Reconhecimento de Emoções através da Fala utilizando Rede Neural Convolucional

Guilherme de S. Peixoto[1], **José E. B. de S. Linhares**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Campus Manaus Zona Leste – Manaus, AM – Brasil

guilhrmpeixoto@gmail.com, breno.linhares@ifam.edu.br

Abstract. *The rapid development of Artificial Intelligence has improved emotion recognition systems through speech using neural networks. This article aims to develop a model trained from a convolutional neural network for the recognition of emotions in Brazilian Portuguese (pt-BR) audio. The validation of the model was carried out based on tests using a data set created with audios that cover the linguistic varieties of Brazil. The tests showed an unsatisfactory performance for a classifier, but it was possible to identify that existing linguistic variations in a language can affect the overall performance of a model.*

Resumo. *O rápido desenvolvimento da Inteligência Artificial aprimorou os sistemas de reconhecimento de emoções através da fala utilizando redes neurais. Este artigo tem como objetivo o desenvolvimento de um modelo treinado a partir de uma rede neural convolucional para o reconhecimento de emoções em áudios do idioma português do Brasil (pt-BR). A validação do modelo foi realizada a partir de testes utilizando um conjunto de dados criado com áudios que abrangem as variedades linguísticas do Brasil. Os testes apresentaram um desempenho insatisfatório para um classificador, porém foi possível identificar que as variações linguísticas existentes em um idioma podem afetar o desempenho geral de um modelo.*

1. Introdução

O rápido desenvolvimento em Inteligência Artificial (IA) e os avanços recentes em diversas tecnologias proporcionaram a melhoria de sistemas automatizados. Com o avanço da tecnologia de automação, a vida do ser humano tem se tornado cada vez mais confortável, acessível e menos exigente em todos os campos [Abdulraheem et al. 2020]. Nestes sistemas, é fundamental que as máquinas possam reconhecer as necessidades humanas para prover melhores soluções automatizadas. Desse modo, realizaram-se diversas pesquisas com o objetivo de propor e desenvolver sistemas de reconhecimento automático de emoções utilizando Rede Neural Artificial (RNA).

Tradicionalmente, em uma RNA as emoções podem ser analisadas através de diversas modalidades como, por exemplo, pela fala, expressões faciais e sinais psicológicos [Rumagit et al. 2021]. Entretanto, o método de Reconhecimento de Emoções através da Fala (REF) tem se tornado gradativamente o principal método nos estudos de reconhecimento de emoções. No processo de comunicação e expressões humanas, a fala não contém apenas informações semânticas, mas também evidenciam informações ricas como as emoções do falante [Lu 2022]. Comparando com outras modalidades, os sinais

da fala se mostram muito mais versáteis [Rumagit et al. 2021]. O REF é uma parte da computação afetiva, que tem como objetivo a análise de emoções na voz humana. Ou seja, nesse ramo de pesquisa ocorre a tentativa de fazer com que o computador reconheça as emoções expressas em um determinado enunciado [Atmaja et al. 2022].

Porém, a generalização dos modelos desenvolvidos ainda é um problema no reconhecimento de emoções. Segundo [Atmaja et al. 2022], a forma de contribuição da informação emocional de um idioma específico pode diferir de outros, sendo necessário a investigação do efeito da informação linguística em diferentes idiomas para acelerar a implementação dos estudos do REF multilíngue. Desse modo, é necessário o treinamento de um modelo com conjuntos de dados diferentes e de idioma distintos para encontrar os ajustes necessários para o reconhecimento de emoções em um idioma específico. Por entender a necessidade de encontrar os ajustes específicos em um idioma e ainda existir uma escassez de modelos treinados com conjunto de dados no idioma português, está sendo desenvolvido, neste trabalho, um modelo treinado utilizando a arquitetura de uma rede neural convolucional, do inglês *Convolutional Neural Network* (CNN), com um conjunto de dados no idioma português do Brasil (pt-BR) para o seu treinamento, em que existe uma diversidade linguística por região, utilizando um método não licenciado e disponível em repositórios de hospedagem de código.

