Leveraging Small Languages Models for Unsupervised Speech Emotion Recognition : Using a Lexical-Paralinguistics Hybrid Approach.

Toh Jian Hong
Nanyang Technological University
jtoh066@e.ntu.edu.sg

Abstract

Recent advancement in Small Languauge Models (SLMs) offer promising avenues for machine learning problems. Unsupervised speech emotion recognition (SER) remains a challenging task due to complex interplay between acoustic cues and linguistic content. In this study, we explore a hybrid framework that leverages both lexical features – derived from sentence embedding on transcriptions using small language models (SLMs) – and paralinguistic features extracted directly from audio signals. We hypothesize the benefit of the hybrid approach is that by leveraging both how something is said and what is being said, we are able to improve the model. For instance, the words "I can't believe this" may be uttered in anger or excitement, paralinguistics alone may struggle to tell them apart without lexical content. By concatenating the two modalities, we strive to improve the clustering of speech without relying on labelled data. Through empirical evaluation across multiple configuration and feature scores, we demonstrate that integrating lexical information can lead to a more semantically coherent cluster compared to just paralinguistic information alone. Highlighting the potential of hybrid approach for unsupervised SER task in other domains.

Contents

1. Introduction

Recent advancement in Natural Language Processing (NLP) has led to the development of powerful yet efficient Small Language Models (SLMs). The industry trend and advancement has democratized access to these cost effective SLMs, presenting an intriguing possibility. Can these language models solve traditional machine learning problems such as speech emotion recognition (SER) while circumventing the need for labeled data ?

So why not an end to end SLM prediction model ? The answer is Prosody Ignorance. Just using a pure SLM would discard critical acoustic emotion cues from paralinguistic features and this would greatly limit the model's clustering ability. Hence we adopt the approach to fuse the two features to reach a cohesive model.

Traditionally, SER systems rely on supervised learning with annotated datasets, prioritizing paralinguistic features over lexical content. Though effective, the traditional approach has costly data dependence. This supervised approach can create significant barriers to entry as labelled SER datasets require costly and extensive annotation making them difficult to come by in other languages and domains.

This project challenges the notion by proposing an unsupervised hybrid SER framework that synergizes lexical features and paralinguistic cues. Our main objective would be to determine whether the incorporation of a SLM to the unsupervised approach would improve the performance of clustering for the unsupervised model, in hopes of circumventing the need for labelled data and traditional supervised models.

2. Background

Speech emotion recognition as a field has developed and evolved over time, encompassing deep learning techniques. In this section, we review both supervised and unsupervised approaches to SER, providing a foundation for our exploration into unsupervised learning methods.

2.1 Supervised Learning

Traditionally, supervised learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and variations like Long Short-Term Memory (LSTM) networks have been used. They adopt the standard backpropagation-based training process, where the models predictions are compared to the validation labels and models weights are updated iteratively to reduce loss.

In contrast our study explores unsupervised learning, which does not require labelled data and therefore does not involve backpropagation through clustering.

2.2 Unsupervised Learning - K Means Clustering Algorithm

K-Means is a centroid based clustering algorithm that partitions the data into K clusters, where each data point would belong to the cluster with the nearest mean (centroid). Advantages are that it is quick and scalable. However this predefined number also means that emotions like sadness that may exist on a spectrum of intensity will be over simplified into a single cluster losing their nuances. (Kavlakoglu & Winland, 2024)

2.3 Unsupervised Learning - HDBSCAN.

Hierarchical Density Based Spatial Clustering of Application with Noise (HDBSCAN) is a clustering technique improved upon the DBSCAN. It constructs a hierarchy of clusters based on density and extracts a flat clustering from it. Its strength being that it is able to identify clusters of varying densities, making it well suited for complex, high-dimensional data. Unlike the K-means, HDBSCAN does not specify the number of clusters beforehand, allowing us to explore and discover the subtle and previously undefined nuances of emotions in speech.

2.4 Small Language Model

SLMs are streamlined versions of LLMs, optimized for efficiency and suitability in resource-constrained environments. Typically ranging from 1 million to 10 billion parameters, SLMs are significantly smaller in scale compared to LLMs. Despite their reduced size, they retain essential natural language processing (NLP) capabilities such as contextual understanding and semantic representation. This makes them a practical choice for applications that require robust language understanding, like SER, without the computational demands of larger models. (Small Language Models (SLM): A Comprehensive Overview, 2025)

3. Methodology

This chapter outlines the methodology employed for implementing the unsupervised SER system. The core of our system is based on the combination of multiple models for feature extraction and clustering. The methodology is designed in theory to leverage both lexical and paralinguistic cues from speech data for SER . This hybrid approach attempts to integrate the SLM embedding and OpenSMILE for comprehensive feature extraction.

