Posthoc Disentanglement of Textual and Acoustic Features in Self-Supervised Speech Encoders

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, Afra Alishahi, Ivan Titov

ICML International Conference On Machine Learning · Tilburg University · University of Amsterdam · The University of Edinburgh

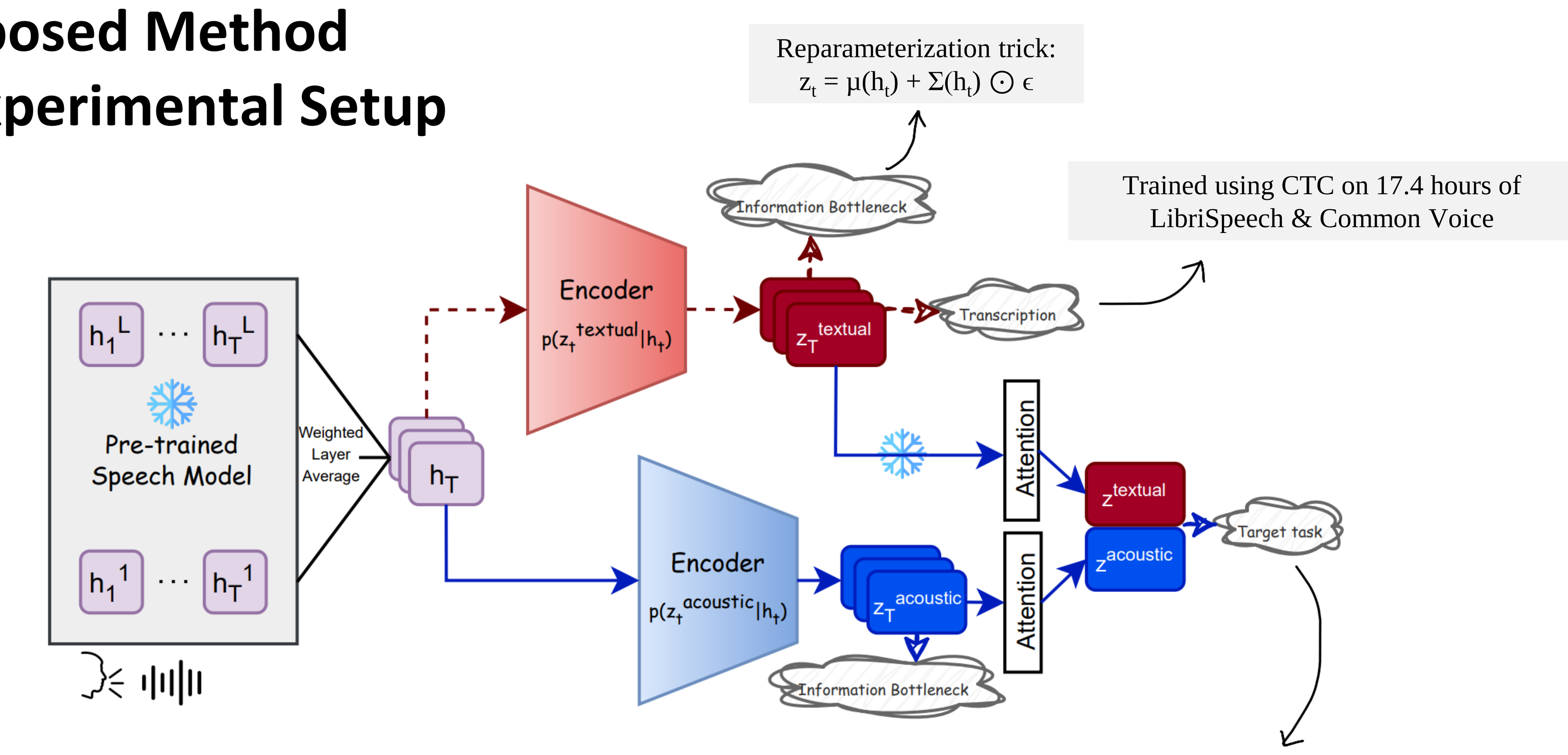
Motivation

In speech, both content (what can be transcribed as text) and acoustic features could contribute to a target task.

