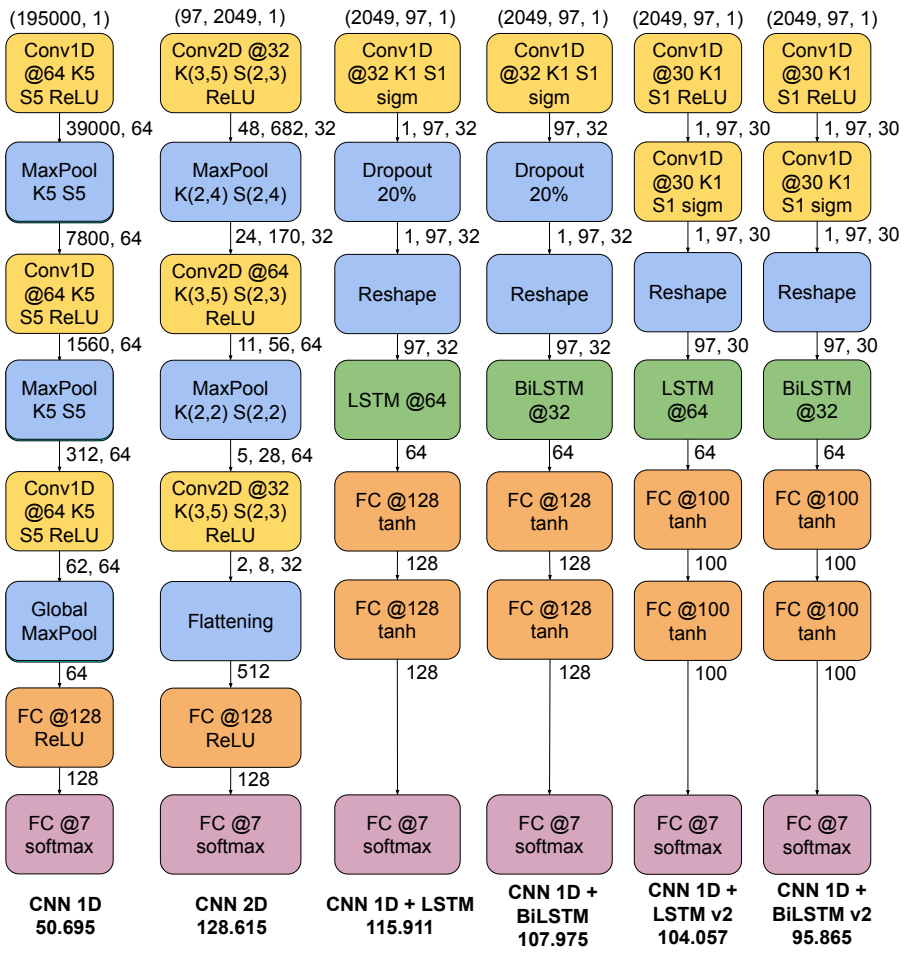


Figura 2. Arquiteturas de redes neurais propostas.

sobre os sete sotaques. Os modelos com indicador *v2* se diferenciam por aplicarem duas camadas convolucionais em sequência.

As arquiteturas de redes neurais foram treinadas usando o otimizador RMSProp para minimizar o erro de classificação dos sotaques (*cross-entropy loss* (Goodfellow et al., 2016)). O treinamento foi feito por 250 épocas usando taxa de aprendizado de 10^{-4}. Ao final do treinamento, são selecionados os pesos com melhor acurácia no conjunto de validação. Esta estratégia é similar à técnica de *early stopping* (Goodfellow et al., 2016). Estes pesos são usados para realizar a avaliação final usando o conjunto de teste. Os códigos foram desenvolvidos usando a linguagem Python e biblioteca TensorFlow (versão 2.4).

5. EXPERIMENTOS

O experimento realizado neste trabalho teve como objetivo comparar a performance das redes neurais com métodos utilizados em trabalhos anteriores na base de dados Braccent. A Figura 4 apresenta as matrizes de confusão obtidas usando as redes para classificar as amostras do conjunto de teste. Nas matrizes, os sotaques foram abreviados como BA (baiano), CR (carioca), FL (fluminense), MN (mineiro), ND (nordestino), NT (nortista) e SU (sulista). A célula da linha l e coluna c contém o número e amostras classificadas como sendo da classe c sendo que a classe verdadeira era l. A coloração das células representa o número de amostras normalizado para cada classe verdadeira.

