Reconhecimento de Emoções na Fala entre Idiomas: Inglês *versus* Alemão

Erick K. Komati[1], **Karin S. Komati**[1]

[1]Coordenação de Informática – Instituto Federal do Espírito Santo (IFES)
Av. dos Sabiás, 330 – 29.166-630 – Serra – ES – Brasil

ekkomati@hotmail.com, kkomati@ifes.edu.br

Abstract. *The present study analyzed the generalization ability of a speech emotion recognition (SER) model in a cross-corpus scenario, training a one-dimensional convolutional neural network with English datasets and testing it on a German dataset. The training used the CREMA-D, RAVDESS, SAVEE, and TESS datasets, while the EmoDB dataset was employed for testing. The model achieved an accuracy of 0.61 on the training data, but its performance dropped to 0.30 when tested on the German dataset. This decline in performance highlights the limitations of SER in the face of linguistic and cultural differences.*

Resumo. *O presente estudo analisou a capacidade de generalização de um modelo de reconhecimento de emoções (SER) na fala em um cenário cross-corpus, treinando uma rede neural convolucional unidimensional com bases em inglês e testando em uma base em alemão. O treinamento utilizou as bases CREMA-D, RAVDESS, SAVEE e TESS, enquanto o teste empregou a base EmoDB. O modelo alcançou 0,61 de acurácia nos dados de treinamento, mas seu desempenho caiu para 0,30 ao ser testado na base em alemão. Essa queda de desempenho evidencia as limitações do SER diante de diferenças linguísticas e culturais.*

1. Introdução

A fala constitui um meio de comunicação que, além de transmitir conteúdo semântico, veicula informações emocionais por meio de elementos fonéticos e paralinguísticos [Hook et al. 2019]. Tais informações podem influenciar a interpretação da mensagem e auxiliar na compreensão do estado afetivo do locutor. O reconhecimento automático de emoções na fala (SER, do inglês Speech Emotion Recognition) busca identificar esses estados emocionais de forma computacional e tem sido aplicado em diferentes contextos, incluindo monitoramento veicular, sistemas de apoio terapêutico, centrais de atendimento e dispositivos móveis [Zaman et al. 2023].

O aprendizado profundo (*deep learning*) tem se consolidado como uma abordagem relevante para o reconhecimento de emoções, pois permite a extração automática de representações de alto nível diretamente dos sinais de áudio, reduzindo a necessidade de engenharia manual de atributos [Zaman et al. 2023]. Entre as arquiteturas disponíveis, as redes neurais convolucionais (CNNs) têm sido amplamente empregadas na análise de sinais unidimensionais, como o áudio, devido à sua capacidade de identificar padrões temporais e espectrais relevantes para a classificação emocional [Peixoto and Linhares 2023].

Apesar dos avanços obtidos, o SER apresenta limitações relacionadas à variabilidade entre locutores, diferenças linguísticas, contextos culturais e à sobreposição

de diferentes emoções em uma mesma fala. Esses fatores afetam a robustez e a generalização dos modelos, especialmente em cenários multilíngues, nos quais variações culturais podem influenciar a expressão vocal das emoções [El Ayadi et al. 2011]. A maior parte dos estudos concentra-se em bases monolíngues, o que restringe a compreensão sobre como diferenças culturais e linguísticas impactam o desempenho dos sistemas [Retta et al. 2023].

O presente estudo investiga a capacidade de generalização de um modelo baseado em CNN treinado com dados em inglês e avaliado em uma base distinta em alemão. Para o treinamento, foram utilizadas as bases CREMA-D, RAVDESS, SAVEE e TESS, todas compostas por amostras em inglês. Para o teste, empregou-se a base EmoDB, em alemão, com o objetivo de avaliar o desempenho do modelo frente a mudanças de língua e contexto cultural. Adotou-se a estratégia de validação cruzada entre bases (*cross-database validation*), que possibilita mensurar o impacto dessas variações no desempenho do modelo. As amostras foram processadas com técnicas consolidadas de extração de características, incluindo Mel-Frequency Cepstral Coefficients (MFCC), MelSpectrogram, Chroma Short-Time Fourier Transform (Chroma STFT), Root Mean Square Energy (RMS) e Zero-Crossing Rate (ZCR).

