Avaliação do uso de *embeddings* para o reconhecimento de emoções na fala

Pedro Munhoz, Guilherme Cavazzotto, Larissa Guder, Luan Dopke, Dalvan Griebler

[1] Escola Politécnica, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Porto Alegre – RS – Brasil

{p.munhoz, guilherme.sanches, larissa.guder, luan.dopke}@edu.pucrs.br

dalvan.griebler@pucrs.br

Abstract. *The present article aims to evaluate the usage embedding-based feature extraction methods, for the task of speech emotion recognition. In order to do so, the IEMOCAP dataset was utilized and 9 classifier models were trained and tested on 11 different feature sets. As a result, it was observed that the feature extraction method "trillsson5" resulted in the combination with the highest accuracy, which suggests that embedding-based models may outperform others in the task.*

Resumo. *O presente artigo tem como objetivo avaliar o uso de métodos de extração de atributos baseados em embeddings para a tarefa de reconhecimento de emoções. Para isso, foi utilizado o conjunto de dados IEMOCAP e 9 modelos classificadores foram treinados e testados com 11 conjuntos de atributos diferentes. Como resultado, foi observado que o modelo trillsson5 de extração de atributos resultou na combinação de melhor acurácia, o que sugere que os modelos baseados em embeddings podem superar os demais na tarefa.*

1. Introdução

A área de reconhecimento de emoções na fala engloba métodos e técnicas voltados ao processamento e à classificação de sinais de fala, com o objetivo de identificar e interpretar as emoções presentes nesses sinais. Pode-se representar as emoções de duas formas: dimensional (contínua) e categórica (discreta). Na abordagem dimensional, as emoções são representadas em escalas contínuas, como valência (grau de positividade ou negatividade) e excitação (nível de ativação ou intensidade), permitindo avaliar a intensidade de cada estado emocional [Russell 1980]. Já na abordagem categórica, as emoções são comumente classificadas em categorias discretas, como felicidade, tristeza, raiva, medo, nojo, surpresa e medo [Lieskovská et al. 2021].

Em geral, as diferentes formas de representação quantitativa de sinal de fala podem ser divididas em dois grupos: as formas manuais e as que geram *embeddings*. As primeiras são baseadas em aspectos acústicos do sinal. Já as segundas são extraídas através da utilização de modelos pré-treinados de aprendizado de máquina e têm apresentado bons resultados nos últimos anos [Feng et al. 2024].

Dentre a literatura analisada por [Hashem et al. 2023], há poucos estudos que exploram o uso de *embeddings* como forma de representação de sinais de fala para o treinamento de modelos cuja finalidade é o reconhecimento de emoções conforme a abordagem categórica, usando o conjunto de dados IEMOCAP (*Interactive Emotional Dyadic Motion Capture Database*) [Busso et al. 2008]. Em trabalho prévio, foi avaliado o uso de *embeddings* para a previsão de emoções segundo o modelo dimensional, no mesmo conjunto de dados [Guder et al. 2024].

Tendo isso em vista, o presente artigo tem como objetivo investigar o impacto do uso de métodos de extração de atributos baseados em *embeddings* para a tarefa de reconhecimento de emoções em sinais de fala. Para isso, foi utilizado o conjunto de dados IEMOCAP, por conta do fato de ser ele o mais comumente usado para a testagem de sistemas que são o estado da arte [Lieskovská et al. 2021].

Como trabalho relacionado, [Purohit et al. 2023] buscou realizar uma comparação similar. Mas em seu estudo, apenas o conjunto de atributos "Compare" foi usado como demonstrativo dos métodos manuais. Além disso, apenas dois classificadores foram utilizados. Este trabalho expande esse escopo ao utilizar outros dois conjuntos de atributos e nove classificadores distintos.

2. Metodologia e Desenvolvimento

O conjunto de dados escolhido para a realização dos experimentos é o IEMOCAP. Nele, há aproximadamente 12 horas de áudios de falas pronunciadas por atores dos sexos masculino e feminino, que estão anotados conforme a emoção no formato categórico [Busso et al. 2008]. Na tabela 1 estão as classes de emoções presentes no conjunto de dados, bem como o número de áudios cuja emoção é representada pela classe em questão.