A acurácia total (AC) e a média macro de precisão (PR), revocação (RE), e F1-*score* (F1) são apresentadas na Tabela 1 (métricas por classe podem ser derivadas pelo leitor a partir das matrizes de confusão). As três primeiras linhas, contendo as métricas dos modelos GMM-UBM, GMM-SVM e *iVector*, foram extraídas do trabalho de Batista et al. (2019) (Apêndice A1, Pág. 107, Tabela 6). Valores de precisão não foram reportados pelos autores. No trabalho, a avaliação foi feita usando o método de 10-*fold cross validation*. O custo de realizar este protocolo de avaliação usando as arquiteturas propostos seria muito alto e, por isto, utilizamos uma divisão simples em treino, validação e teste. As demais linhas da tabela são os resultados para as arquiteturas propostas.

Os resultados apontam que todos as arquiteturas exceto a CNN 1D e a CNN 2D levam à resultados superiores aos alcançados em Batista et al. (2019). A análise da matriz de confusão associada à estes modelos mostra que eles aprenderam a responder a classe mais frequente, isto é, a classe sulista. Os resultados permitem concluir ainda que os modelos que iteram sobre faixas completas de frequências alcançam resultados em todas as métricas exceto o f1-*score* que é surpreendentemente alto para o

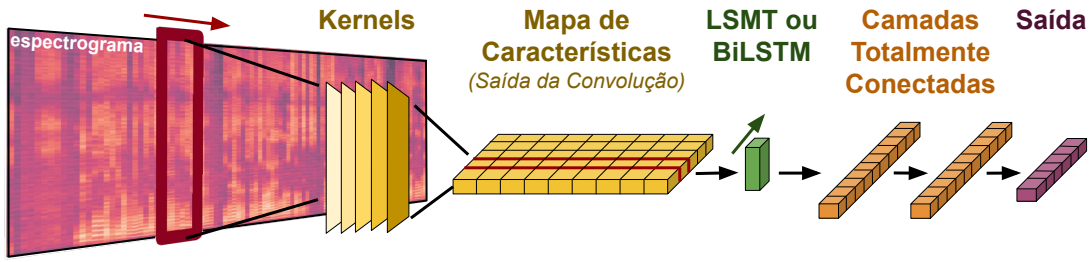
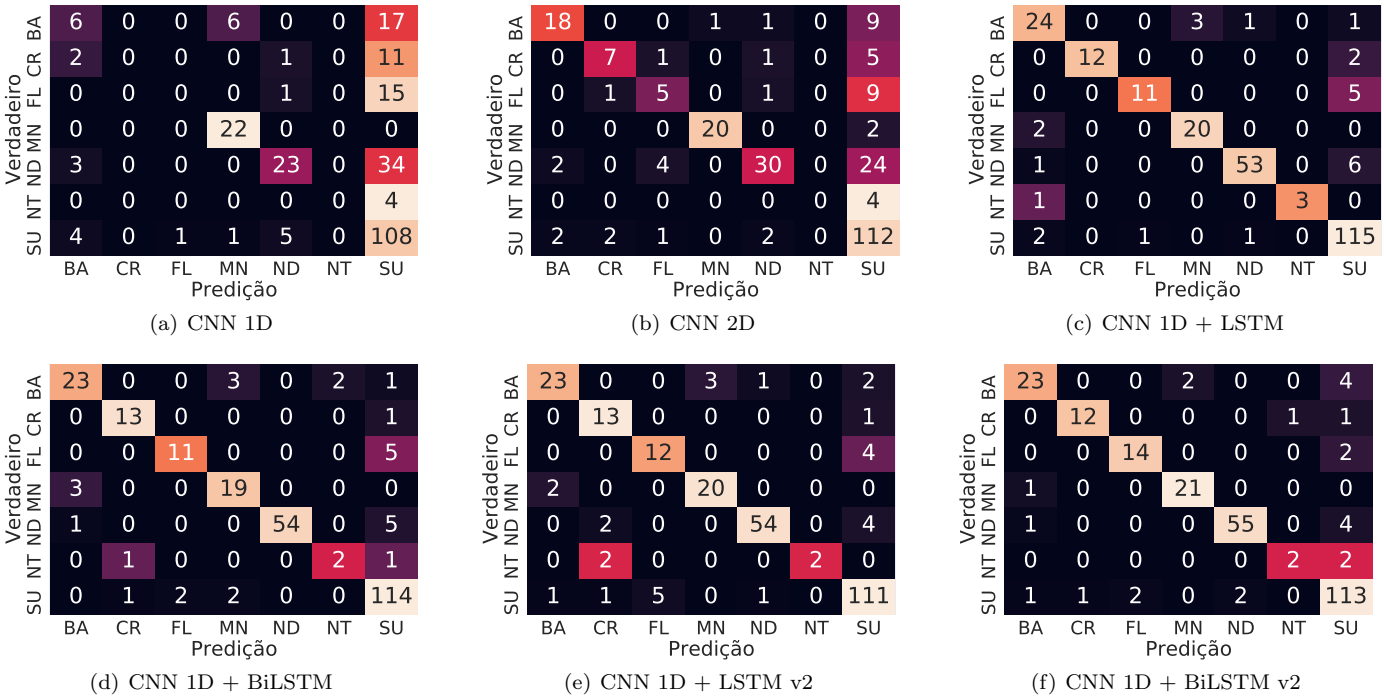