A estrutura do artigo é a seguinte: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve as bases de dados e os métodos; a Seção 4 relata e discute os resultados obtidos; por fim, a Seção 5 conclui o artigo e indica direções para pesquisas futuras.

2. Trabalhos Correlatos

Cross-Corpus Speech Emotion Recognition [Zhang et al. 2021] é o reconhecimento de emoções na fala quando o modelo é treinado e testado em corpora diferentes, que podem variar em idioma, cultura e condições de gravação. O objetivo é criar sistemas capazes de generalizar para contextos distintos, aproximando-se de situações reais. Essa tarefa é desafiadora porque diferenças entre corpora geralmente reduzem o desempenho dos modelos, exigindo técnicas para extrair características invariantes ao domínio.

O estudo de [Zehra et al. 2021] apresenta um sistema de reconhecimento de emoções na fala (SER) voltado para cenários multilíngues e com múltiplos conjuntos de dados (*cross-corpus*), baseado em aprendizado por conjunto (*ensemble learning*) com técnica de votação por maioria. A pesquisa empregou quatro corpora de idiomas distintos: SAVEE (inglês), Urdu, EMO-DB (alemão) e EMOVO (italiano). As amostras de áudio foram processadas para a extração de características espectrais e prosódicas, incluindo MFCCs. O conjunto de classificadores foi composto por três métodos: J48 (árvore de decisão), Floresta Aleatória e Sequential Minimal Optimization (SMO), combinados por meio de votação majoritária. Nos experimentos *cross-corpus*, o treinamento com dados em Urdu resultou, quando testado no corpus alemão, em medida F1 máxima de 0,61; no corpus em inglês, a maior medida F1 foi de 0,44; e no corpus em italiano, de 0,59. Quando o modelo foi testado em Urdu, as maiores medidas F1 obtidas foram de 0,60 (treinamento em alemão), 0,46 (treinamento em inglês) e 0,59 (treinamento em italiano).

O artigo de [Retta et al. 2023] investiga o reconhecimento de emoções da fala (SER) em contextos multilíngues e entre diferentes conjuntos de dados (*cross-corpus*), com foco na língua amárica. Os autores utilizaram o conjunto de dados ASED (Amharic Speech Emotion Dataset), que foi criado por eles em um trabalho anterior, e o compara-

ram com os conjuntos de dados RAVDESS (inglês), EMO-DB (alemão) e URDU (urdu). A pesquisa realizou três experimentos principais usando três modelos de aprendizado profundo: AlexNet, ResNet50 e uma variante proposta do VGG chamada VGGE. Os resultados do primeiro experimento, focado em SER monolíngue, sugeriram que o SER em amárico e inglês tem uma dificuldade similar, enquanto o alemão é mais difícil e o urdu é mais fácil. No segundo experimento, que envolveu treinamento em uma língua e teste em outra, Amharic⇔ German, Amharic⇔ English, and Amharic⇔Urdu, os resultados indicaram que o uso de inglês ou alemão como língua de treinamento (fonte) para o amárico (alvo) gerou os melhores resultados, sugerindo uma maior similaridade entre essas línguas em termos de SER. O terceiro e último experimento demonstrou que o treinamento com múltiplos conjuntos de dados não amáricos para testar no amárico obteve uma acurácia significativamente maior (4.07% superior) do que o treinamento com apenas um único conjunto de dados não amárico.