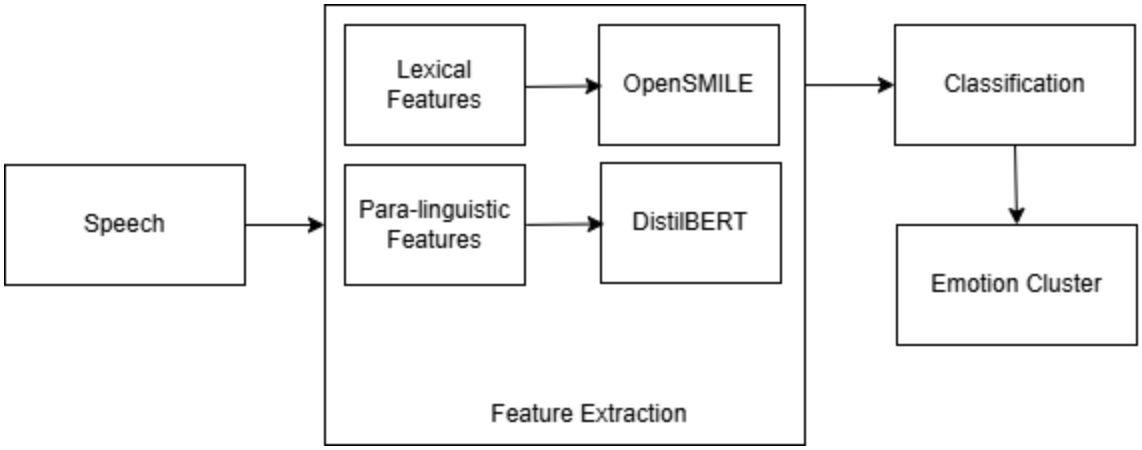


Figure 1 : The Hybrid approach pipeline

3.1 Model implementation

The Audio file is first passed through and undergone paralinguistic extraction, followed by a transcription of the audio file itself, and a SLM model proceeds to embed the transcription with lexical cues. These features are then combined into a combined feature set where standardization and dimensionality reduction is applied before clustering. Lastly we used various performance metrics at each point to select the best performing model.

3.1.2 Feature Extraction

Paralinguistic features - refer to aspects of speech that convey meaning through characteristics like tone, pitch, and rhythm, rather than through the textual content itself. To explore these features, we compared two tools: Librosa and OpenSMILE.

In our initial experiments, we extracted a variety of acoustic features, such as Mel-frequency cepstral coefficients (MFCCs), using both libraries. However, after comparing the performance of these tools using our evaluation metrics (Figure 2), we found that OpenSMILE outperformed Librosa.

This could be because OpenSMILE is specifically designed for speech and paralinguistic analysis, offering a more extensive and specialized range of features tailored for tasks like emotion recognition, speaker profiling, and affective computing. In contrast, while Librosa is excellent for audio analysis in music and general sound processing, it lacks the focused feature sets and optimizations that are critical for analyzing paralinguistic cues.

As a result, we opted for OpenSMILE's eGeMAPSv01a configuration, which includes a compact set of 88 features that have been extensively validated for their relevance in paralinguistic tasks, making it an ideal choice for our study.

Lexical features - features convery through textual content. The transcription of each audio clip was performed using Wav2Vec2. The Wav2Vec2 model was used to transcribe the audio into

text which was then passed through a SLM – specifically DistilBERT – to generate sentence embedding, representing the semantic content of the speech. These generated embedding help represents what is being said in terms of meaning, emotionally challenging words or contextual sentiment – which can be useful in distinguishing emotions like anger vs excitement, or sadness vs boredom, that might sound acoustically similar.

3.1.3 Feature Fusion

To leverage both paralinguistic and lexical features obtained, we concatenated the features to create a combined feature set, This approach allows the model to jointly reason over what is being said and how it is said. From a neurocognitive perspective, this mirrors the dual-stream model of speech processing in the human auditory cortex, where lexical and prosodic information are processed in parallel along the ventral and dorsal streams respectively (Hickok, 2012).

