# Identification of Emotions in Spoken Language Using Deep Learning

**Yasmin Maria Muniz de Oliveira[1], Carlos Maurício Seródio Figueiredo[1]**

[1]Laboratório de Sistemas Inteligentes (LSI) – Escola Superior de Tecnologia (EST) – Universidade do Estado do Amazonas (UEA) – Manaus, AM – Brasil

{ymmdo.snf19,cfigueiredo}@uea.edu.br

**Abstract.** Emotions are one of the pillars of human communication, especially spoken language. In emotional speech, they can be identified by inherent attributes of the voice, such as pitch, frequency, intensity etc. In this paper, a model based on artificial data augmentation and Deep Learning, more specifically a Convolutional Recurrent Neural Network, was proposed to automate this emotion identification task by being trained on the RAVDESS database with cross-validation technique. Evaluated by the accuracy and F1-Score metrics, the model achieved in them 76.25% and 76% on average and a maximum of 83.33% and 80%, respectively, which are slightly better results than those presented in related research.

**Resumo.** As emoções constituem um dos pilares da comunicação humana, especialmente da linguagem falada. Na fala emocional, as emoções podem ser identificadas por atributos inerentes à voz, como *pitch*, frequência, intensidade etc. Neste artigo, foi proposto um modelo baseado em aumento artificial de dados e *Deep Learning*, mais especificamente uma Rede Neural Recorrente Convolucional, para automatizar essa tarefa de identificação de emoções ao ser treinado na base de dados RAVDESS com técnica de validação cruzada. Avaliado pelas métricas acurácia e F1-score, o modelo atingiu nelas, respectivamente, 76,25% e 76% em média e uma máxima de 83,33% e 80%, o que são resultados levemente melhores que os apresentados em pesquisas relacionadas.

## 1. Introdução

A linguagem falada é a manifestação oral de ideias. É por meio dela que, desde tempos arcaicos, os seres humanos comunicam-se diretamente entre si para compartilhar pensamentos e emoções, sobretudo por meio de uma língua em comum [Zhang *et al*. 2018]. Hodiernamente, ainda, é evidente o aumento e a relevância do contato verbal humano com dispositivos digitais, como as assistentes virtuais [Koolagudi and Rao 2012]. A identificação automática das emoções naturalmente expressas pela fala, logo, destaca-se como um atrativo artifício para melhorar não apenas o convívio em sociedade como também essa crescente interação homem-máquina [Kwon *et al.* 2003].

Ao serem vocalizadas, as emoções humanas, tais como felicidade, tristeza e raiva, apresentam padrões constituídos por alterações típicas nos atributos inerentes à voz, entre os quais estão o *pitch*, a intensidade, a velocidade e a frequência [Koolagudi and Rao 2012]. Mediante a análise e a aprendizagem desses atributos, as emoções podem, portanto, ser identificadas por algoritmos de Inteligência Artificial, sobretudo por aqueles do subcampo do aprendizado profundo (*Deep Learning*).

Com a finalidade de reconhecer padrões e aprender abstrações em conjuntos de dados ao longo de múltiplas camadas de neurônios artificiais, as técnicas de aprendizado profundo são amplamente aplicadas em processamento de imagens, sons e vídeos [Han, Yu and Tashev 2014]. Redes Neurais Convolucionais (CNN), por exemplo, apresentam excelente desempenho em tarefas que envolvem a classificação de imagens, como a de espectrogramas obtidos de sinais sonoros, demonstrada por [Zhang *et al*. 2018], [Aharon *et al*. 2017], [N. Cummins *et al*. 2017] e [Qirong *et al*. 2014]. Por outro lado, Redes Neurais Recorrentes (RNN) são destinadas a lidar com informações temporais, conforme demonstram [Fayek, Lech and Cavedon 2017], [Jalal *et al*. 2019] e [Lee and Tashev 2015].

Este artigo, portanto, proporá um modelo de aprendizado profundo que integra camadas convolucionais e camadas recorrentes para reconhecer as propriedades típicas das emoções destacadas em cada amostra de áudio do conjunto de dados utilizado. A seguir, a Seção 2 consiste da apresentação de trabalhos relacionados a essa tarefa. A Seção 3 é destinada a evidenciar os materiais e métodos que foram empregados para chegar aos então resultados obtidos, dispostos na Seção 4. As considerações finais desse trabalho são desenvolvidas na Seção 5, por desfecho.