Figura 3. Ilustração do modelo CNN 1D + LSTM/BiLSTM.



(a) CNN 1D (b) CNN 2D (c) CNN 1D + LSTM

(d) CNN 1D + BiLSTM (e) CNN 1D + LSTM v2 (f) CNN 1D + BiLSTM v2

Figura 4. Matrizes de confusão para as diferentes arquiteturas de redes neurais.

Tabela 1. Comparação de modelos na classificação de sotaques brasileiros.

Modelo	AC	PR	RE	F1
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
CNN 1D + BiLSTM	0.89	0.82	0.80	0.81
CNN 1D + LSTM v2	0.89	0.87	0.82	0.83
CNN 1D + BiLSTM v2	**0.91**	0.87	**0.84**	0.85

método GMM-UBM considerando os valores de acurácia e revocação.

Para avaliar a velocidade de aprendizado e a generalização das redes, a Figura 5 compara a evolução de acurácia para os conjuntos de treino e validação durante o treinamento dos modelos. A acurácia da CNN 1D no conjunto de treino indica que a rede ainda estava aprendendo quando o treinamento foi interrompido. Com mais tempo de treinamento ou ajuste de hiper-parâmetros, o modelo talvez alcançasse resultados equivalentes aos dos outros. Efeito similar, mas menos intenso, pode ser observado nos modelos CNN 1D + LSTM e CNN 1D + LSTM v2. Vale notar que a CNN 1D + LSTM alcançou os melhores resultados no conjunto de teste se considerarmos todas as métricas. A curva da acurácia da CNN 2D no conjunto de validação indica que ela sofreu *overfitting*, uma vez que ela convergiu para valores de acurácia inferiores aos demais modelos com performance similar no conjunto de treino. Os modelos CNN 1D + BiLSTM e CNN 1D + BiLSTM v2 apresentaram os crescimentos mais acentuados de acurácia nos conjuntos de treino e validação. Contudo, os resultados no conjunto de teste foram ligeiramente menores que os da CNN 1D + LSTM, o que pode indicar que os modelos precisam de regularização adicional para combater *overfitting*.

6. CONCLUSÃO

Neste trabalho, a base de dados Braccent foi usada para comparar diversas arquiteturas de redes neurais no problema de classificação de sotaques. As redes foram comparadas com os métodos GMM-UBM, GMM-SVM e *iVector* (Batista et al., 2019). Todas as arquitetura exceto a CNN 1D e a CNN 2D alcançaram acurácia e revocação melhores que os *baselines*. A arquitetura que em geral alcançou melhores resultados foi a CNN 1D + LSTM. Este modelo

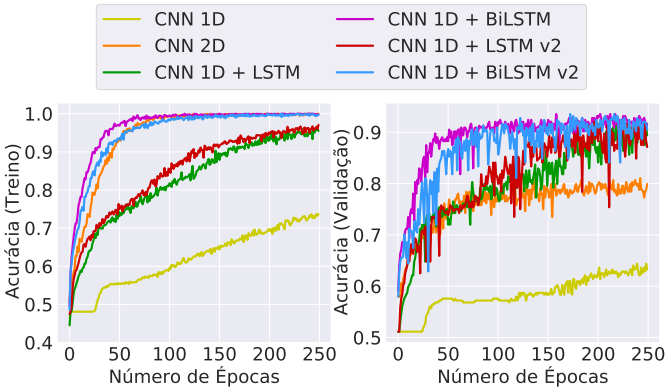


Figura 5. Evolução da acurácia nos conjuntos de treino e validação ao longo do treinamento das redes neurais.

utilizou uma nova abordagem de convolução sobre faixas de frequências do espectrograma de áudio.