Tabela 1. Distribuição das classes no conjunto de dados

Classe	xxx	fru	neu	ang	sad	exc	hap	sur	fea	oth	dis
Número de Instâncias	2507	1849	1708	1103	1084	1041	595	107	40	3	2

Dando ênfase à anotação manual do IEMOCAP, ele foi projetado para conter emoções bem definidas, mas manifestações ambíguas ou mistas são comuns. Logo, para simplificar a anotação, os avaliadores poderiam selecionar múltiplas emoções, sendo usada a votação majoritária para definir a classe final. Quando ocorreu empate, registrou-se a classe como "xxx".

Conforme [Lieskovská et al. 2021], é comum a utilização das classes "raiva", "neutro", "tristeza" e uma combinação das classes "excitação" e "alegria", em artigos que visam propor modelos de classificação de emoções utilizando o conjunto de dados em questão. Por isso, por fins de comparação, para este estudo , apenas essas quatro classes foram utilizadas.

Como pré-processamento para esta pesquisa, todos os áudios foram re-amostrados para uma *sample rate* de 16 kHz, e convertidos para o formato *mono*. A extração de atributos é uma etapa fundamental no reconhecimento de emoções na fala, pois define como o sinal de fala será representado para a classificação. Para fins de comparação, foram utilizados atributos extraídos manualmente, como as propostas nos conjuntos ComParE_2016, eGeMAPSv02 e pAA [1], que são definidas com base em conhecimento prévio acústico (como energia e pitch). Mais recentemente, modelos baseados em aprendizado profundo passaram a gerar representações automáticas, conhecidas como *embeddings*, que capturam informações de alto nível diretamente do sinal de fala, sem necessidade de especificar manualmente quais atributos extrair. Dentre as opções existentes, testamos os modelos *VGGish*, *TRILL*, *TRILLsson5*, *FRILL*, *Whisper*, *HuBERT*, *WavLM*, e *Wav2Vec2*. A extração de atributos foi feita a cada arquivo de áudio separadamente e, após o processo de extração, foi aplicada uma padronização a cada instância de conjunto de atributos através do módulo *StandartScaler*[2].

[1] As bibliotecas estão disponíveis em: <https://github.com/audeering/opensmile-python/> (para as duas primeiras); e <https://github.com/tyiannak/pyAudioAnalysis>. Acesso em: 22 de agosto de 2025.

[2] Ver: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>. Acesso em: 06 de outubro de 2025

Os modelos de classificação têm como objetivo mapear as representações extraídas dos áudios, sejam elas atributos extraídos manualmente ou *embeddings*, para classes emocionais definidas anteriormente. Ou seja, recebem como entrada números que representam características e produzem como saída a predição da emoção correspondente. Neste artigo, foram utilizados diferentes modelos de classificação. Entre eles estão o *XGBoost*[3]; a Floresta Aleatória; as árvores de decisão; o *CatBoost*[4]; o *LightGBM*[5], uma versão menos custosa computacionalmente do *XGBoost*; o *Multi-layer Perceptron* (MLP); o k-Vizinhos Mais Próximos (KNN); a Regressão Logística; e a Máquina de Vetores de Suporte[6]. Para a execução de todos os classificadores, foram utilizados os seus parâmetros padrão.

O mapa de calor apresentado na Figura 1 mostra a acurácia entre classificadores (linhas) e a forma de extração de atributos utilizada (colunas), sendo as três primeiras manuais e as demais *embeddings*. O trillson5 com MP obteve a maior acurácia (0,77), pois fornece representações ricas que se ajustam bem a modelos não lineares. No geral, os *embeddings* acabaram por ter uma média de acurácia similar à dos atributos manuais: ambas resultaram em aproximadamente 60%.

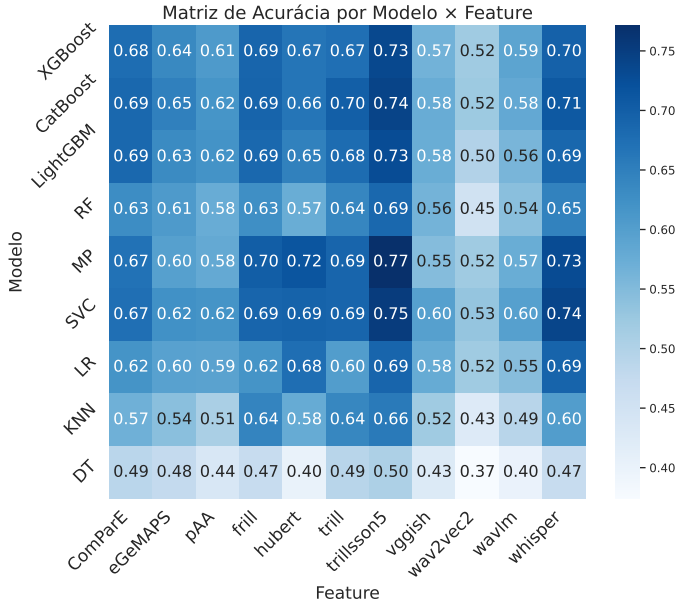

Figura 1. Mapa de calor com as acurácias atingidas através de cada execução.