Estudos de sistemas de REF visam criar métodos eficientes e em tempo real para detectar as emoções de usuários de telefones celulares, operadores e clientes de *call center*, e muitos outros usuários de comunicação humano-máquina [Lech et al. 2020], além de melhorar sistemas de automação residencial. Nos sistemas de automação residencial, onde ocorre o controle da iluminação, temperatura, sistemas multimídia e eletrodomésticos conectados em uma infraestrutura comum, a adaptação com base nas ações do usuário e dos arredores é necessária para tornar estes sistemas artificialmente inteligentes, possibilitando prever ações futuras e também minimizando a interação do usuário. Porém, os sistemas tradicionais não são tão eficazes neste cenário de adaptação [Jaihar et al. 2020].

Sistemas automatizados com base na emoção do usuário tornam a automação residencial mais artificialmente inteligente [Jaihar et al. 2020]. Máquinas que são capazes de entender emoções podem prover respostas emocionais mais apropriadas e exibem personalidades emocionais [Lech et al. 2020]. Assim, se torna necessária a evolução natural de técnicas e algoritmos neste ramo de pesquisa. O trabalho tem como justificativa, portanto, os benefícios que estes sistemas trarão à sociedade, estimulando a maior interação humano e máquina, e para isto, é fundamental o reconhecimento de emoções.

O artigo está organizado da seguinte forma: A Seção 2 descreve os principais trabalhos relacionados. A Seção 3 explica cada etapa do sistema desenvolvido. A Seção 4 descreve os experimentos realizados e apresenta os resultados obtidos em cada experimento. Na Seção 5, é feita a conclusão do trabalho e a apresentação dos trabalhos futuros.

2. Trabalhos Relacionados

Os autores [Abdelhamid et al. 2022] desenvolveram uma abordagem que consiste em um algoritmo para aumento de dados, uma rede neural que combina as arquiteturas CNN e *Long-Short Term Memory* (LSTM) e apresentaram uma abordagem de otimização dos parâmetros da rede neural. O algoritmo de aumento de dados produz amostras de treina-

mento adicionais colocando cuidadosamente frações de ruído nas amostras limpas. Para a extração de características, foi utilizado o espectro log-Mel. Para validar a metodologia, os autores utilizaram quatro conjuntos de dados: RAVDESS, Emo-DB, SAVEE e IEMOCAP. O modelo alcançou $99,47\%$, $99,76\%$, $99,50\%$ e $98,13\%$ de precisão de classificação com base nesses conjuntos de dados, respectivamente.

[Singh and Prasad 2023] adaptaram um método de reconhecimento da emoção através da fala para que o modelo fosse capaz de diferenciar os gêneros, ou seja, o modelo base reconhece apenas as emoções e no modelo proposto para cada emoção é considerado separadamente os gêneros. O modelo é capaz de identificar oito emoções: surpresa, felicidade, desgosto, tristeza, raiva, calmo, medo e neutro. A arquitetura utilizada é a CNN, com o método *Mel Frequency Cepstral Coefficients* (MFCC) para realizar as extrações de características. O conjunto de dados utilizado no estudo foi o RAVDESS, construído com amostras de áudio no idioma Inglês. A acurácia do modelo proposto foi de $72,07\%$.

[Mustaqeem and Kwon 2020] apresentaram um novo modelo baseado na arquitetura CNN com intuito de diminuir a complexidade computacional, utilizando uma estrutura simples com poucas camadas de convolução. O modelo reduz a amostragem dos mapas em vez de agrupar as camadas, feito especialmente para problemas de sistemas REF utilizando espectrogramas. Os autores adotaram uma estratégia de pré-processamento que elimina ruídos seguido pela remoção de partes silenciosas dos conjuntos de dados IEMOCAP e RAVDESS. A rede neural obteve uma acurácia de $81,75\%$ utilizando o conjunto de dados IEMOCAP, já com o conjunto RAVDESS a acurácia alcançada foi de $79,50\%$.