## 2. Trabalhos Relacionados

O reconhecimento de emoções em fala, apesar de ser abordado há tempos consideráveis, persiste como um intrigante desafio da área de Inteligência Artificial, haja vista sua magnitude e complexidade. Consiste basicamente de duas etapas: a extração de características da fala e a classificação das emoções com base nelas. Entre as prosódicas (como o ritmo) e as espectrais (como o *pitch*), as características espectrais são as mais atrativas em desafios como esse, principalmente porque permitem a representação dos sinais de áudio em espectrogramas. Essa representação é majoritariamente realizada por meio do emprego dos *Log Frequency Power Coefficients* (LFPC), *Linear Prediction Cepstral Coefficients* (LPCC) e *Mel-frequency Cepstral Coefficients* (MFCC), tal como atestam [Nwe, Foo and Silva 2003], [Kwon *et al.* 2003] e [Koolagudi and Rao 2012].

Já na etapa de classificação, com o advento do aprendizado de máquina, a tarefa em questão foi uma das várias revolucionadas pelos algoritmos classificatórios que emergiram. Entre eles, a Máquina de Vetores de Suporte (SVM) é a que concentra maior destaque por conta de sua fácil implementação e pouca exigência de processamento [Kwon *et al.* 2003]. Em posição seguinte, o Modelo Oculto de Markov (HMM), mesmo que não tão vantajoso, já foi bastante popular e ainda compõe vários sistemas híbridos, como constatam [Nwe, Foo and Silva 2003] e [Schuller, Rigoll and Lang 2003]. Ressaltam-se, ainda, o algoritmo de K-vizinhos Mais Próximos (KNN) e o Modelo de Misturas Gaussianas (GMM), ambos já amplamente testados nessa tarefa, sobretudo em processos comparativos [Schuller, Rigoll and Lang 2003].

Embora essas técnicas apresentem resultados significativos, a eficiência delas ainda proporciona certa margem a ser superada. Razão pela qual já não são mais tão utilizadas quanto antigamente no desafio tratado, o advento do aprendizado profundo irrompeu justamente a fim de preencher lacunas como essa. Sobressaem-se os algoritmos de Redes Neurais, como Máquinas de Aprendizado Extremo (ELM), Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN). ELMs são alimentadas com segmentos da entrada já rotulados por outro classificador para determinar o estado emocional emitido pela entrada inteira [Han, Yu and Tashev 2014].

Já CNNs, propostas essencialmente para a classificação de imagens, são conhecidas por extrair as melhores informações possíveis acerca das emoções estampadas nos espectrogramas dos sinais de fala, como mostram [Zhang *et al*. 2018], [Aharon *et al*. 2017], [N. Cummins *et al*. 2017] e [Qirong *et al*. 2014]. Por outro lado, RNNs são, nesse caso, mais eficientes em treinamento do tipo "sequência para um", no qual, após analisar quadro a quadro da sequência de entrada, preveem o rótulo do último deles, conforme apresentado por [Fayek, Lech and Cavedon 2017], [Jalal *et al*. 2019] e [Lee and Tashev 2015].

Entretanto, uma das abordagens que mais se sobressai ultimamente na tarefa de identificar emoções em fala é, na verdade, a que combina Redes Neurais Convolucionais com Redes Neurais Recorrentes, como demonstram [Jalal *et al*. 2019] e [Mustaqeem *et al*. 2020]. Enquanto as primeiras se focam em extrair características distintivas dos espectrogramas, as últimas se preocupam em assimilar informações temporais de cada um deles. Essa será a abordagem adotada e aperfeiçoada nesse trabalho a fim de elevar o desempenho sobre a base de dados utilizada.

## 3. Materiais e Métodos

Esta seção objetiva relatar os materiais e métodos empregados na elaboração e avaliação dos experimentos empreendidos por meio de três tópicos específicos. O primeiro visa explanar acerca das transformações efetuadas na base de dados selecionada. O segundo apresenta os detalhes das etapas da execução da tarefa em questão. Por fim, o terceiro esboça a arquitetura do modelo proposto. A figura 1 ilustra a visão geral desse sistema.

