Impacto do Ajuste Fino na Redução de Dimensionalidade para Reconhecimento Multimodal de Emoções na Fala

Larissa Guder[1,2], **Dalvan Griebler**[1], **Felipe Meneguzzi**[1,3]

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

[2]Laboratório de Pesquisas Avançadas para Computação em Nuvem,
Sociedade Educacional Três de Maio (SETREM)
Três de Maio – RS – Brasil

[3]University of Aberdeen – Aberdeen – Scotland

lariguder10@gmail.com, dalvan.griebler@pucrs.br

felipe.meneguzzi@abdn.ac.uk

Abstract. *The objective of this work is to evaluate the impact of fine-tuning on the dimensionality reduction of the MiniLM L3 sentence embedding for dimensional speech emotion recognition, using a multimodal approach that combines acoustic and textual information. Through fine-tuning, it was possible to achieve a threefold increase in the value of the Concordance Correlation Coefficient for the valence dimension.*

Resumo. *O objetivo desse trabalho é avaliar o impacto do ajuste fino na redução da dimensionalidade do embedding de sentença MiniLM L3, para a tarefa de reconhecimento dimensional de emoções na fala, através de uma abordagem bimodal que combina informações acústicas e textuais. O ajuste fino resultou em um aumento de 3x no Coeficiente de Correlação de Concordância para a dimensão de valência.*

1. Introdução

As nossas emoções desempenham um papel subjetivo e controverso, sendo essenciais para a nossa sobrevivência psíquica. Compreender as emoções dos outros e como elas são expressas é fundamental para que nos relacionemos enquanto sociedade. Por exemplo, enquanto o medo é um regulador protetivo natural e auxilia na tomada de decisões, a raiva permite-nos estabelecer limites e desenvolver o nosso senso de justiça. Com base nisso, o reconhecimento de emoções é mais uma perspectiva do que uma ciência exata.

Além das formas utilizadas para determinar emoções na psicologia, duas abordagens têm sido utilizadas para reconhecer emoções usando *deep learning*: classes discretas e dimensionais [Lieskovská et al. 2021]. Nas classes discretas, são utilizadas as seis emoções consideradas essenciais por [Ekman 1999]: raiva, nojo, medo, felicidade, tristeza e neutro, onde o modelo deve classificar a entrada de acordo com a classe mais correlacionada. Por outro lado, [Russell 1980] define uma abordagem dimensional através do modelo circumplexo de afeto. O modelo circumplexo considera duas dimensões: excitação e valência. Cada dimensão possui um valor que varia de -1 a 1. A excitação está relacionada à tonalidade calma ou estimulante da fala, enquanto a valência representa o quão agradável ou desagradável ela é. Com a pontuação de cada dimensão, é possível correlacionar com uma emoção específica. Por exemplo, medo e raiva podem ser definidos como de baixa valência e alta excitação. [Mehrabian 1996] adiciona a dimensão de dominância, que representa como a emoção influencia o comportamento de uma pessoa.

É importante que os modelos reconheçam as emoções e respeitem a diversidade idiossincrática de cada pessoa.

A escassez de dados para o treino e teste de modelos de *deep learning* dificulta o crescimento da tarefa de reconhecimento de emoções na fala [de Lope and Graña 2023]. Os conjuntos de dados existentes possuem uma quantidade limitada de dados disponíveis; são menos diversos do que o necessário ou diferem bastante dos dados do mundo real. Mesmo ao focar apenas no reconhecimento de emoções na fala, é necessário considerar que a percepção humana das emoções envolve múltiplos sentidos, sendo multimodal [Geetha et al. 2024]. Dessa forma, para superar essa limitação e extrair mais informações apenas dos dados de fala, o uso de informações textuais pode melhorar a precisão dos classificadores. O objetivo principal é avaliar o efeito da utilização de ajuste fino em relação ao uso de PCA, conforme pesquisa prévia [Guder et al. 2024], para reduzir o tamanho da dimensão para o modelo de *embedding* de sentença, MiniLm L3.