O trabalho desenvolvido por [Sun 2020] propõe um modelo de arquitetura CNN com a habilidade de consumir apenas os dados brutos dos conjuntos de dados sem a necessidade de utilização de algum método de extração de *features*. O modelo é treinado para classificar as emoções e diferenciar os gêneros de forma independente, desse modo, o algoritmo é dividido em três etapas. A primeira etapa é responsável por treinar o modelo diferenciando o gênero, a segunda etapa o modelo é treinado para reconhecer a emoção, e na última etapa o classificador combina as duas informações em uma emoção final. Este estudo utilizou três conjuntos de dados: um conjunto de dados no idioma Chinês, CASIA; um conjunto de dados no idioma Inglês, IEMOCAP; e um conjunto de dados no idioma Alemão, Emo-DB. O modelo alcançou $84,26\%$, $71,50\%$ e $90,30\%$ de acurácia com base nos conjuntos de dados apresentados, respectivamente.

3. Metodologia

Neste artigo, desenvolveu-se um modelo treinado de rede neural convolucional para reconhecimento de emoções através da fala em áudios do idioma português do Brasil (pt-BR). A pesquisa foi realizada seguindo duas fases, sendo primeira a fase de Treinamento da Rede e depois a fase de Testes da Rede. Conforme apresentado no diagrama em blocos da Figura 1, na fase de Treinamento da Rede, realizou-se o pré-processamento dos áudios presentes no conjunto de dados de entrada escolhido e extraídas as características acústicas da fala para realizar o treinamento do modelo da CNN. Na fase de Testes da Rede, foram extraídas as características acústicas dos áudios do conjunto de entrada para realizar a predição dos dados, utilizando o modelo treinado na fase anterior.

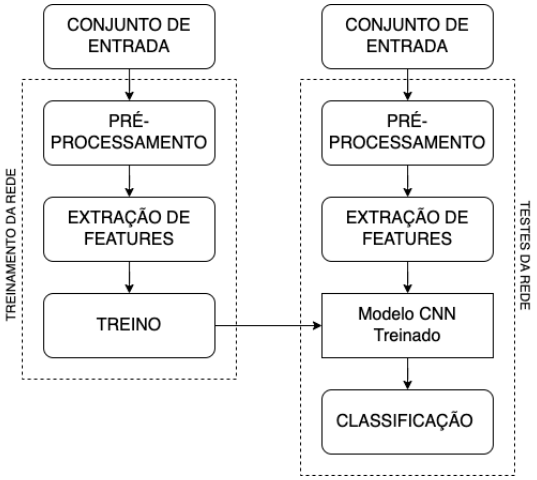

Figura 1. Diagrama em blocos das etapas da metodologia proposta.

3.1. Conjunto de entrada

Cada conjunto de dados escolhidos foram organizados, sendo atribuídos classes para cada áudio e, conforme cenários estabelecidos nos experimentos, as classes utilizadas sofrem adaptações. Para a fase de Treinamento, o conjunto de dados de entrada utilizado é o emoUERJ e, para a fase de Testes, foi criado um conjunto de dados, que tem como característica a diversidade linguística presente em cada região do Brasil.

O conjunto de dados emoUERJ foi desenvolvido pela Universidade do Estado do Rio de Janeiro com o objetivo de suprir a carência de conjuntos em português existentes e utilizar no desenvolvimento de modelos de REF [Bastos Germano et al. 2021]. O conjunto foi elaborado a partir de dez frases, disponibilizado para oito atores, sendo divididos igualmente entre os gêneros e cada ator teve a liberdade de escolher a frase para a gravação dos áudios em quatro emoções: alegria, raiva, tristeza ou neutra. Finalizado o processo de geração de áudios, totalizou-se 377 áudios distribuídos, conforme apresentado na Tabela 1.

Tabela 1. Distribuição das emoções no conjunto emoUERJ.

Emoção	Homem	Mulher	Total
Felicidade	41	50	91
Raiva	52	42	94
Neutro	46	46	92
Tristeza	51	49	100
Total	190	187	377

Para os testes do modelo treinado, foi utilizado um conjunto de dados criado a partir de vídeos coletados no *YouTube*, com a classificação validada de forma pessoal para as três seguintes emoções: raiva, felicidade e neutra. Das cinco regiões brasileiras, foram selecionadas quatro regiões, em razão das dificuldades encontradas na seleção de vídeos na seguinte região: a região Norte. Cada região apresenta uma particularidade linguística, desse modo, foi escolhida uma variação específica de cada região. As variações linguísticas selecionadas foram:

- **Baiano:** variação linguística presente na região Nordeste;
- **Mineiro:** variação linguística presente na região Centro-Oeste;
- **Carioca:** variação linguística presente na região Sudeste;
- **Gaúcho:** variação linguística presente na região Sul.