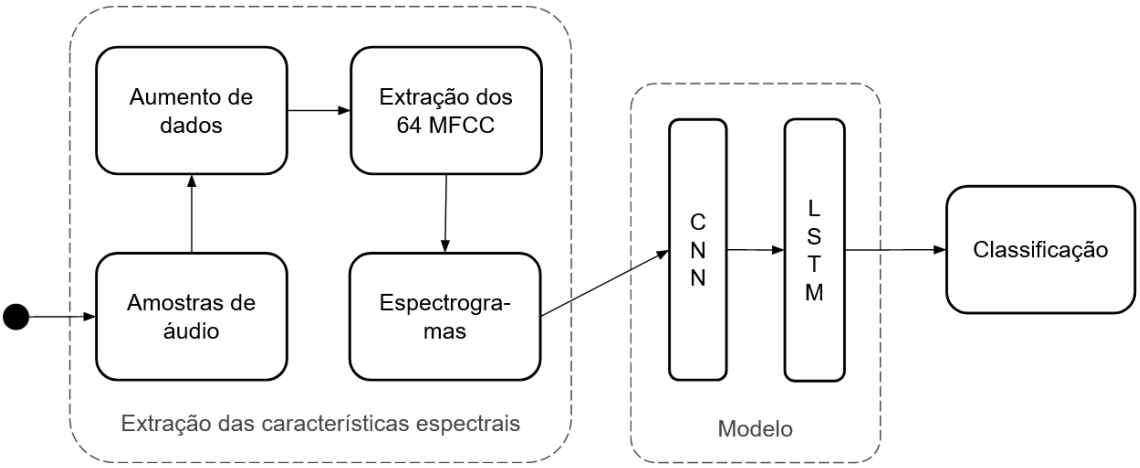


**Figura 1. Visão geral do sistema do modelo proposto**

### 3.1. Base de Dados

A base de dados selecionada para obter as amostras de áudio necessárias para a realização deste trabalho denomina-se RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*). Em adição à facilidade de acessá-la, o conteúdo em língua inglesa e o fato de ainda não ter sido tão explorada quanto outras bases de dados foram os motivos dessa escolha.

Ao todo, a base RAVDESS engloba 7356 gravações de 24 atores, dos quais 12 são homens e 12 são mulheres. Cada um performa 60 sentenças faladas, além de 44 sentenças cantadas. Há, ainda, três formatos de apresentação dessas gravações: áudio e

vídeo, apenas vídeo e apenas áudio [Livingstone and Russo 2018]. Para os propósitos deste trabalho, o enfoque deu-se somente sobre os áudios das sentenças faladas pelos 24 atores, o que resulta em 1440 amostras distintas.

As amostras de áudio, exportadas em formato WAV a 48000 Hz, têm como rótulo as emoções calma, felicidade, medo, nojo, raiva, tristeza e surpresa, além do estado de neutralidade, o que totaliza 8 classificações excludentes. No geral, há equilíbrio na frequência de distribuição das amostras, com 192 para cada emoção. A única exceção é o estado neutro, com apenas 96 amostras. Portanto, optou-se por não interferir na proporção dos dados extraídos e nem na quantidade de classes.

### 3.2. Etapas da Tarefa

A tarefa em questão trata-se de uma classificação multiclasse efetuada por um modelo de aprendizado profundo supervisionado. Além da etapa de treinamento, teste e validação do modelo proposto, a tarefa também exigiu uma etapa de pré-processamento para preparar os dados para a etapa seguinte.

Com base nos resultados de trabalhos anteriores quanto ao formato dos dados de entrada, o passo primordial da tarefa foi, portanto, transformar as amostras de áudio em espectrogramas, extraindo delas as características mais essenciais para a tarefa por intermédio de 64 MFCC (*Mel-frequency Cepstral Coefficients*). Além disso, a fim de evitar *overfitting* do modelo e potencializar a capacidade de generalização do mesmo, foi conduzido um aumento artificial dos dados de treino, fundamentado em variações aleatórias de características como *pitch,* velocidade e deslocamento temporal. O *pitch* (altura) refere-se a como a frequência fundamental sonora é captada pelo ouvido humano. Frequências baixas são interpretadas como sons graves e frequências altas como sons agudos. Então, cada áudio do conjunto de dados sofreu um deslocamento de *pitch* aleatório no intervalo de 30238 Hz a 76195 Hz. Já quanto à velocidade, cada áudio foi acelerado ou desacelerado em 0,25 a 2,5 vezes em relação à sua velocidade normal. O deslocamento temporal deu-se, por fim, em um intervalo de 10 a 150 posições na *array* gerada no carregamento de cada áudio. De tal maneira, a partir de uma amostra de áudio, foram obtidas mais oito amostras, totalizando 7776 amostras para o conjunto de treino. Essa constituiu a etapa de pré-processamento dos dados.