3. Materiais e Metódos

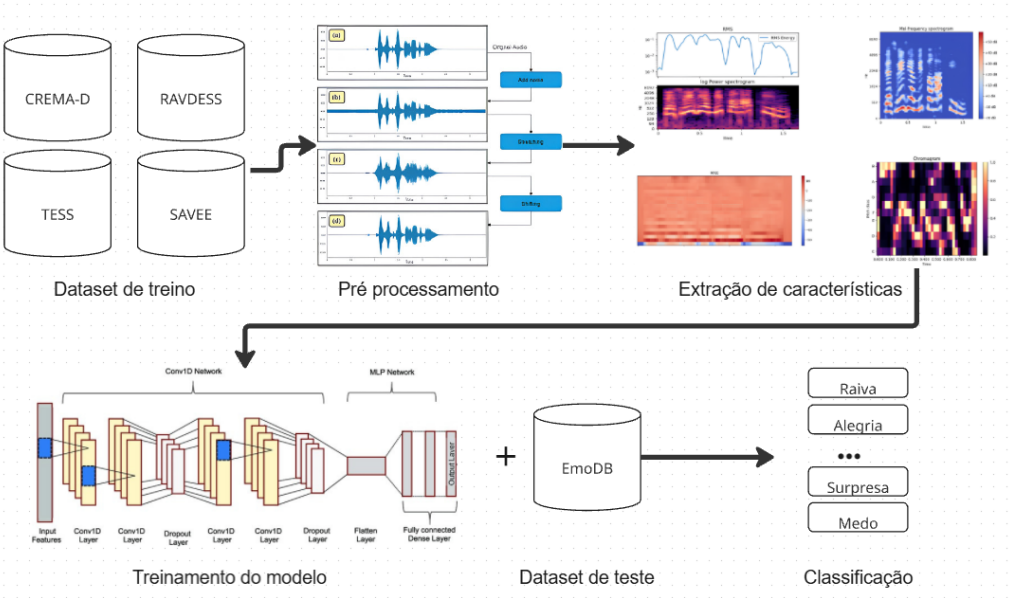

Figura 1. *Pipeline* do experimento.

A arquitetura do *pipeline* de reconhecimento de emoção da fala (SER) deste trabalho (Figura 1) teve a seguinte sequência de etapas:

- Base de dados de Treinamento: A primeira etapa utiliza quatro conjuntos de dados (*corpora*) distintos para treinar o modelo: CREMA-D, RAVDESS, TESS e SAVEE. Esses conjuntos de dados são combinados para formar a base de treinamento do sistema.
- Pré-processamento: A partir dos dados de áudio originais, são aplicadas técnicas de aumento de dados (*data augmentation*) para aumentar a quantidade e a variedade do conjunto de treinamento. Isso é ilustrado pelas transformações `Add noise` (adicionar ruído), `Stretching` (alongamento no tempo) e `Shifting` (deslocamento no tempo) do sinal de áudio original.

- Extração de Características: Os sinais de áudio pré-processados são convertidos em representações numéricas (vetores de características) que o modelo pode processar.
- Treinamento do Modelo: O modelo é uma rede neural convolucional de 1 dimensão (Conv1D Network).
- Base de dados de Teste: Após o treinamento, o modelo é avaliado utilizando um conjunto de dados diferente e em uma língua diferente, o EmoDB.
- Classificação: A etapa final é a classificação, onde o modelo faz uma previsão da emoção contida no áudio de teste.

3.1. Bases de Dados

Para avaliar a capacidade de generalização do modelo diante de variações linguísticas e culturais, foi adotada a técnica de Cross-Corpus Speech Emotion Recognition. Nesse procedimento, o modelo é treinado com bases em um idioma e testado em uma base independente, de outro idioma, possibilitando uma análise mais robusta da transferência entre domínios distintos.

Neste trabalho, foram utilizadas amostras de voz provenientes de cinco bases de dados distintas. Quatro delas, gravadas em inglês, foram empregadas no treinamento do modelo: CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset [Cao et al. 2014], RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [Livingstone and Russo 2018], SAVEE (Surrey Audio-Visual Expressed Emotion) [Jackson and Haq 2014] e TESS (Toronto Emotional Speech Set) [Pichora-Fuller and Dupuis 2020]. A quinta base utilizada, EmoDB (Berlin Database of Emotional Speech) [Burkhardt et al. 2005], contém gravações em alemão e foi destinada exclusivamente à etapa de teste, composta por expressões emocionais simuladas por atores profissionais. As emoções presentes em cada base e seu tempo médio de áudios (em segundos) estão listadas a seguir:

- **CREMA-D**: 6 emoções: raiva, nojo/desgosto, medo, alegria, neutro e tristeza; 3,70s;
- **RAVDESS**: 8 emoções: raiva, nojo/desgosto, medo, alegria, neutro, tristeza, **surpresa** e **calmo**; 2,54s;
- **SAVEE**: 7 emoções: raiva, nojo/desgosto, medo, alegria, neutro, tristeza e **surpresa**; 3,84s;
- **TESS**: 8 emoções: raiva, nojo/desgosto, medo, alegria, neutro, tristeza, **surpresa** e **agradável**; 2,06s;
- **EmoDB**: 7 emoções: raiva, nojo/desgosto, medo, alegria, neutro, tristeza e **tédio**; 2,78s.

As bases de dados utilizadas compartilham um conjunto central de seis emoções comuns: raiva, nojo, medo, alegria, neutro e tristeza, presentes em todas elas. As diferenças residem nas emoções adicionais incluídas por cada corpus. A RAVDESS e a TESS incorporam as emoções “surpresa” e “calmo” (na RAVDESS) ou “agradável” (na TESS), enquanto a SAVEE inclui apenas “surpresa” como extra. Já a EmoDB acrescenta “tédio” ao conjunto básico, e a CREMA-D mantém-se restrita às seis emoções centrais, sem adicionar categorias exclusivas. A Figura 2 apresenta a distribuição das amostras por emoção da base de dados de treino, isto é, a quantidade de áudios por emoção da junção das 4 bases de dados.

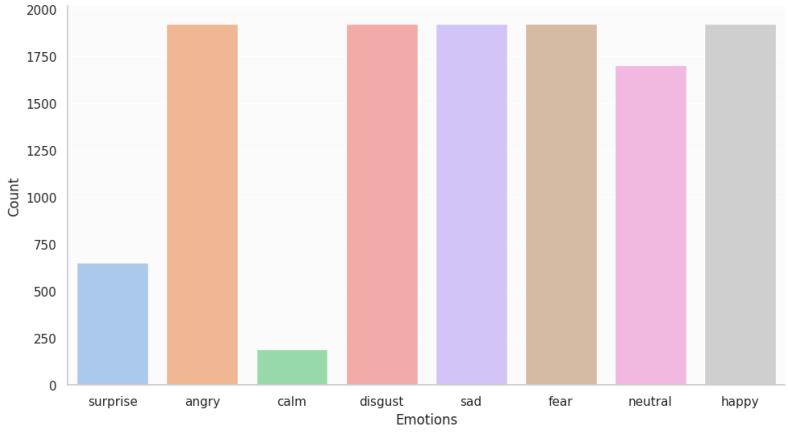

Figura 2. Gráfico de barras com quantidade de cada emoção

O RAVDESS é um conjunto multimodal com gravações de fala e canto em inglês norte-americano neutro. Contou com 24 atores profissionais (12 homens e 12 mulheres) e 7.356 gravações de duas frases lexicalmente idênticas. Há duas modalidades: a *speech* e a *song*. A porção usada neste trabalho corresponde à fala emocional apenas em áudio, totalizando 1.440 arquivos (*60 ensaios por ator* × *24 atores*). Cada emoção, exceto a neutra, tem dois níveis de intensidade (normal e forte).

A base CREMA-D é composta por 7.442 clipes gravados por 91 atores de diferentes idades e etnias; a base inclui seis emoções universais. Os clipes estão disponíveis em áudio, vídeo e audiovisual, e foram avaliados por 2.443 participantes via *crowdsourcing*, reduzindo vieses individuais e ampliando a diversidade de avaliadores, que atribuíram emoções categóricas e intensidade em uma escala contínua. Os clipes foram categorizados em três grupos: *matching* (emoção percebida coincide com a intenção do ator), *non-matching* (divergência entre percepção e intenção) e *ambiguous* (ausência de consenso entre avaliadores). Respostas com tempos de reação superiores a 10 segundos foram descartadas para eliminar avaliações possivelmente desatentas, resultando na exclusão de 3,6% das avaliações totais.