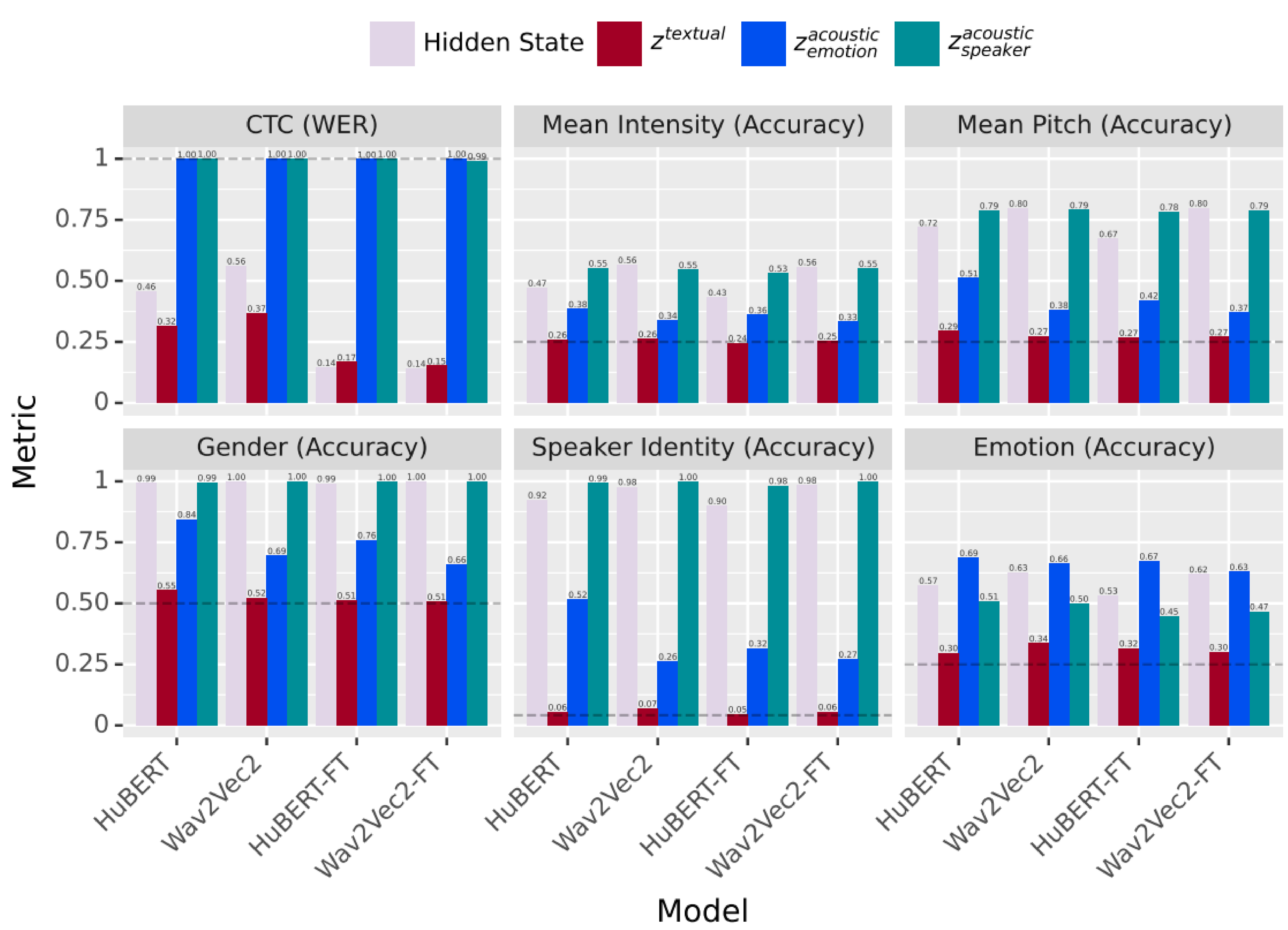
We propose a **cascaded framework** based on Information Bottleneck that disentangles the textual and acoustic features of speech representations while satisfying the following desiderata:

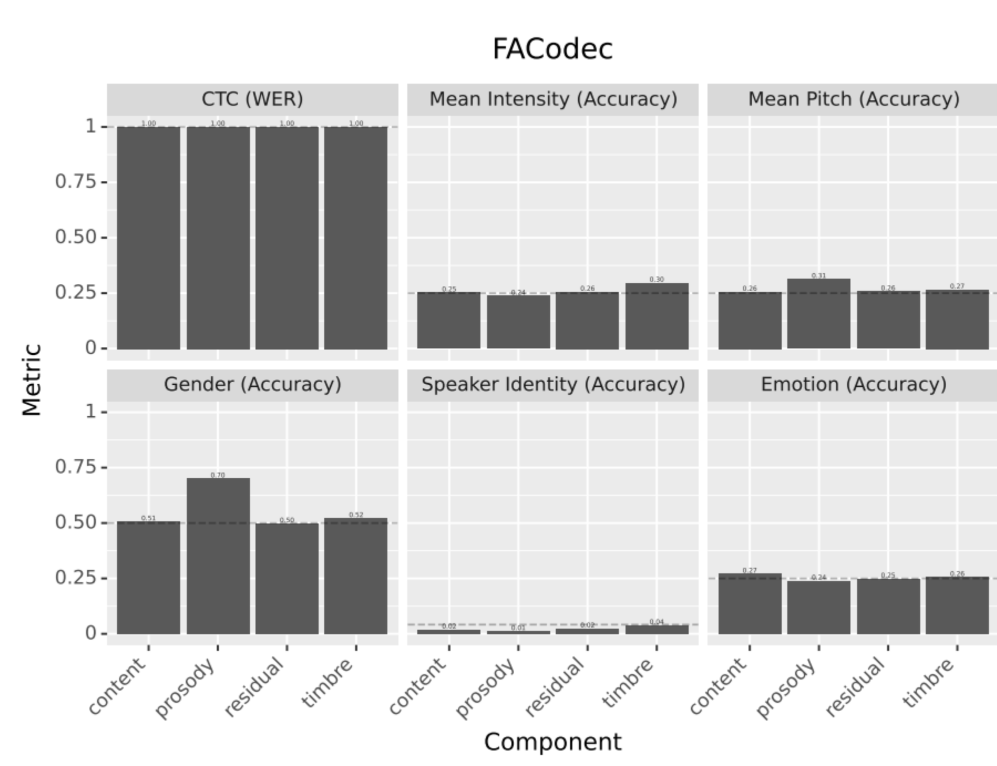
- ✓ **Post-hoc**: The approach must disentangle representations learned by pre-existing, pre-trained models, with minimal data or computation.
- ✓ **Model-agnostic**: The approach must be general and applicable to models with varying architectures, sizes, and learning objectives.
- ✓ **Task-relevant**: The approach should allow for acoustic features to emerge based on their relevance to a target downstream task, rather than being based on pre-specified static components (such as pitch, timbre, and speaker).