Para a etapa seguinte, foi empregada a técnica da validação cruzada *holdout* para obter do total de amostras disponíveis uma proporção de 60% de amostras para o treinamento, 30% para o teste e 10% para a validação. Essa segregação foi realizada aleatoriamente dez vezes, de modo a adquirir-se dez partições para cada conjunto, associadas em trios. O modelo foi, então, treinado e validado em cada uma delas para obter-se a média do desempenho dele.

### 3.3. Arquitetura do Modelo Proposto

As Redes Neurais Convolucionais (CNN), capazes de modelar padrões com elevada eficácia e resistência a distorções, são amplamente utilizadas em tarefas relacionadas à fala [Mustaqeem *et al*. 2020]. Neste trabalho, para extrair as características discriminativas dos espectrogramas dos áudios pré-processados, portanto, foram utilizadas duas camadas convolucionais. Cada uma foi acompanhada por uma operação de *max pooling* com uma janela de entrada 2x2 e por uma operação *dropout* a uma taxa de 0,5 a fim de reduzir os ruídos dos dados e prevenir *overfitting.* Ambas tiveram uma

janela de convolução 5x5 e a função *relu* como função de ativação. Quanto à quantidade de filtros, a primeira camada dispôs de 64 filtros e a segunda de 128.

Em seguida, com o intuito de analisar as informações contextuais ao longo dos quadros temporais de cada entrada, foi aplicada uma camada recorrente do tipo *Long Short-term Memory* (LSTM). Isso porque, em razão da célula de memória especial, ela apresenta maior efetividade no armazenamento de tais informações [Jalal *et al*. 2019]. Com as funções de ativação sigmoidal (para a recorrência) e tangente hiperbólica, atuou com 512 filtros no total a uma taxa de *dropout* de 0,7.

Por último, para determinar a saída final, uma camada densamente conectada regular foi integrada à arquitetura do modelo com 128 filtros e com a função de ativação tangente hiperbólica. Tal qual a camada anterior, também foi acompanhada por uma operação *dropout* com uma taxa de 0,7 para a redução das unidades de entrada.

Ademais, para o vetor da constante *bias*, foi adotada, em todas as camadas, uma função regularizadora com um fator de regularização do tipo L2 igual a $10^{-5}$. Em adição, na primeira camada do modelo, foi destinada uma inicialização com distribuição normal à matriz de pesos.

Por desfecho, com o otimizador que implementa o algoritmo *RMSprop* a uma taxa de aprendizagem de $10^{-4}$, o modelo foi, então, treinado com *batches* de tamanho 20 durante 200 épocas em todos os experimentos efetuados.

## 4. Resultados e Discussões

Nesta seção, estão especificados os resultados obtidos com o emprego dos materiais e métodos retratados na seção anterior. Conforme o progresso dos experimentos realizados com o modelo proposto, encontra-se dividida em três partes: o desempenho do modelo sem a aplicação de técnicas para aumentar artificialmente os dados, o desempenho do modelo com a aplicação de tais técnicas e, por fim, a média do desempenho do modelo nas 10 repartições efetuadas na validação cruzada.

### 4.1. Desempenho do Modelo sem Aumento de Dados

Assegurada a arquitetura do modelo a ser aplicada neste trabalho, o primeiro experimento realizado a fim de avaliar o desempenho dela teve como entrada os dados obtidos de uma única validação cruzada *holdout*, em proporções de 60% para o treino, 30% para o teste e 10% para a validação.

A matriz de confusão com as acurácias alcançadas em cada classe de emoção gerada na validação do modelo está ilustrada na Figura 2, o que culminou numa acurácia total de 0,6806.