Embora os resultados obtidos neste trabalho sejam promissores e demonstrem o potencial de redes neurais artificiais na tarefa de classificação de acentos, em trabalhos futuros avaliaremos de o ganho de performance se mantém em outras bases de dados, inclusive de outros idiomas.

Como apontado por Batista et al. (2019), a performance dos modelos nem sempre generaliza para outras bases de dados. Uma explicação para este fenômeno é que os modelos podem aprender padrões nos áudios que não necessariamente são relacionados aos sotaques. Para avaliar se este efeito acontece com redes neurais artificias, em trabalhos futuros realizaremos avaliações *cross-datasets*.

Por fim, em estudos futuros analisaremos os pesos aprendidos pelo modelo CNN 1D + LSTM com o objetivo de compreender o porquê de sua performance superior. Técnicas de explicação de redes neurais serão utilizadas para investigar as características de frequência extraídas pelos *kernels* e avaliar se estes aprendem a detectar padrões de frequências característicos para cada sotaque. Esta análise pode levar à novos conhecimentos tanto na área de análise da fala quanto para o desenvolvimento de novas técnicas de classificação automática de sotaques.

7. AGRADECIMENTOS

Os autores agradecem à Nathalia Alves Rocha Batista, ao seu orientador Prof. Dr. Lee Luan Ling e coorientador Prof. Dr. Tiago Fernandes Tavares por ter dado acesso à base de dados para este trabalho.

REFERÊNCIAS

Ahmed, A. et al. (2019). Vfnet: A convolutional architecture for accent classification. In *2019 IEEE 16th India Council International Conference (INDICON)*, 1–4. IEEE.

Assis, E. et al. (2019). redição de ações judiciais de consumo de energia não registrado usando a rede LSTM. In *Anais do 14º Simpósio Brasileiro de Automação Inteligente (SBAI 2019)*, 1178–1183.

Badhon, S.S.I. et al. (2021). Bengali accent classification from speech using different machine learning and deep learning techniques. In *Soft Computing Techniques and Applications*, 503–513. Springer.

Batista, N.A.R. et al. (2019). *Estudo sobre identificação automática de sotaques regionais brasileiros baseada em modelagens estatísticas e técnicas de aprendizado de máquina*. Master's thesis, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas (Unicamp).

Batista, N.A.R. et al. (2018). Detecção automática de sotaques regionais brasileiros: A importância da validação cross-datasets. In *Anais do XXXVI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT)*, 939–944. Sociedade Brasileira de Telecomunicações, Campina Grande, PB.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, 53–61.

Brescancini, C.R. (2017). Sobre o potencial discriminante das propriedades de voz/fala na tarefa de comparação de locutores: um estudo de caso. *Revista da Anpoll*, 1(42), 12–27.

Goodfellow, I. et al. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6), 602–610.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Lazaridis, A. et al. (2014). Swiss french regional accent identification. In *Odyssey*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.

Rose, P. (2002). *Forensic speaker identification*. cRc Press.

Salau, A.O., Olowoyo, T.D., and Akinola, S.O. (2020). Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model. In *Advances in Computational Intelligence Techniques*, 1–16. Springer.

Shi, X. et al. (2021). The accented english speech recognition challenge 2020: open datasets, tracks, baselines, results and methods. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6918–6922. IEEE.

Wang, W., Zhang, C., and Wu, X. (2020). Sar-net: A end-to-end deep speech accent recognition network. *arXiv preprint arXiv:2011.12461*.

Weninger, F. et al. (2019). Deep learning based mandarin accent identification for accent robust ASR. In *Proceedings of INTERSPEECH*, 510–514.

Wu, Y., Mao, H., and Yi, Z. (2018). Audio classification using attention-augmented convolutional neural network. *Knowledge-Based Systems*, 161, 90–100.

Ynoguti, C. (1999). *Reconhecimento de Fala Contínua Utilizando Modelos Ocultos de Markov*. Ph.D. thesis, Faculdade de Engenharia Elétrica, Unicamp.

Yu, D. and Deng, L. (2016). *Automatic Speech Recognition*. Springer.

Zhang, Z. et al. (2021). Accent recognition with hybrid phonetic features. *arXiv preprint arXiv:2105.01920*.