A distribuição das emoções para os áudios coletados por sexo está apresentado na Tabela 2, em que apenas a variação linguística gaúcha feminina na emoção raiva tem 6 áudios coletados, enquanto que os restantes apresentam 10 áudios para cada emoção, variação linguística e gênero.

Tabela 2. Distribuição das emoções no conjunto de testes.

Emoção	Homem	Mulher	Total
Felicidade	40	40	80
Raiva	40	36	76
Neutro	40	40	80
Total	120	116	236

3.2. Pré-processamento

O bloco de Pré-processamento é realizado em duas etapas: carregamento de áudios e aplicação de efeitos. Para prosseguimento dos blocos seguintes, é necessário carregar, como primeira etapa, um arquivo de áudio como uma série temporal de ponto flutuante. A segunda etapa é opcional e é realizada conforme as estratégias adotadas na aplicação da técnica de aumento de dados na fase de Treinamento. Como exemplos de efeitos, tem-se a aplicação de ruído (*noise*), alongamento (*stretch*), deslocamento (*shift*) e afinação (*pitch*). Define-se a escolha do efeito a partir da experimentação e da análise da contribuição para o aumento da precisão do modelo.

3.3. Extração de *Features*

O bloco de Extração de *Features* é uma etapa importante para o desenvolvimento e treinamento de uma rede neural. O método escolhido para extrair os coeficientes de um espectro de áudio é o MFCC. Após a extração, o algoritmo fornece um vetor com a sequência MFCC, porém é necessário realizar a limpeza dos dados, adequando o tipo dos dados do vetor para o treinamento e testes do modelo. A limpeza é feita substituindo os valores nulos presentes no vetor pelo valor zero.

3.4. Treino

Com todos os vetores de características obtidos no bloco de Extração de *Features*, organiza-se o conjunto de treinamento, sendo dividido em duas partes: uma porcentagem dos dados são utilizados para o treinamento do modelo e o restante para a validação.

O modelo é criado, conforme apresentado na Tabela 3, que mostra os tipos de camadas e funções de ativação. Para a análise de série temporal, é recomendada a utilização da camada do tipo convolução 1D. A entrada da primeira camada de convolução tem o formato (*batch-size*, *input-dim*), onde o primeiro parâmetro é o tamanho do conjunto de treinamento, ou seja, o número de vetores e o segundo é a dimensão do vetor de características. A rede tem 8 camadas de convolução e 2 de agrupamento, na qual a operação

utilizada é a *max polling*. A função de ativação utilizada após as camadas de convolução é a função *Rectified Linear Unit* (ReLU), enquanto na última camada da rede é utilizada a função de ativação *Softmax*. Na arquitetura do modelo, também são utilizadas 2 camadas de *dropout* e 2 camadas de *batch-normalization*. As camadas de normalização em lote, do inglês *batch-normalization*, facilitam a convergência durante o treinamento e fornecem um efeito de regularização que ajuda a evitar o *overfitting* [Lopez-Martin et al. 2020]. Também com o objetivo de evitar o *overfitting*, as camadas que aplicam a técnica *dropout* removem temporariamente alguns neurônios, juntamente com todas as suas conexões de entrada e saída. Portanto, com essa técnica a rede é treinada de diversas formas tornando-a mais robusta [Srivastava et al. 2014].

Tabela 3. Tipos de camadas e funções de ativação da CNN utilizada.

Camadas e Funções de Ativação	Tipo
conv1d_1	Conv1D
activation_relu_1	Activation
conv1d_2	Conv1D
batch_normalization_1	BatchNormalization
activation_relu_2	Activation
dropout_1	Dropout
max_pooling1d_1	MaxPooling1D
conv1d_3	Conv1D
activation_relu_3	Activation
conv1d_4	Conv1D
activation_relu_4	Activation
conv1d_5	Conv1D
activation_relu_5	Activation
conv1d_6	Conv1D
batch_normalization_2	BatchNormalization
activation_relu_6	Activation
dropout_2	Dropout
max_pooling1d_2	MaxPooling1D
conv1d_7	Conv1D
activation_relu_7	Activation
conv1d_8	Conv1D
activation_relu_8	Activation
flatten	Flatten
dense	Dense
activation_softmax	Activation