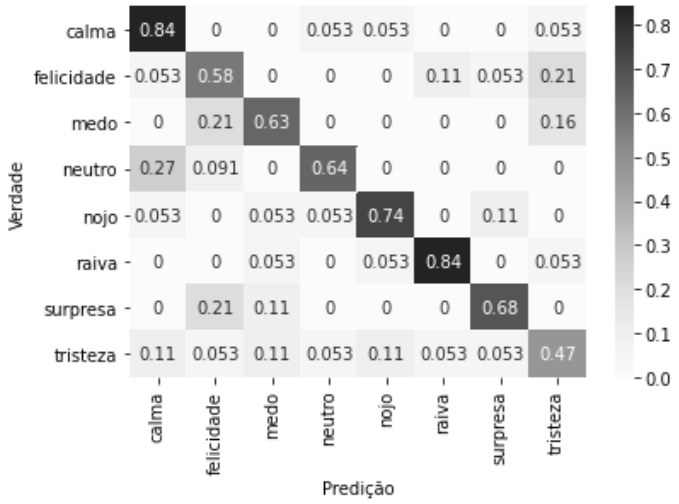


**Figura 2. Matriz de confusão da validação do modelo sem aumento de dados**

Já a Tabela 1 contém os resultados atingidos nas demais métricas sobre as quais o modelo foi avaliado.

**Tabela 1. F1-score, precisão e revocação obtidas pelo modelo sem aumento de dados**

| Classe | F1-score | Precisão | Revocação |
|---|---|---|---|
| Calma | 0,76 | 0,7 | 0,84 |
| Felicidade | 0,55 | 0,52 | 0,58 |
| Medo | 0,65 | 0,67 | 0,63 |
| Neutro | 0,67 | 0,7 | 0,64 |
| Nojo | 0,76 | 0,78 | 0,74 |
| Raiva | 0,84 | 0,84 | 0,84 |
| Surpresa | 0,72 | 0,76 | 0,68 |
| Tristeza | 0,49 | 0,5 | 0,47 |
| **Total** | **0,68** | **0,68** | **0,68** |

O propósito de tal experimento foi avaliar como o modelo se comportava sem maiores intervenções nos dados de treino para elevar o desempenho dele, bem como evidenciar as razões para realizá-las. Observa-se que metade das classes obteve menos de 70% em todas as quatro métricas, com destaque para as classes “Felicidade” e “Tristeza”. Em especial nessa última, métricas como acurácia e F1-score sequer atingiram 50%. Em virtude dessa disparidade de resultados entre as classes, ainda distantes do desejável, e especialmente para prevenir *overfitting*, tornou-se evidente a necessidade de aumentar artificialmente os dados de treino utilizados.

## 4.2. Desempenho do Modelo com Aumento de Dados

Com o objetivo de evitar *overfitting* do modelo e potencializar a capacidade de generalização dele, foi conduzido um aumento artificial dos dados do conjunto de treino. Cada amostra foi submetida a uma geração de mais oito amostras, com base em variações aleatórias no *pitch*, na velocidade e no deslocamento temporal do áudio em

questão. Ademais, foi utilizada a mesma especificação da validação cruzada do experimento anterior.

A matriz de confusão com as acurácias alcançadas em cada classe na validação do modelo está ilustrada na Figura 3. Dessa vez, a acurácia total obtida pelo modelo foi de 0,8333.