3.2 Small Language Models (SLM)

SLM democratized the LLMs by offering a resource efficient alternative to large language models, enabling the use of deep contextual embedding without significant computational overhead. In our experiment we evaluated popular models, BERT, DistilBERT and RoBERTA, to determine which yield the most effective representation for emotion recognition in speech. Based on the clustering performance (Figure 7) and our metrics below (Figure 2), the DistilBERT model was selected to be the most effective. It achieved the highest Silhouette score and the second lowest Davies-Bouldin Index and the second highest Calinski - Harabasz Index. Meaning that clusters are pretty well defined.

	Model	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
0	BERT	0.587290	0.633921	11276.040039
1	DistilBERT	0.611604	0.624821	12959.646484
2	RoBERTa	0.593878	0.620112	13825.484375

Figure 2 : Comparison of the 3 SLM during clustering on the combined features dataset

3.2 Dataset

For our study, we have chosen the RAVDESS dataset (Livingstone, 2018) and the CREMA-D dataset (CheyneyComputerScience/CREMA-D, 2023). These datasets were selected to make up for the weakness of the RAVDESS dataset where the same speech is used; this does not provide enough lexical information. The combination will hence challenge our model's adaptability across varying domains within the SER landscape. Together the total combined dataset of RAVDESS and CREAM-D contains 8882 audio files providing a robust foundation for modeling and evaluation.

RAVDESS Dataset : The data set contains speech that is repeated with different tones and intonations. This provides us with essential paralinguistic information.

CREMA-D Dataset : The data set holds various speeches made by mank speakers. This set of different speech provides various key lexical data for our SLM to work with.

3.3 Training Configuration :

After our features were extracted from both the paralinguistic and lexical models, dimensionality reduction and standardisation was performed using UMAP (Uniform Manifold Approximation and Projection) to map the features into a lower-dimensional space. This not only aids in visualization but also helps to improve the clustering process.

RAVDESS + CREMA-D Dataset:

	Method	Feature Set	Clustering	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
0	librosa	KMeans	0.43	0.744	47408.2
1	librosa	HDBSCAN	0.755	0.391	1368.04
2	OpenSMILE	KMeans	0.448	0.681	50399.5
3	OpenSMILE	HDBSCAN	0.783	0.199	31367

Figure 3 : Comparison between the metrics scored and the method and feature set used (only on paralinguistic features)

We experimented with the two clustering algorithms, HDBScan and K-Means. Although HDBSCAN initially yielded better quantitative results on OpenSMILE - Only features (Figure-3). HDBSCAN achieved the highest silhouette score and a lowest DBI. Further inspection revealed that it had the tendency to cluster data efficiently, with a lot of it being noise points, looking at the UMAP plot on the OpenSMILE-Only features (Figure 4) the majority of the data is clustered into one group while the rest is noise points. HSBSCAN was able to identify some clusters in the Librosa feature set however we can still see that a large chunk of it is noise points. This outcome, while numerically attractive when focusing on the metrics, lacked practical utility for our task. Consequently, we opted for the K-Means algorithm, which provided more interpretable and balanced cluster assignments even with relatively worse metrics (Figure 4).

3.4 Evaluation Metrics:

To ensure a comprehensive evaluation of our model's performance, we will be utilizing a range of metrics for our analysis.

1. **Silhouette score :** Measure of how similar a point is to its own cluster compared to other clusters. Ranges from -1 to 1. Closer to 1 = better defined clusters. Negative = Likely assigned to the wrong cluster. Close to 0 = overlapping clusters.
2. **Davies-Bouldin Index (DBI)** : Ratio of intra-cluster distance to inter-cluster separation. A score closer to 0 indicates better clustering.
3. **Calinski-Harabasz Index (CHI) :** Also known as the Variance Ratio Criterion (VRC). Ratio of between-cluster dispersion to within-cluster dispersion. The higher the better.