Após a criação do modelo, é preciso fazer a configuração necessária para o treinamento, utilizando o método de compilação. Basicamente, definem-se 3 quesitos para o modelo: o otimizador que a rede vai utilizar para aprender o problema; a função de perda (*loss*); e a métrica a ser avaliada pelo modelo durante o treinamento e testes. Utilizou-se o *Stochastic Gradient Descent* (SGD) ou Gradiente Descendente Estocástico como otimizador, o *Categorical Cross Eentropy* ou Entropia Cruzada Categórica como função de perda e a Acurácia para a métrica.

Para evitar o desperdício de tempo aguardando um modelo que não apresenta evolução na taxa de aprendizagem, foi configurada a seguinte lista para o treinamento do modelo:

- ***ReduceLROnPlateau:*** tem como objetivo reduzir a taxa de aprendizagem quando atingir um platô. A métrica monitorada para identificar um platô é a perda do conjunto de validação, um fator aplicado de $0,9$ com uma paciência de 20, significa que aguarda 20 épocas sem melhora até que a taxa de aprendizagem seja reduzida aplicando o fator.
- ***EarlyStopping:*** interrompe o treinamento quando uma determinada métrica não é atingida. A métrica de desempenho do modelo escolhida é a perda do conjunto de validação, com uma paciência de 50 épocas. Caso não ocorra uma diminuição no valor da perda durante 50 épocas, o treinamento do modelo é interrompido.
- ***ModelCheckPoint:*** o modelo que apresenta a menor taxa de perda para o conjunto de testes é salvo para ser aplicado no conjunto de testes.

3.5. Classificação

Com o modelo treinado e os vetores de características do conjunto de testes, pode-se gerar as previsões de saída para as amostras de entrada utilizando o mecanismo de predição. Após a predição sobre o conjunto de testes, o resultado obtido será expresso em uma coluna com as predições. Desse modo, é possível comparar a coluna de predições com a coluna de valores reais, com o objetivo de avaliar o modelo treinado. Para auxiliar na avaliação do modelo, foram utilizadas as seguintes métricas:

- **Acurácia:** percentual de acertos sobre todos os valores;
- **Precisão:** é a capacidade do classificador de não rotular uma amostra negativa como positiva;
- **Sensibilidade:** é a capacidade do classificador de encontrar todas as amostras positivas.
- ***F1-score*:** é uma maneira de visualizarmos as métricas Precisão e Sensibilidade juntas, calculando a média harmônica.

4. Resultados e Discussões

4.1. Procedimentos experimentais

A estratégia adotada para a realização dos experimentos foi escolher, primeiramente, os efeitos a serem utilizados na aplicação da técnica de aumento de dados. Com os efeitos escolhidos, foram realizados os experimentos em dois cenários específicos: treinamento do modelo com a diferenciação do gênero e o treinamento sem a diferenciação do gênero. Em ambos experimentos os parâmetros utilizados foram:

- **Conjunto de treinamento:** 75% dos dados;
- **Conjunto de validação:** 25% dos dados;
- **Taxa de aprendizagem:** $0,00005$.

No experimento 1, na qual não existe uma diferenciação entre os gêneros, as classes foram definidas de acordo com as emoções presentes no conjunto de treinamento: alegria, raiva, tristeza e neutra. Com a escolha das classes, é necessário realizar a estratégia de definição dos efeitos na aplicação da técnica de aumento de dados, com o

objetivo de obter o melhor desempenho do modelo. Foram realizados alguns testes e o aumento de dados com a aplicação dos efeitos *noise*, *shift* e *shift* resultaram nos melhores desempenhos. Na Figura 2, apresenta-se a curva de perda e a curva da acurácia durante o treinamento neste cenário do experimento. Ao fim do treinamento, a acurácia para este experimento foi de $91,51\%$.