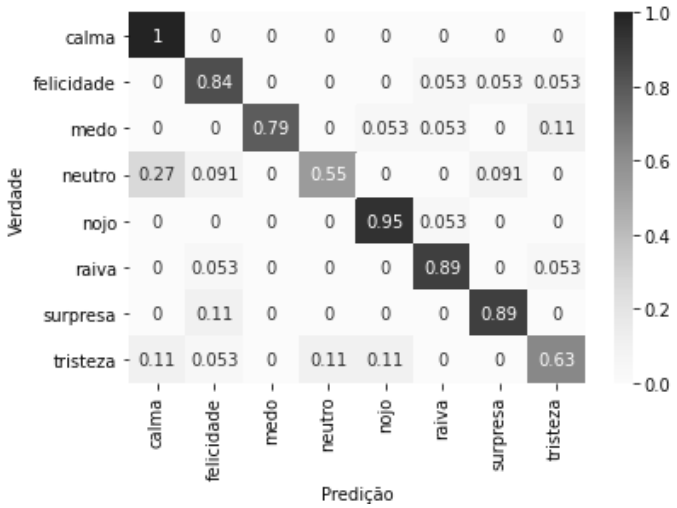


**Figura 3. Matriz de confusão da validação do modelo com aumento de dados**

Observou-se um crescimento expressivo não apenas na acurácia total como também na acurácia obtida na maioria das classes, variando de 79% a 100%. As únicas exceções foram as classes “Neutro” e “Tristeza”. A redução da acurácia para a classe “Neutro” pode ser explicada pelo fato de que, na base de dados original, ela possuía uma quantidade de amostras equivalente à metade da quantidade de amostras das demais classes. Como o aumento de dados foi efetuado de forma proporcional à quantidade original de amostras em cada classe, essa diferença tornou-se maior em termos absolutos, o que pode ter comprometido o desempenho do modelo com essa classe. Já a classe “Tristeza”, por outro lado, apesar de não ter atingido o resultado médio, teve um incremento de mais de 15% na métrica de acurácia em comparação ao experimento anterior. Além disso, a acurácia da classe “Felicidade” impulsionou-se em 26% com relação à obtida anteriormente, o que evidencia a melhora do desempenho geral do modelo nesse experimento.

Já a Tabela 2 ilustra os resultados obtidos pelo modelo nas demais métricas. Facilmente nota-se maior estabilidade neles quando comparados aos do experimento anterior. Na métrica F1-score, a maioria das classes atingiu 80% ou mais e nenhuma menos do que 60%. A classe “Neutro” foi a única que apresentou uma queda, tal como ocorreu na métrica acurácia. Porém, no caso da F1-score, foi uma queda de apenas 4%. Já as classes “Tristeza” e “Felicidade”, nas quais o modelo apresentou pior desempenho no experimento anterior, sofreram um aumento de 20% e 25%, respectivamente, na métrica F1-score.

**Tabela 2. F1-score, precisão e revocação obtidos pelo modelo com aumento de dados**

| Classe | F1-score | Precisão | Revocação |
|---|---|---|---|
| Calma | 0,88 | 0,79 | 1 |

| | | | |
|---|---|---|---|
| Felicidade | 0,8 | 0,76 | 0,84 |
| Medo | 0,88 | 1 | 0,79 |
| Neutro | 0,63 | 0,75 | 0,55 |
| Nojo | 0,9 | 0,86 | 0,95 |
| Raiva | 0,87 | 0,85 | 0,89 |
| Surpresa | 0,89 | 0,89 | 0,89 |
| Tristeza | 0,69 | 0,75 | 0,63 |
| **Total** | **0,83** | **0,83** | **0,84** |

### 4.3. Desempenho Médio do Modelo em 10 Partições Distintas dos Dados

Por desfecho, para constatar o desempenho geral do modelo, a técnica de validação cruzada *holdout* foi aplicada dez vezes sobre a base de dados utilizada, gerando, assim, 10 *splits* distintos com os conjuntos de treino, teste e validação nas mesmas proporções descritas na subseção 4.1 dessa mesma seção. Com um aumento artificial de dados baseado em variações aleatórias de características espectrais em cada conjunto de treino, o modelo proposto foi, em seguida, avaliado em uma cada das dez partições. As médias e os respectivos desvios-padrão dos resultados obtidos para as métricas de acurácia, F1-score, precisão e revocação podem ser analisados na Tabela 3.

**Tabela 3. Médias e desvios-padrão das métricas acurácia, F1-score, precisão e revocação obtidas pelo modelo em cada partição.**

| Partição | Acurácia | F1-score | Precisão | Revocação |
|---|---|---|---|---|
| 1 | 0,7986 | 0,79 | 0,80 | 0,80 |
| 2 | 0,7431 | 0,73 | 0,74 | 0,74 |
| 3 | 0,7500 | 0,75 | 0,79 | 0,75 |
| 4 | 0,8333 | 0,83 | 0,84 | 0,83 |
| 5 | 0,7361 | 0,73 | 0,75 | 0,74 |
| 6 | 0,7431 | 0,74 | 0,76 | 0,74 |
| 7 | 0,7917 | 0,79 | 0,80 | 0,79 |
| 8 | 0,7431 | 0,75 | 0,77 | 0,74 |
| 9 | 0,7292 | 0,73 | 0,76 | 0,73 |
| 10 | 0,7569 | 0,76 | 0,78 | 0,76 |
| **Média** | **0,7625 (± 0,03381)** | **0,7600 (± 0,03333)** | **0,7790 (± 0,03327)** | **0,7620 (± 0,03327)** |