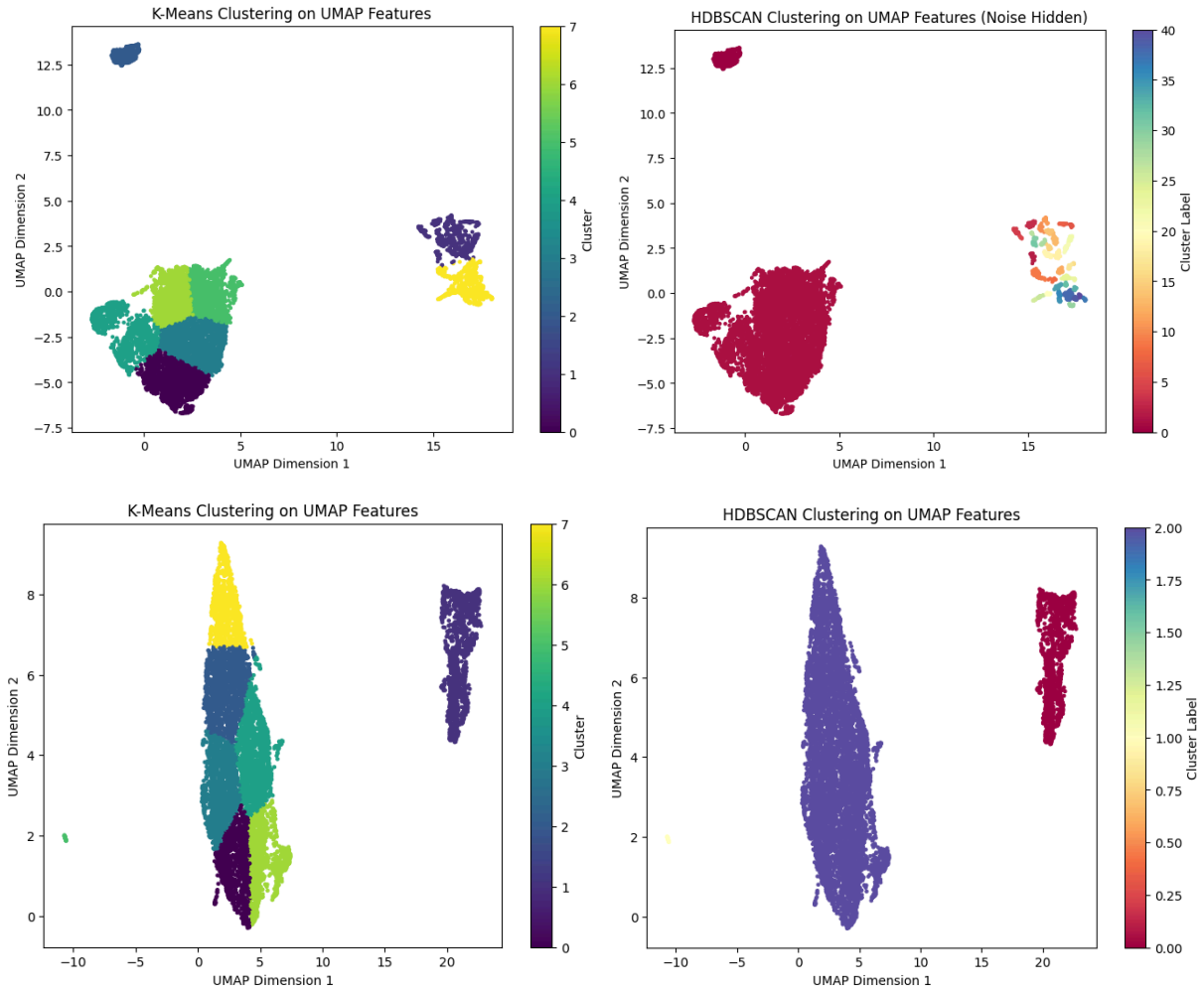


Figure 4 : Top row contains Librosa feature extraction, Bottom row contains OpenSMILE feature extractions (only on paralinguistic features)