A base de dados SAVEE contém 480 gravações no formato .wav, de alta qualidade, de quatro atores homens falantes nativos do inglês britânico e com idades entre 27 e 31 anos. Cada ator interpreta sete emoções utilizando frases foneticamente equilibradas selecionadas do corpus TIMIT. Para a validação, as gravações foram avaliadas por 10 participantes sob três condições diferentes: áudio, vídeo e audiovisual.

A base TESS contém 2.800 arquivos de áudio gerados por duas atrizes nativas da língua inglesa com idades de 26 e 64 anos. As gravações foram realizadas em ambiente controlado, utilizando 200 palavras diferentes incorporadas em frases padronizadas no formato “Say the word ___”. Cada palavra foi pronunciada com uma das sete emoções. A seleção dessas emoções segue categorias comumente adotadas em estudos de reconhecimento de emoções na fala. As emoções atribuídas às amostras foram validadas por avaliadoras com audição normal. Cada gravação foi escutada e classificada em uma das categorias emocionais previstas, o que permite o uso da base como referência para estudos de reconhecimento de emoções com dados de voz.

3.1.1. EmoDB

Neste estudo, a base de dados de testes é o Berlin Database of Emotional Speech (EmoDB) [Burkhardt et al. 2005]. O protocolo de gravação dessa base estabelece que cada amostra de áudio expresse apenas uma emoção específica, interpretada por atores profissionais. A alta qualidade das gravações dispensa a aplicação de etapas de redução de ruído. As amostras foram registradas com taxa de amostragem de 48 kHz e posteriormente convertidas (*downsample*) para 16 kHz.

As emoções contempladas na base são: raiva, tédio, tristeza, alegria, ansiedade/medo, neutra e desgosto. Participaram das gravações 10 atores (cinco homens e cinco mulheres), cada um interpretando pelo menos uma das sete emoções em uma entre 10 sentenças predefinidas. A validação das emoções simuladas foi realizada por avaliadores humanos, que julgaram a qualidade emocional de cada interpretação. Para essa etapa, cada áudio foi reproduzido apenas uma vez, e o avaliador deveria identificar a emoção expressa e avaliar o quão convincente era a performance. Apenas os áudios que obtiveram taxa de reconhecimento acima de 80% e naturalidade superior a 66% seguiram para as próximas etapas. Nessas etapas adicionais, de caráter classificatório, avaliou-se a intensidade da emoção e o padrão de acentuação silábica, sendo permitido ouvir os áudios mais de uma vez.

Após essa filtragem, o conjunto final passou a conter 535 áudios. A identificação de cada amostra é codificada no nome do arquivo, informação que também pode ser utilizada para validar os resultados do sistema de reconhecimento. Cada nome de arquivo possui sete caracteres, que indicam, respectivamente: o número do ator, o código da sentença, a emoção expressa e a versão da gravação. Por exemplo, no arquivo `03a01Fa.wav`, o código `03` indica o ator número 3, `a01` corresponde à sentença, `F` representa a emoção *Freude* (felicidade, em alemão) e `a` indica a primeira versão da sentença gravada pelo ator. Algumas sentenças têm múltiplas versões, porém somente a última permanece na base.

3.2. Métodos

Os métodos para extração de características e classificador foram os mesmos propostos por [Burnwal 2021]. A extração de características é realizada a partir das amostras de áudio utilizando a biblioteca *librosa*. As representações extraídas têm como objetivo destacar aspectos relevantes do sinal para reconhecimento de emoções. Cada amostra é processada para compor um vetor de características com 162 elementos, resultante da concatenação de diferentes atributos acústicos, estruturados da seguinte forma:

- Bloco 1 (ZCR): Contém 1 valor. Quantifica o número de vezes em que o sinal cruza o ponto de valor nulo, sendo um parâmetro que fornece indícios sobre a natureza percussiva ou contínua do som, especialmente útil para segmentação de voz e detecção de transições bruscas no espectro.
- Bloco 2 (Chroma STFT): Contém 12 valores. Representa a intensidade média de cada uma das 12 classes de tom, indicando a intensidade relativa das classes de altura musical com base na transformada de Fourier de curto tempo.
- Bloco 3 (MFCC): Contém 20 valores. Representa os 20 coeficientes cepstrais de frequência de Mel médios do áudio.
- Bloco 4 (RMS): Contém 1 valor. Representa a energia média do áudio.