Observa-se que, em cada partição, a maioria das métricas apresentou resultados na casa dos 70%, com variações relativamente pequenas entre si, o que constata a capacidade de generalização do modelo. Não houve, portanto, discrepâncias severas entre as métricas em cada partição, culminando numa média de 76,25% para a acurácia e 76% para a F1-score.

### 4.4. Resultados Comparativos

A Tabela 4 apresenta uma comparação entre o resultado atingido pelo modelo proposto e o dos trabalhos referenciados sobre a base de dados RAVDESS. Os trabalhos avaliam todas as oito classificações de emoções por meio da análise da única métrica comum a todos eles, a acurácia. No entanto, diferentemente da avaliação realizada pelo trabalho proposto, todos os trabalhos relacionados apresentam a métrica acurácia calculada sobre os dados de teste, sem segregar dados para a validação. Por isso, para que a avaliação dos trabalhos ocorra sobre a mesma base de comparação, foi apresentada para o modelo proposto a acurácia tanto dos dados de teste quanto dos dados de validação. É importante salientar que usar dados de validação segregados dos dados de teste consiste de uma metodologia adequada para reduzir a chance de viés dos modelos, sendo essa mais uma contribuição do trabalho proposto. Quando avaliado sobre os dados de teste, é possível constatar que o modelo proposto obteve um desempenho superior às referências averiguadas. Mesmo considerando os dados de validação, o desempenho dele ainda mostrou-se superior à maioria dos demais trabalhos.

**Tabela 4. Comparação de resultados entre o modelo proposto e os trabalhos referenciados.**

| Modelo | Acurácia obtida (%) |
|---|---|
| Zeng, Mao, Peng e Yi (2019) | 64,48 |
| Paiva (2017) | 67,68 |
| Jalal *et al.* (2019) | 69,40 |
| Bhavan *et al.* (2019) | 75,69 |
| Mustaqeem *et al.* (2020) | 77,02 |
| **Modelo proposto** | **77,71 (teste), 76,25 (valid.)** |

A fim de aprofundar a análise comparativa entre o modelo proposto e o trabalho de referência com o melhor desempenho visto anteriormente [Mustaqeem *et al.* 2020], a Tabela 5 evidencia a performance de ambos em cada classe da base de dados utilizada. Dessa vez, a análise foi efetuada com a métrica F1-score sobre os dados do conjunto de teste, uma vez que o trabalho referenciado não emprega um conjunto de validação. Por meio dela, é possível verificar que o modelo proposto superou de forma significativa a referência destacada nas classes "Calma" e "Felicidade". Por outro lado, observa-se que não há uma diferença expressiva na maioria das classes em que o trabalho referenciado mostrou-se melhor. Além do mais, é notável que a classe "Neutro" denota certa dificuldade para ambos os casos e, ainda assim, apresentou um resultado levemente maior no modelo proposto. Por fim, tal como ocorreu com a métrica acurácia, é importante notar que o modelo proposto atingiu uma F1-score total ligeiramente melhor que o trabalho de referência e obteve um equilíbrio maior entre os desempenhos com cada classe.

**Tabela 5. Comparação entre a métrica F1-score atingida pelo modelo proposto e pela referência destacada.**

| Classe | F1-score do modelo proposto | F1-score de Mustaqeem *et al.* (2020) |
|---|---|---|
| Calma | 0,82 | 0,71 |
| Felicidade | 0,72 | 0,60 |

| | | |
|---|---|---|
| Medo | 0,79 | 0,90 |
| Neutro | 0,69 | 0,67 |
| Nojo | 0,81 | 0,92 |
| Raiva | 0,82 | 0,87 |
| Surpresa | 0,82 | 0,85 |
| Tristeza | 0,67 | 0,73 |
| **Total** | **0,7750** | **0,77** |