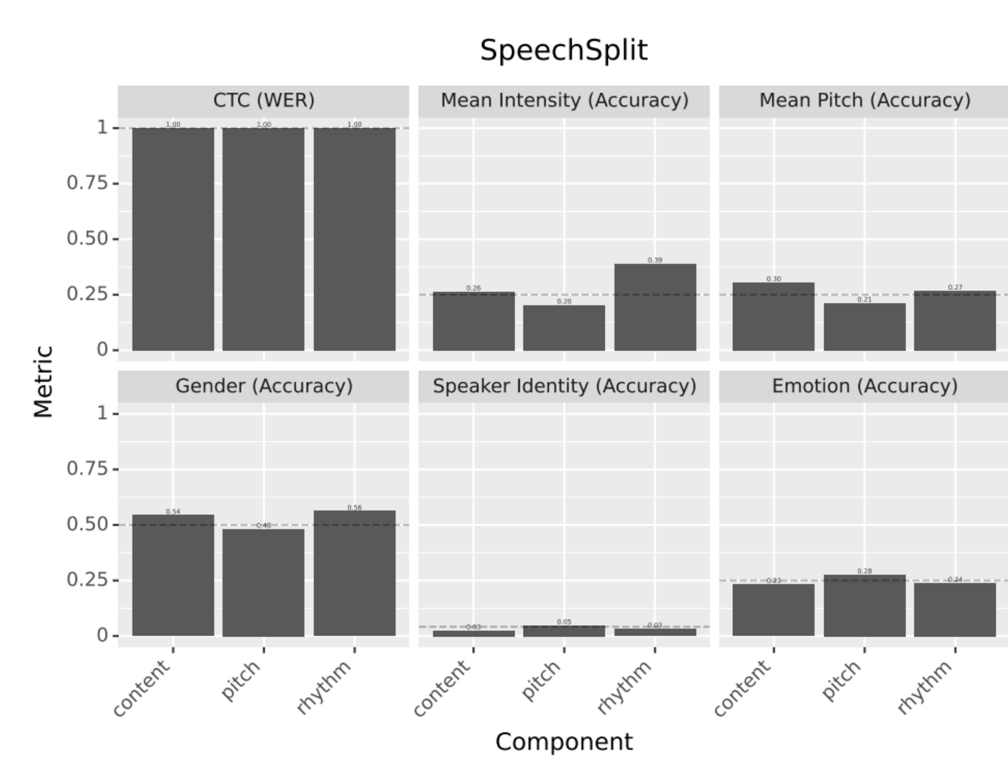
Proposed Method & Experimental Setup



- **Emotion**: IEMOCAP {Angry, Happy, Neutral, Sad}
- **Speaker Identification**: A subset of Common Voice (24 speakers)







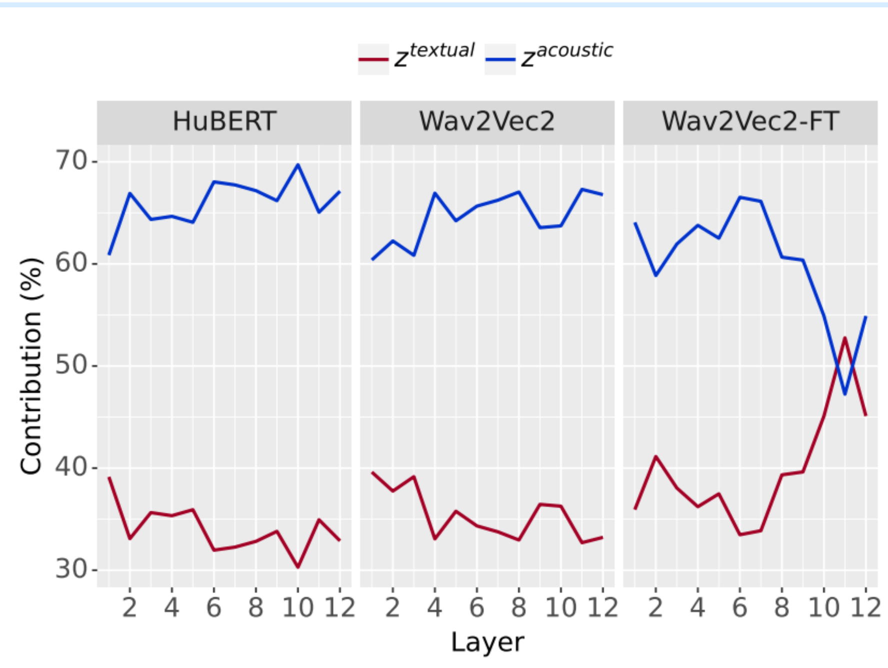
Evaluation of disentanglement

We find the representations are almost perfectly disentangled from each other:

- Textual representations predict transcriptions as well as the original speech representations but fail at predicting acoustic features (**Intensity**, **Pitch**, **Gender**, and **Speaker Identification**).
- Acoustic representations excel at predicting acoustic features but perform randomly at transcriptions.

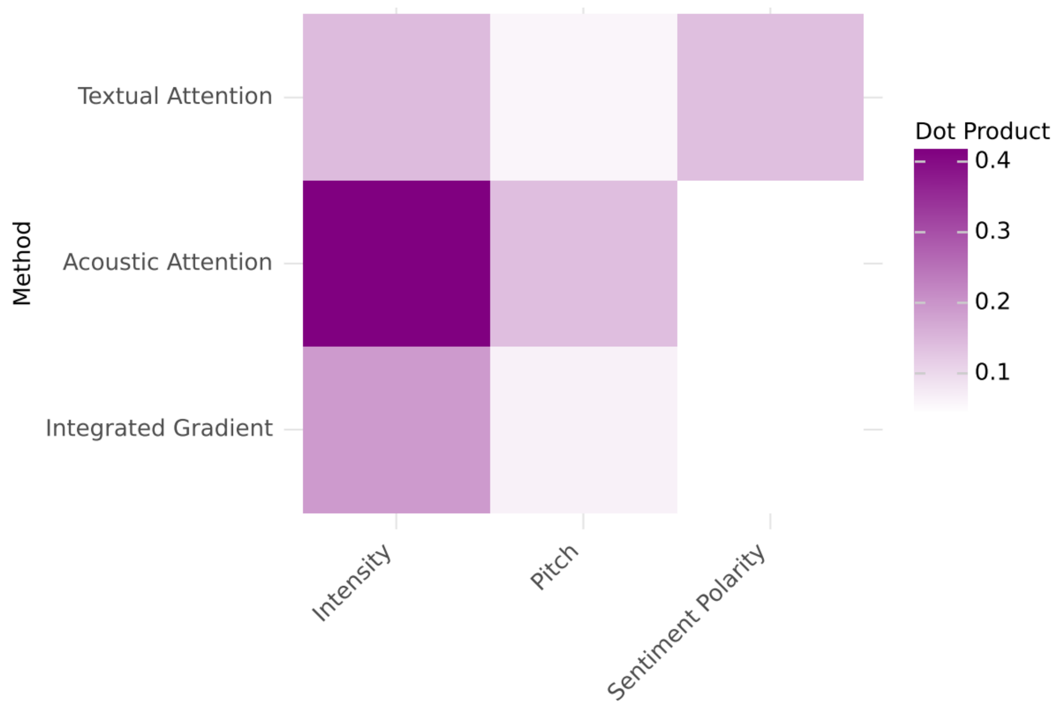
Layerwise Emotion Contribution

Using probing analysis, our framework reveals something we couldn't show without disentangling: **Through layers, the acoustic contribution to emotion recognition significantly decreases when models are fine-tuned on ASR, while the textual contribution increases.**



Disentangled Feature Attribution

- The proposed disentanglement framework can serve as a feature attribution method: **it allows us to determine whether a frame's contribution is textual or acoustic.**
- Acoustic attention captures peaks and valleys in acoustic features, while textual attention focuses on word polarity. Both have higher agreement with features than **Integrated Gradient** scores.



Reproducibility

The textual encoder learned in stage 1 is independent of the target task and can be applied to new downstream tasks; try it out on GitHub:

Paper

Code