2. Metodologia

Para avaliar os experimentos, foi utilizado o conjunto de dados IEMOCAP (The Interactive Emotional Dyadic Motion Capture) [Busso et al. 2008]. O IEMOCAP contém informações multimodais, combinando vídeo, fala, captura de movimento facial e transcrições textuais. O uso de diferentes fontes de informação pode levar a predições mais robustas. Para isso, dentre as *features* presentes no conjunto de dados, foram utilizadas a fala e as transcrições. O IEMOCAP fornece anotações para cada elocução (*utterance*). As pontuações para as três dimensões variam de 1 a 5 e foram normalizadas para uma escala de -1 a 1 . O conjunto de dados contém aproximadamente 12 horas de fala. Como o IEMOCAP não contém informações sobre a proporção de divisão, ele foi dividido nas proporções de 60/20/20 para treino, teste e validação, utilizando 42 como *seed*.

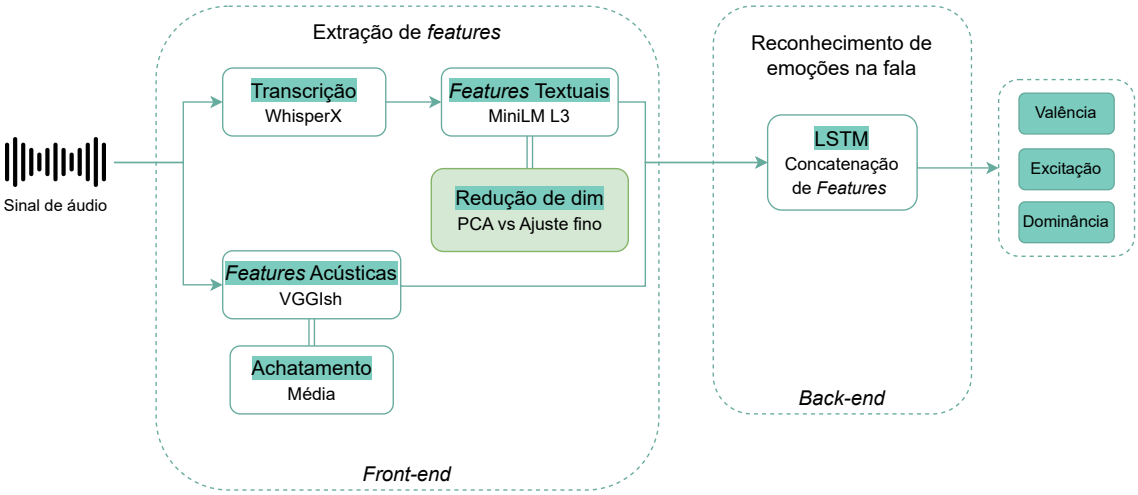

Figura 1. Arquitetura completa. Em verde claro está destacado o processo que será alvo nesse trabalho

A arquitetura proposta de ponta a ponta (*end-to-end*) consiste em dois blocos: *front-end* e *back-end*. O front-end é responsável por extrair as *features* do sinal de entrada, enquanto o back-end é responsável por processar a informação vinda do *front-end* e predizer a saída. A arquitetura é detalhada na Figura 1.

Dado o sinal de entrada, ele é transformado em uma forma mono *waveform* e é reamostrado para uma taxa de amostragem de 16 kHz. A partir disso, dois conjuntos de

features são extraídos. Para que a extração de recursos textuais seja possível, primeiramente, o sinal de áudio é transcrito utilizando o modelo WhisperX. Depois, é utilizado o modelo MiniLM L3 para gerar a representação de sentença. Para a extração de *features* acústicas, o modelo VGGIsh é utilizado. Por gerar um *embedding* a cada segundo de áudio, foi aplicado o achatamento do vetor para gerar uma única representação de dimensão (1,128). A rede *back-end* utiliza uma rede LSTM para processar os dados de entrada. A primeira camada concatena ambos os conjuntos de características (features). Utiliza-se a ordem $(audio, texto)$. Após a camada de entrada, emprega-se uma camada de normalização em lote (*batch normalization*) para padronizar as características. São utilizadas apenas duas camadas LSTM, a primeira com 128 unidades e a segunda com 256, seguidas por uma camada densa com 64 unidades. Aplica-se um *dropout* com uma probabilidade de $0,25$ após a camada densa. A saída é uma camada densa com três valores correspondentes às dimensões de valência, excitação e dominância. A função de ativação utilizada é a $tanh$ e o otimizador é o Adam, com uma taxa de aprendizado de $0,001$. Para realizar a concatenação dentro da rede LSTM, é necessário que as duas *features* tenham a mesma dimensionalidade. Como o modelo MiniLM L3 gera representações com 384 dimensões, duas abordagens foram avaliadas: aplicar o algoritmo *Principal Component Analysis* (PCA) nas representações geradas e realizar o ajuste fino do modelo, alterando a camada de saída para o tamanho de 128. O ajuste fino utilizado é baseado na disponibilização da biblioteca sentence transformer [1], a qual foca na tarefa de similaridade semântica. Para isso, o conjunto de dados AllNLI [2] foi utilizado.