- Bloco 5 (MelSpectrogram): Contém 128 valores. Representa as 128 médias de frequências do espectrograma na escala de Mel.

3.2.1. Classificador

A arquitetura da rede neural convolucional de 1 dimensão (CNN1D), conforme detalhada na Figura 3, é dividida em duas partes principais: uma seção de extração de características e uma seção de classificação.

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 162, 256)	1,536
max_pooling1d (MaxPooling1D)	(None, 81, 256)	0
conv1d_1 (Conv1D)	(None, 81, 256)	327,936
max_pooling1d_1 (MaxPooling1D)	(None, 41, 256)	0
conv1d_2 (Conv1D)	(None, 41, 128)	163,968
max_pooling1d_2 (MaxPooling1D)	(None, 21, 128)	0
dropout (Dropout)	(None, 21, 128)	0
conv1d_3 (Conv1D)	(None, 21, 64)	41,024
max_pooling1d_3 (MaxPooling1D)	(None, 11, 64)	0
flatten (Flatten)	(None, 704)	0
dense (Dense)	(None, 32)	22,560
dropout_1 (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 8)	264

Figura 3. Arquitetura da CNN1D.

O processo se inicia com a entrada dos dados com 3 (três) conjuntos da combinação de camada convolucional `conv1d` com camada em uma camada *max pooling* (`max_pooling1d`). A primeira camada convolucional aplica 256 filtros a um vetor de entrada de 162 valores e, em seguida, a camada de *max pooling* reduz a dimensionalidade do vetor de características pela metade. A segunda camada convolucional aplica 256 filtros para um vetor de 81 posições, que é reduzido à metade pela camada de *max pooling*. Na sequência, uma terceira camada convolucional diminui o número de filtros para 128, e uma quarta para 64, enquanto as camadas de *max pooling* subsequentes continuam a reduzir o formato do vetor. A transição para a parte de classificação é feita por uma camada *flatten*, que transforma em um vetor 1D de 704 valores (resultado da multiplicação de 11 por 64). Por fim, duas camadas densamente conectadas realizam a classificação. A primeira camada densa tem 32 neurônios, e a segunda, que é a camada de saída, possui 8 neurônios, indicando que o modelo foi projetado para classificar em 8 classes de emoção distintas. Para evitar o *overfitting*, duas camadas de *dropout* são incluídas na arquitetura.

Os hiperparâmetros utilizados para a arquitetura da CNN foram: função de ativação ReLU, otimizador Adam, tamanho de lote 64, 50 iterações e o agendador de taxa de aprendizado `ReduceLROnPlateau` com fator 0,4, paciência de 2 iterações e taxa de aprendizado mínima 10^{-7}. O `ReduceLROnPlateau` reduz a taxa de aprendizado quando a métrica monitorada (*loss*) deixa de melhorar após o número de pocas definido pela paciência; a cada ajuste, a taxa é multiplicada pelo fator (0,4) e não reduzida abaixo do limite mínimo, o que ajuda a evitar a estagnação do treinamento. As métricas de precisão, revocação e medida F1 são obtidas a partir da função `classification_report` da biblioteca *Scikit-learn*.

4. Experimentos, resultados e discussão

Os resultados obtidos com a base utilizada para o treinamento do modelo são apresentados na Tabela 1. Esses valores refletem o desempenho geral alcançado durante o ajuste e aprendizado dos padrões emocionais. Na etapa de treinamento, utilizando exclusivamente bases em língua inglesa, o modelo apresentou desempenho consistente, alcançando acurácia global de 0,61 sobre 9.122 amostras. Entre as classes emocionais, destacaram-se os melhores resultados para surpresa (F1 = 0,82) e raiva (F1 = 0,73), enquanto desgosto e alegria obtiveram valores mais baixos de F1 (0,51 e 0,57, respectivamente). As métricas macro e ponderadas mantiveram-se próximas (F1 $\approx$ 0,64 e 0,61), sugerindo um desempenho relativamente equilibrado entre as diferentes classes, ainda que com variações perceptíveis conforme a emoção analisada.