## 5. Considerações Finais

A união entre Redes Neurais Convolucionais e Redes Neurais Recorrentes da arquitetura empregada revelou-se satisfatória quanto aos resultados obtidos na identificação das oito classes presentes nos áudios de fala da base de dados RAVDESS. Haja vista a comparação entre os experimentos conduzidos, é notável que a técnica de aumento artificial de dados foi fundamental para o aperfeiçoamento do modelo. Com o incremento na quantidade de amostras de áudio dos conjuntos de treino, baseado em variações aleatórias do *pitch*, velocidade e deslocamento temporal delas, foram obtidos os melhores resultados da pesquisa.

O modelo proposto mostrou-se homogêneo nas oito classes de emoções avaliadas, à exceção das classes “Neutro” e “Tristeza”, que apresentaram resultados um pouco inferiores. Destaca-se que o trabalho proposto seguiu uma metodologia de avaliação mais consistente ao dividir dados de teste e validação e apresentar o resultado para diferentes partições de dados, o que não feito nos demais trabalhos da literatura.

A pesquisa prosseguirá com o aperfeiçoamento da arquitetura do modelo a fim de alcançar maior regularidade entre as métricas de cada classe, além da aplicação de novos experimentos à etapa de pré-processamento dos dados.

## Agradecimentos

Este trabalho foi apoiado pela FAPEAM (Fundação de Amparo à Pesquisa do Estado do Amazonas) por meio do programa de Iniciação Científica PAIC/UEA-2020/2021 do edital nº 45/2020 e recebeu suporte da FAPEAM e do CNPq por meio do Programa PPP 04/2017.

## Referências

Aharon, S. *et al.* (2017). Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In *Interspeech 2017*, pages 1089-1093. DOI: 10.21437/Interspeech.2017-200.

Bhavan, A. *et al.* (2019). Bagged support vector machines for emotion recognition from speech. *Elsevier*, 184: 104886. DOI: https://doi.org/10.1016/j.knosys.2019.104886.

Cummins, N. *et al.* (2017) An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech. In *Proceedings of the 25th ACM international conference on Multimedia (MM '17)*, pages 478–484. DOI: 10.1145/3123266.3123371.

Fayek, H. M., Lech, M. and Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Elsevier*, 92: 60-68. DOI: 10.1016/j.neunet.2017.02.013.

Han, K., Yu, D. and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*, pages 23-227.

Jalal, M. A. *et al.* (2019). Learning temporal clusters using capsule routing for speech emotion recognition. In *Interspeech 2019*, pages 1701-1705. DOI: 10.21437/interspeech.2019-3068.

Koolagudi, S. G. and Rao, K. S (2012). Emotion recognition from speech: a review. In *International Journal of Speech Technology 15*, pages 99–117. DOI: 10.1007/s10772-011-9125-1.

Kwon, O. W. *et al.* (2003). Emotion recognition by speech signals. In *Eurospeech 2003*, pages 125-128.

Lee, J. and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech 2015*, pages 1537-1540.

Livingstone, S. R. and Russo, F. A (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS ONE*, 13(5): e0196391. DOI: 10.1371/journal.pone.0196391.

Mustaqeem *et al.* (2020). Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access*, 8: 79861-79875. DOI: 10.1109/ACCESS.2020.2990405.

Nwe, T. L., Foo, S. W. and Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Elsevier Speech Communications Journal*, 41(4): 603-623. DOI: 10.1016/S0167-6393(03)00099-2.

Paiva, E. C. (2017). Reconhecimento de emoção através da voz para integração em uma aplicação web. *Universidade Federal de Uberlândia*.

Qirong, M. *et al.* (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE*, 16(8): 2203-2213. DOI: 10.1109/TMM.2014.2360798.

Schuller, B., Rigoll, G. and Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings (ICASSP'03)*, pages II-1. DOI: 10.1109/ICASSP.2003.1202279.

Zeng, Y., Mao, H., Peng, D. and Yi, Z. (2019). Spectrogram based multi-task audioclassification. *Multimedia Tools Appl.*, 78(3): 3705–3722.

Zhang, Y. *et al.* (2018). Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* pages 1771-1775. DOI: 10.23919/APSIPA.2018.8659587.