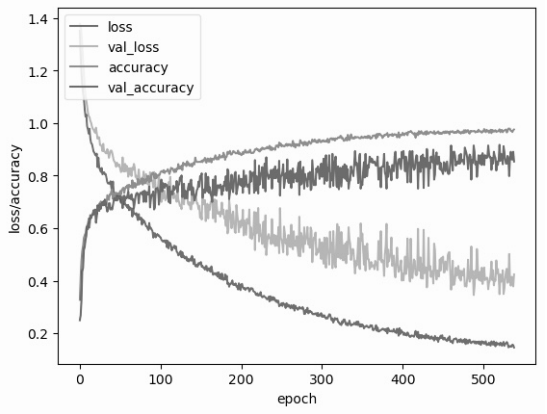

Figura 2. Desempenho do modelo no experimento 1 após a definição dos efeitos.

O experimento 2 tem como objetivo verificar qual a influência da diferenciação do gênero no treinamento de um modelo. Para isso, as classes foram definidas levando em consideração os gêneros masculino e feminino, do conjunto de treinamento. Desse modo, soma-se 8 classes. Para a realização do próximo experimento, foi definida a utilização dos mesmo efeitos do experimento 1. Isto representa um aumento no conjunto de treinamento de 377 para 1507 arquivos. Na Figura 3, observa-se a curva do valor da perda e a curva da acurácia durante o treinamento do modelo neste novo cenário do experimento. Ao fim do treinamento, a acurácia para este experimento foi de $94,70\%$.

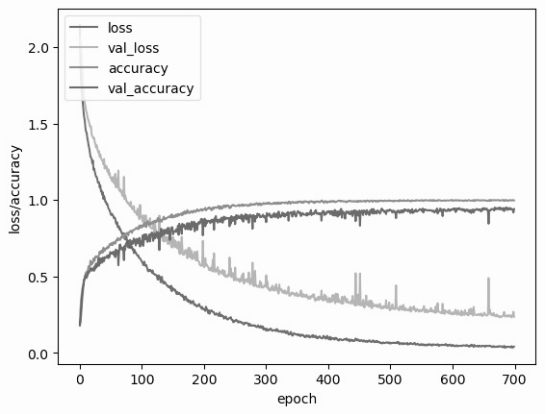

Figura 3. Desempenho do modelo no experimento 2 após a definição dos efeitos.

4.2. Análise dos Resultados

Com os modelos treinados após a aplicação dos efeitos nos dois cenários (sem diferenciação dos gêneros e com diferenciação dos gêneros), foram realizados os testes com o conjunto de dados criado. Esses testes visam entender se as variações linguísticas

presentes no idioma português do Brasil (pt-BR) são um fator importante para o desenvolvimento de sistemas de REF. Nesta seção, serão discorridos os resultados dos testes para cada variação linguística (baiano, carioca, gaúcho e mineiro) nos dois experimentos.

A Tabela 4 apresenta as métricas de desempenho, precisão (P), sensibilidade (S), F1-score (F1) e acurácia para cada teste realizado no experimento 1 com as amostras de áudios das variações linguísticas selecionadas.

Tabela 4. Resultados dos testes realizados no experimento 1.

	Baiano			**Carioca**			**Gaúcho**			**Mineiro**		
	P	S	F1	P	S	F1	P	S	F1	P	S	F1
raiva	1.00	0.15	0.26	1.00	0.20	0.33	0.33	0.06	0.11	0.18	0.10	0.13
alegria	0.00	0.00	0.00	0.64	0.45	0.53	0.00	0.00	0.00	0.24	0.30	0.27
neutra	0.24	0.45	0.32	0.44	0.55	0.49	0.33	0.35	0.34	0.25	0.25	0.25
tristeza	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Acurácia			0.20			0.40			0.14			0.22

De acordo com os resultados apresentados na Tabela 4, o teste da variação linguística Baiana obteve uma acurácia de 20%, o modelo apresentou um resultado insatisfatório no reconhecimento da emoção felicidade, mas foi possível o reconhecimento da emoção raiva e neutra. No teste da variação linguística Carioca, a acurácia obtida foi de 40% e o modelo reconheceu todas as emoções presentes no conjunto de testes; no geral, o teste da variação linguística carioca foi o que apresentou o melhor desempenho para o reconhecimento das emoções. Dentre todos os testes, o teste da variação linguística Gaúcha obteve a menor acurácia, com 14%; e da mesma forma que aconteceu no teste da variação baiana, não houve nenhuma classificação positiva para a emoção felicidade. O desempenho do modelo foi de 22% analisando a acurácia para o teste da variação linguística Mineira; a emoção mais reconhecida foi a felicidade, seguido pela emoção neutra.