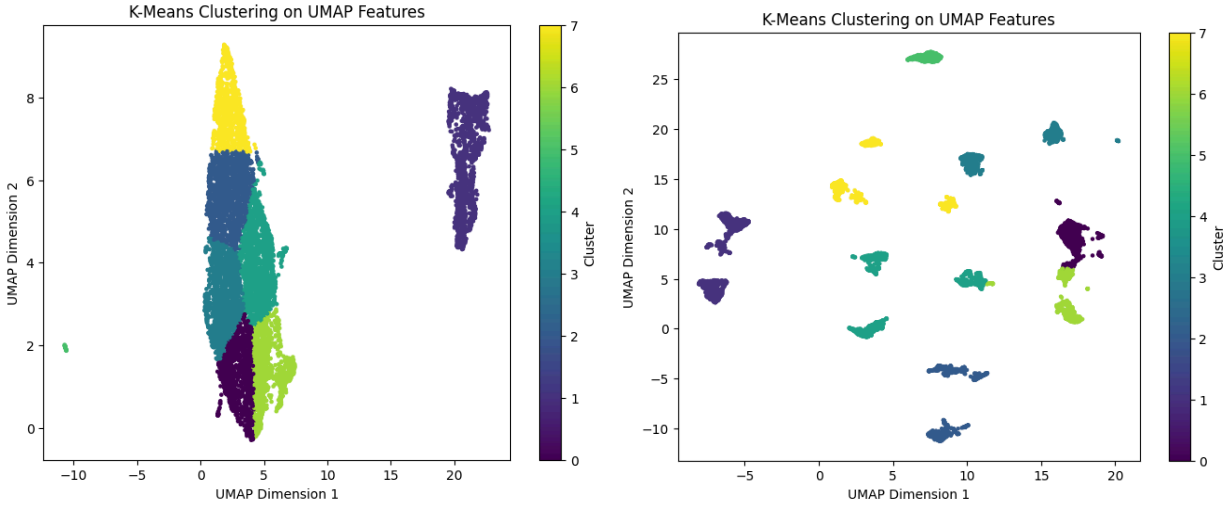
4. Experiments

The goal of this experiment was to evaluate the effectiveness of combining paralinguistic and lexical features for unsupervised SER. Specifically focusing on assessing how feature extractors and clustering algorithms influence the performance of the system, with the ultimate aim of identifying optimal combinations that maximize clustering quality and the system's ability to identify emotions.

4.1 Ablation Study : Performance of SLMs incorporated features compared with Base Models.

We compared the performance of models that used paralinguistic features alone with models that integrated lexical features using Small Language Models (SLMs) like DistilBERT. We focused on measuring the impact of these feature combinations on the quality of clustering and the overall effectiveness of the unsupervised system in recognizing emotions.

Baseline : The baseline model consisted of a system where only paralinguistic features (extracted using OpenSMILE-Only) were used for clustering, with KMeans as the clustering algorithm.



Silhouette Score (opensmile + KMeans): 0.4480889
Davies-Bouldin Index (opensmile + KMeans): 0.6807895024155116
alinski-Harabasz Index (opensmile + KMeans): 50399.48

Silhouette Score (opensmile + KMeans): 0.59267473
Davies-Bouldin Index (opensmile + KMeans): 0.5946751875304629
alinski-Harabasz Index (opensmile + KMeans): 16527.426

Figure 5 : KMeans Clustering on Paralinguistic features only on the left and Combined Features on the right.

The baseline model performs well for an initial attempt, with a reasonably distinct cluster, although some overlap remains, this could mean that some data points correspond to similar emotions. This overlap also suggests the system may benefit from additional information beyond paralinguistic features alone. After incorporating lexical features via a SLM, we saw an improvement in clustering performance (Figure 5). The SLM provided efficient and informative embeddings for the transcription text, which captures the semantic and syntactic element of speech, helping the model distinguish between emotional states that share similar paralinguistic cues but differ in content.

We applied the UMAP for dimensionality reduction as well. This was followed by K-Means clustering. As a result, we did observe an improvement in clustering quality, reflected in the significant increase in Silhouette score from 0.44 to 0.59 and a lowered Davies-Bouldin Index from 0.68 to 0.59. Alinksi-Harabasz score unfortunately decreased.

These results highlight the improvement made by the combining of paralinguistic and lexical features in unsupervised emotion recognition tasks. The enhanced feature set allows the system to more accurately differentiate emotional states by capturing both the emotional tone and meaning behind it.

After verifying that our model was able to improve with the SLM, we opted to optimized this through comparison between three different SLMs, BERT, DistilBERT and RoBERTa. From our

experiments we managed to improve the silhouette score. Our best model is DistilBERT and OpenSMILE, achieving a Silhouette value of 0.612 (Figure 6).