Tabela 1. Métricas de desempenho por emoção nas bases em inglês e alemão.

	Inglês				Alemão			
Classe	**Precisão**	**Revocação**	**medida-F1**	**Suporte**	**Precisão**	**Revocação**	**medida-F1**	**Suporte**
raiva	0,78	0,69	**0,73**	1396	0,49	0,80	**0,61**	127
tristeza	0,58	0,68	**0,62**	1470	0,26	0,23	0,24	62
medo	0,63	0,51	0,57	1443	0,23	0,32	0,27	69
alegria	0,53	0,62	0,57	1450	0,12	0,20	0,15	71
neutro	0,55	0,57	0,56	1265	0,50	0,05	0,09	79
desgosto/nojo	0,54	0,48	0,51	1461	0,04	0,04	0,04	46
surpresa	0,85	0,79	**0,82**	495	0,00	0,00	0,00	0
calma	0,62	0,86	**0,72**	142			—	
tédio			—		0,00	0,00	0,00	81
Média Macro	0,63	0,65	0,64	9122	0,23	0,23	0,20	535
Média Ponderada	0,61	0,61	0,61	9122	0,27	0,30	0,24	535
Acurácia Global	0,61 (sobre 9122 amostras)				0,30 (sobre 535 amostras)			

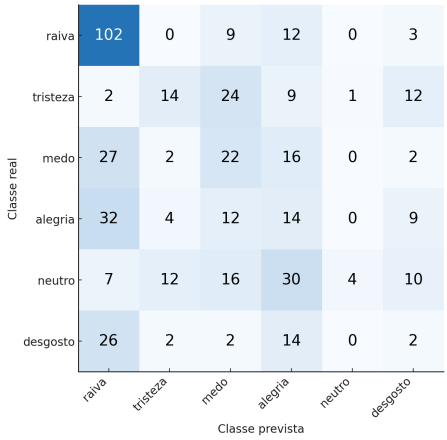

Figura 4. Matriz de confusão da base EmoDB.

Os resultados obtidos sobre a base de teste também se encontram na Tabela 1 e na matriz de confusão (Figura 4). A base EmoDB não foi utilizada durante o treinamento e as métricas dessa etapa fornecem uma visão sobre a robustez e aplicabilidade do modelo no reconhecimento de emoções em cenários distintos. O desempenho do modelo sofreu queda acentuada com relação à base de dados em inglês, com acurácia global reduzida para 0,30 em um total de 535 amostras. As classes tédio e surpresa não foram corretamente reconhecidas, apresentando F1 igual a zero. Entre as emoções, a raiva destacou-se

como a única que manteve desempenho relativamente alto em ambos os cenários, com F1 de 0,73 no treinamento e 0,61 no teste, sugerindo que suas características acústicas apresentam maior consistência entre inglês e alemão. Já as demais emoções tiveram métricas consideravelmente baixas na base de teste, evidenciando a dificuldade do modelo em generalizar para um idioma e contexto cultural distintos e reforçando a influência de fatores linguísticos e prosódicos na tarefa de reconhecimento de emoções em cenários *cross-corpus*.

Um resultado consistente foi que a classe "raiva" manteve desempenho superior em transferência de idioma. Isso pode dever-se a características prosódicas compartilhadas entre inglês e alemão, tais como intensidade, pitch, energia. Estudos mostraram que ouvintes de línguas distintas identificam raiva em gravações de uma língua estrangeira mesmo sem entender o idioma, baseando-se apenas em prosódia [Matos et al. 2024]. Esses achados sugerem que tais características funcionam como pistas universais ou próximas entre inglês e alemão para a emoção "raiva".