A Tabela 5 apresenta as métricas de desempenho, precisão (P), sensibilidade (S), F1-score (F1) e acurácia para cada teste realizado no experimento 2 com as amostras de áudios das variações linguísticas selecionadas. O sufixo *m_* representa as classes do gênero masculino e o sufixo *f_* representa as classes do gênero feminino.

Tabela 5. Resultados dos testes realizados no experimento 2.

	Baiano			**Carioca**			**Gaúcho**			**Mineiro**		
	P	S	F1	P	S	F1	P	S	F1	P	S	F1
f_raiva	0.50	0.20	0.29	0.43	0.20	0.35	1.00	0.17	0.29	0.21	0.30	0.25
f_alegria	0.50	0.10	0.17	0.11	0.10	0.11	0.00	0.00	0.00	0.20	0.20	0.20
f_neutra	0.21	0.60	0.31	0.23	0.70	0.34	0.09	0.20	0.13	0.11	0.20	0.14
f_tristeza	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
m_raiva	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
m_alegria	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.10	0.12	0.86	0.60	0.71
m_neutra	0.33	0.50	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.10	0.12
m_tristeza	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Acurácia			0.23			0.18			0.07			0.22

De acordo com os resultados apresentados na Tabela 5, o teste da variação linguística baiana obteve o melhor desempenho entre os testes realizados com este mo-

delo, alcançando uma acurácia de 23%; o modelo não conseguiu classificar de forma correta 2 classes, *m_anger* e *m_happiness*, ou seja, não houve um desempenho considerável bom para as classes isoladas do gênero masculino. No teste da variação linguística carioca, as amostras das classes isoladas do gênero feminino foram as que obtiveram uma melhor classificação, diferenciando-se das amostras das classes do gênero masculino que não obtiveram nenhum valor correto na classificação; para este teste da variação linguística carioca, a acurácia do modelo obtida foi de apenas 18%. Assim como no teste da variação linguística gaúcha do modelo treinado no experimento 1, o teste da variação com o modelo do experimento 2 obteve a menor acurácia dentre todos os testes, com o valor de 7%. Por fim, o teste da variação linguística mineira apresentou uma acurácia de 22%; a única classe que não obteve um valor correto de classificação foi a classe *m_anger*, e o melhor desempenho foi na classificação de *m_neutral*.

Alguns dos fatores que podem ter influenciado para o baixo desempenho são as características do conjunto de testes, em que o viés da classificação da emoção não foi validada e não houve um tratamento de ruído. Em ambos os modelos, observou-se que as variações linguísticas presentes em um idioma podem alterar o desempenho geral de um sistema REF. A diferença entre a voz masculina e a voz feminina mostrou ser um fator importante para o desenvolvimento de sistemas REF, caso o objetivo seja saber qual o gênero em uma amostra de áudio, além de classificar a emoção. Então, o modelo precisa ser capaz de identificar as diferentes intensidades encontradas em cada emoção de ambos os gêneros.

5. Conclusão

Neste trabalho, buscou-se desenvolver um modelo de rede neural treinado para um sistema de REF utilizando uma arquitetura de rede neural amplamente utilizada na área de análise da emoção, a CNN. O método não licenciado utilizado para a realização da pesquisa foi desenvolvido para o reconhecimento de emoções através da fala em áudios do idioma Inglês[1]. A adaptação feita produziu um modelo com um bom desempenho, pois nos dois experimentos a acurácia obtida, na fase de treinamento, foi acima de 90%. Durante o desenvolvimento do trabalho, notou-se que ainda existe uma insuficiência relacionada ao volume dos conjuntos de dados no idioma português, sendo necessária a aplicação da técnica de aumento de dados. Durante os experimentos, foram identificados que os efeitos aplicados para aumentar o conjunto de dados de treinamento e validação influenciavam no desempenho geral do modelo. Aplicaram-se alguns efeitos e foi analisada a sua eficiência na melhoria do desempenho do modelo desenvolvido. Não foram todos os efeitos que obtiveram um bom resultado, mas os que apresentaram um aumento significativo do desempenho do modelo foram os efeitos *noise* e *shift*.