3. Resultados

Para a avaliação, foi considerado o Coeficiente de Correlação de Concordância (CCC). Quanto mais perto de 1, melhor o resultado. Como baseline foi considerado o trabalho relacionado proposto por [Atmaja and Akagi 2020], que avalia no mesmo conjunto de dados o uso de uma rede LSTM, considerando para a representação textual o *embedding* de palavras GloVe combinado com o uso da biblioteca pAA para a extração de *features* de áudio.

Tabela 1. Comparação entre o baseline e as duas propostas apresentadas.

Modo	CCC			
	Valência	Excitação	Dominância	Média
Baseline [Atmaja and Akagi 2020]	0.418	0.571	0.500	0.496
PCA [Guder et al. 2024]	0.1431	**0.5915**	**0.5899**	0.4415
Ajuste Fino	**0.444**	0.5731	0.5728	**0.5299**

Enquanto [Atmaja and Akagi 2020] foca em *embedding* de palavras, com GloVe, o foco deste trabalho é na captura do significado da sentença através do *embedding* do MiniLM L3. O MiniLM L3 foi testado na tarefa de Análise de Sentimento e obteve um bom desempenho no conjunto de dados *Stanford Sentiment Treebank* (SST) [Socher et al. 2013]. O *embedding* textual visa aprimorar a dimensão de valência, uma vez que a tarefa é semelhante à análise de sentimento, abrangendo perspectivas que vão da negativa à positiva. Através do uso do ajuste fino, é possível ajustar produzir diretamente os *embeddings* com 128 dimensões, isso permite manter o máximo de informação possível. Ao alterar de forma externa ao modelo, acaba-se por perder informações importantes no processo. Dessa forma, foi possível obter resultados superiores ao *baseline*.

[1] O acesso pode ser feito em: `https://github.com/UKPLab/sentence-transformers`

[2] O acesso pode ser feito em: `https://huggingface.co/datasets/sentence-transformers/all-nli`

4. Conclusões

O objetivo principal desta proposta foi avaliar o efeito do ajuste fino em comparação com o uso de PCA para reduzir a dimensionalidade do modelo de embedding de sentença, MiniLM L3. A utilização do ajuste fino em relação ao PCA permitiu um ganho de aproximadamente três vezes na valência. Em comparação com o baseline, obteve-se um ganho 6,83%. Visto que o VGGIsh foi treinado originalmente com dados de sons focando na tarefa de classificação, pode-se obter um ganho maior ao utilizar modelos de *embedding* de áudio para predição de dimensões. O conjunto de dados IEMOCAP é atuado, como trabalhos futuros sugerem, pela aplicação desta abordagem no conjunto de dados MSP-PODCAST, que possui dados de fala natural. A falta de conjuntos de dados em português impede a replicação deste trabalho na língua portuguesa.

Referências

Atmaja, B. and Akagi, M. (2020). Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing*, 9.

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*, 42:335–359.

de Lope, J. and Graña, M. (2023). An ongoing review of speech emotion recognition. *Neurocomputing*, 528:1–11.

Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Powers, M. J., editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley.

Geetha, A., Mala, T., Priyanka, D., and Uma, E. (2024). Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Information Fusion*, 105.

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.

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

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, 14:261–292.

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

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.