Levando isso em consideração, é possível afirmar que formas de extração de atributos baseadas em *embeddings* não necessariamente acarretam desempenhos melhores para modelos de classificação de emoções em sinais de fala. Todavia, vale ressaltar que, dentre as formas de extração de atributos baseadas em *embeddings* aqui utilizadas, existe uma enorme variação, por um lado, de forma de representação (por exemplo: de número de dimensões do vetor produzido) e, por outro, de acurácia média obtida. Tendo isso em vista, é possível que, caso seja feito algum tipo de otimização dos hiper-parâmetros passados para os modelos (com o uso de *Grid Search*, por exemplo), possivelmente haverá uma mudança significativa nos níveis de acurácia obtidos. Além disso, embora nem

[3]Disponível em: <https://xgboost.readthedocs.io/en/stable/>. Acesso em: 22 de agosto de 2025.

[4]Disponível em: <https://catboost.ai/>. Acesso em: 22 de agosto de 2025.

[5]Disponível em: <https://lightgbm.readthedocs.io/en/stable/>. Acesso em: 22 de agosto de 2025.

[6]Com exceção dos três primeiros apresentados, todos foram implementados através do uso de biblioteca *Scikit-learn*. Ver: <https://scikit-learn.org/stable/>. Acesso em 22 de agosto de 2025

todas as formas de extração de atributos baseadas em *embeddings* tenham resultado em acurácias melhores que as manuais, aquelas que obtiveram as melhores acurácias foram baseadas em *embeddings*. Isso sugere que as baseadas em *embeddings* têm o potencial de superarem as manuais na tarefa de reconhecimento de emoções, caso os modelos sejam otimizados para essa tarefa. No caso do modelo que levou à melhor combinação em termos de acurácia, o trillsson5, por conta do fato de ser um modelo que, através de um mecanismo chamado "distilação de conhecimento", se baseia em um modelo pré-treinado com sinais de fala [Shor and Venugopalan 2022], não é surpreendente que, para essa tarefa, ele tenha tido uma boa eficácia.

3. Conclusão

Conclui-se que o modelo *Trillsson5* combinado com o Perceptron Multi-Camadas apresentou o melhor desempenho em termos de acurácia, e que, apesar do fato de que as formas de extração de atributos baseadas em *embeddings* não terem superado as manuais, as primeiras têm o potencial de superarem as segundas na tarefa de reconhecimento de emoções caso o modelo certo seja selecionado. Ademais, uma otimização dos hiperparâmetros passados aos modelos poderia impactar significativamente o desempenho dos mesmos e, por isso, sugere-se que, para estudos futuros, o mecanismo de *Grid Search* seja utilizado para que hiper-parâmetros mais apropriados sejam selecionados, e para que a comparação possa ser feita de forma mais fidedigna.

Referências

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Feng, T., Hebbar, R., and Narayanan, S. (2024). Trust-ser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11201–11205.

Guder, L., Aires, J., Meneguzzi, F., and Griebler, D. (2024). Dimensional Speech Emotion Recognition from Bimodal Features. In *Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 579–590, Porto Alegre, RS, Brasil. SBC.

Hashem, A., Arif, M., and Alghamdi, M. (2023). Speech emotion recognition approaches: A systematic review. *Speech Communication*, 154:102974.

Lieskovská, E., Jakubec, M., Jarina, R., and Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10):1163.

Purohit, T., Vlasenko, B., and Magimai-Doss, M. (2023). Implicit phonetic information modeling for speech emotion recognition. In *INTERSPEECH 2023*, Interspeech, pages 1883–1887. Interspeech Conference, Dublin, IRELAND, AUG 20-24, 2023.

Russell, J. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39:1161–1178.

Shor, J. and Venugopalan, S. (2022). Trillsson: Distilled universal paralinguistic speech representations. *arXiv preprint arXiv:2203.00236*.