O conjunto de dados utilizado na fase de testes apresentou um baixo desempenho e foi possível identificar que é necessário treinar um sistema REF com conjuntos de dados em diferentes idiomas que apresentem também as possíveis variações linguísticas. Desse modo, o modelo aprende a reconhecer a emoção em diferentes entonações que podem ser encontradas em diferentes regiões de um país. A partir dos resultados obtidos, também foi possível concluir que o modelo treinado sem a capacidade de diferenciar o gênero presente em uma amostra de áudio é menos complexo e também apresenta um melhor desempenho

[1]Disponível em: https://github.com/N-F-I/Emotion-Recognition-through-speech-signal-using-CNN

na classificação da emoção. Mas, se em um problema real, o conhecimento de um gênero específico for importante, novas pesquisas que tenham como foco esse objetivo precisarão ser realizadas, pois o modelo treinado neste trabalho foi capaz de reconhecer e classificar melhor a voz feminina.

Com a finalidade de promover a continuidade das pesquisas na área de REF e a evolução deste trabalho, apresentam-se, a seguir, algumas propostas para trabalhos futuros: desenvolver um conjunto de dados no idioma português que contenha as variações linguísticas do Brasil validando de forma correta a classificação das emoções; treinar e avaliar os modelos com um conjunto de dados no idioma português que contenha outras emoções básicas classificadas, como a surpresa, o nojo e o medo; e construir um sistema que analisa a semântica e os elementos prosódicos vocais presentes em um enunciado de um falante.

6. Agradecimentos

Este artigo é resultado do projeto de pesquisa e desenvolvimento ARANOUÁ financiado pela Samsung Eletrônica da Amazônia Ltda nos termos da Lei Federal nº 8.387/1991, de acordo com o art. 21 do Decreto nº 10.521/2020. Agradecemos, também, ao Campus Manaus Zona Leste do Instituto Federal do Amazonas (IFAM) pelo suporte financeiro e incentivos para a realização deste trabalho.

Referências

Abdelhamid, A. A., El-Kenawy, E.-S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., and Eid, M. M. (2022). Robust speech emotion recognition using cnn+lstm based on stochastic fractal search optimization algorithm. *IEEE Access*, 10:49265–49284.

Abdulraheem, A., Salih, A., Abdulla, A., M.Sadeeq, M., O. M.Salim, N., Abdullah, H., Khalifa, F., and Abdullah, R. (2020). Home automation system based on iot. *Proceedings of the Technology Reports of Kansai University*, 62:2453.

Atmaja, B. T., Sasou, A., and Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*, 140:11–28.

Bastos Germano, R. G., Pompeu Tcheou, M., da Rocha Henriques, F., and Pinto Gomes Junior, S. (2021). emoUERJ: an emotional speech database in Portuguese.

Jaihar, J., Lingayat, N., Vijaybhai, P. S., Venkatesh, G., and Upla, K. P. (2020). Smart home automation using machine learning algorithms. In *Proceedings of the International Conference for Emerging Technology (INCET)*, pages 1–4.

Lech, M., Stolar, M., Best, C., and Robert, B. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Proceedings of the Frontiers in Computer Science*.

Lopez-Martin, M., Nevado, A., and Carro, B. (2020). Detection of early stages of alzheimer's disease based on meg activity with a randomized convolutional neural network. *Artificial Intelligence in Medicine*, 107:101924.

Lu, X. (2022). Deep learning based emotion recognition and visualization of figural representation. *Proceedings of Frontiers in Computer Science*.

Mustaqeem and Kwon, S. (2020). A cnn-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1).

Rumagit, R. Y., Alexander, G., and Saputra, I. F. (2021). Model comparison in speech emotion recognition for indonesian language. *Proceedings of the Procedia Computer Science*, 179:789–797.

Singh, V. and Prasad, S. (2023). Speech emotion recognition system using gender dependent convolution neural network. *Procedia Computer Science*, 218:2533–2540. International Conference on Machine Learning and Data Engineering.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Sun, T.-W. (2020). End-to-end speech emotion recognition with gender information. *IEEE Access*, 8:152423–152438.