5. Considerações Finais

O presente estudo investigou a capacidade de generalização de um modelo de reconhecimento de emoções na fala (SER) ao ser treinado em um conjunto de dados em inglês e testado em um conjunto de dados de teste em alemão. Os resultados demonstram que, embora o modelo tenha alcançado uma acurácia global de 0,61 no conjunto de treinamento em inglês, essa performance não se manteve ao ser avaliado na base de dados em alemão. A acurácia global obtida na base de teste foi de 0,30, indicando um desempenho limitado na generalização para um idioma e contexto cultural distintos. Esse resultado corrobora a conclusão de outros trabalhos relacionados, que também apontam a dificuldade de generalizar modelos de SER em cenários multilíngues.

Os achados têm aplicações em cenários multilíngues de interação humano-máquina. Em assistentes virtuais e centrais de atendimento, o reconhecimento de emoções entre línguas pode orientar a adaptação das respostas e a priorização de casos conforme o estado do usuário. A estabilidade de classes como "raiva" sugere que modelos treinados em uma língua podem operar por transferência, oferecendo suporte inicial sem exigir, de saída, bases balanceadas por idioma. Essas soluções também podem compor recursos assistivos para pessoas com autismo, que frequentemente apresentam dificuldade em identificar sinais emocionais na fala, com impactos na socialização e na saúde mental [Chatterjee et al. 2015].

Futuras pesquisas poderiam explorar a combinação de um número maior de bases de dados de línguas distintas no treinamento, como demonstrado em outro estudo correlato, aplicar técnicas de aumento de dados mais robustas para mitigar o desequilíbrio entre as classes de emoção e implementar os diferentes experimentos propostos por [Retta et al. 2023].

Agradecimentos

A professora Komati agradece ao CNPq pela bolsa DT-2 (nº 302726/2023-3) e pelo projeto nº 407742/2022-0; e agradece à FAPES pelo projeto nº 1023/2022 P:2022-8TZV6.

Referências

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Burnwal, S. (2021). Speech emotion recognition. https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Chatterjee, M., Zion, D. J., Deroche, M. L., Burianek, B. A., Limb, C. J., Goren, A. P., Kulkarni, A. M., and Christensen, J. A. (2015). Voice emotion recognition by cochlear-implanted children and their normally-hearing peers. *Hearing research*, 322:151–162.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Hook, J., Noroozi, F., Toygar, O., and Anbarjafari, G. (2019). Automatic speech based emotion recognition using paralinguistics features. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, 67(3).

Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

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.

Matos, P. V. d. O. S., Andrade, R. S. S., Rehder, M. I. B. C., Guedes-Granzotti, R. B., Silva, K. d., and César, C. P. H. A. R. (2024). Reconhecimento da prosódia emocional por meio de pseudopalavras do Hoosier Vocal Emotions Collection. *Revista CEFAC*, 26:e3624.

Peixoto, G. d. S. and Linhares, J. E. d. S. (2023). Reconhecimento de emoçoes através da fala utilizando rede neural convolucional. In *Seminário Integrado de Software e Hardware (SEMISH)*, pages 119–130. SBC.

Pichora-Fuller, M. K. and Dupuis, K. (2020). Toronto emotional speech set (tess). *Scholars Portal Dataverse*, 1:2020.

Retta, E. A., Sutcliffe, R., Mahmood, J., Berwo, M. A., Almekhlafi, E., Khan, S. A., Chaudhry, S. A., Mhamed, M., and Feng, J. (2023). Cross-corpus multilingual speech emotion recognition: Amharic vs. other languages. *Applied Sciences*, 13(23):12587.

Zaman, K., Sah, M., Direkoglu, C., and Unoki, M. (2023). A survey of audio classification using deep learning. *IEEE access*, 11:106620–106649.

Zehra, W., Javed, A. R., Jalil, Z., Khan, H. U., and Gadekallu, T. R. (2021). Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 7(4):1845–1854.

Zhang, S., Liu, R., Tao, X., and Zhao, X. (2021). Deep cross-corpus speech emotion recognition: Recent advances and perspectives. *Frontiers in Neurorobotics*, Volume 15 - 2021.