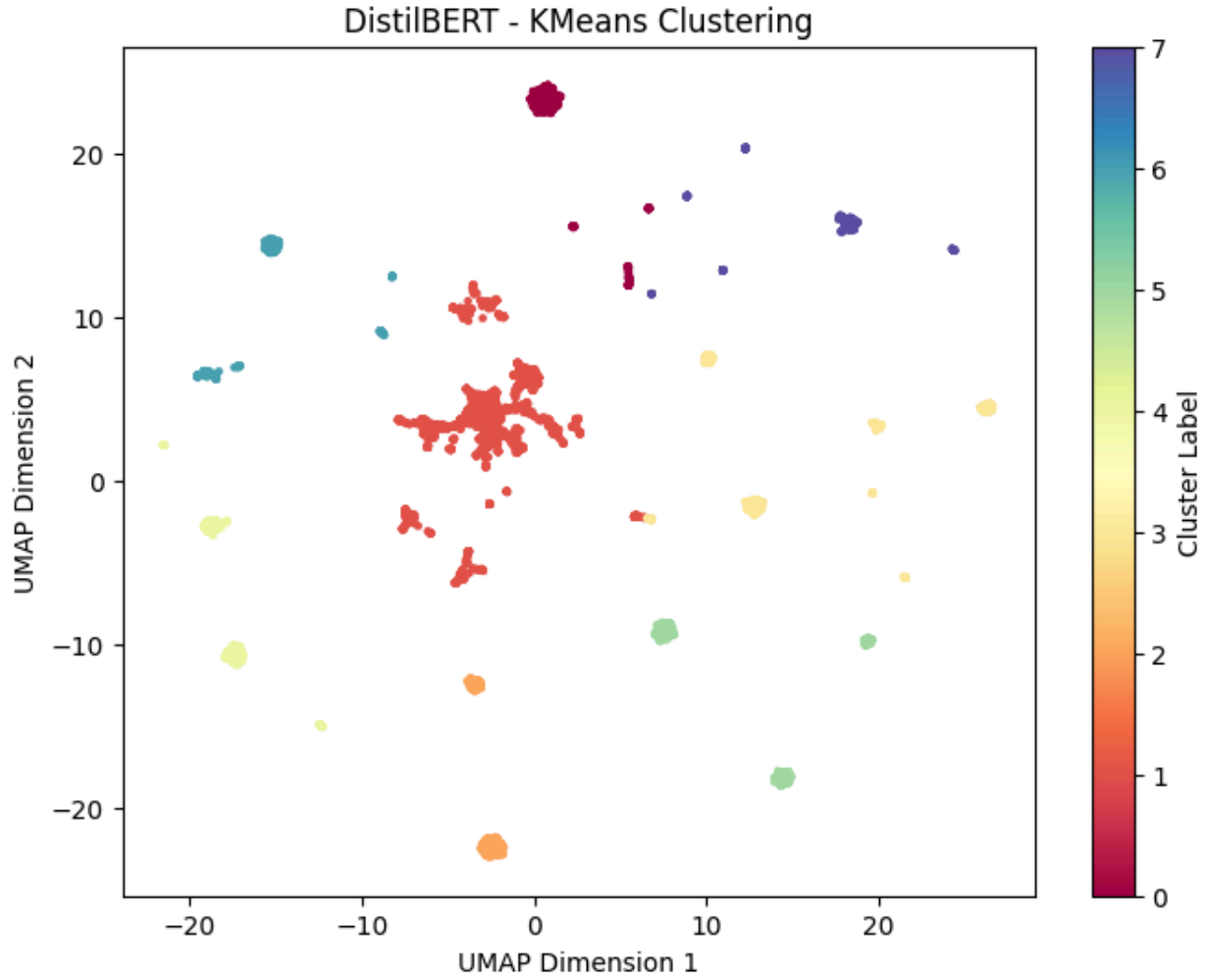


Figure 6 : UMAP of our KMeans-Cluster on OpenSMILE and DistilBERT Features

5. Limitation and Future Work

While our study demonstrates the viability of SLMs for SER, several limitations warrant discussion. Addressing these could yield more robust and generalizable models. We shall attempt to discuss some areas that can be improved and worked on in this section.

5.1 Hyperparameter Tuning

Hyperparameter Tuning is essential in optimizing a model's performance; it demands a significant time- and resource-intensive process. Due to constraints we were not able to perform it enough. Given more time and resources, we can optimize the parameters for the models to achieve better results for evaluation.

5.2 Variety of Dataset

Our dataset was small due to resource constraints, RAVDEES dataset especially did not have much lexical content to work with. Hence we combined it with the CREMA-D dataset. However there is still limited lexical data. The model's accuracy could be improved through increasing the quantity and variety of the speeches.

5.3 Mono-Lingual Bias

Our datasets were limited to speech data in English. Using a plethora of datasets ranging from different languages would test our theory more comprehensively. The ability of our model to adapt to different languages and maybe even other domains remain to be explored as well.

5.4 SLM

In our study we focused only on 3 SLM Models; BERT, DistilBERT, RoBERTa, given more time we can maximize the efficiency and cost by exploring the impact that scaling down a larger variety of these models will bring. More metrics could also be used to evaluate the models as well.

5.5 Training time and computational resources

One important limitation of our study is the lack of analysis on the additional computational and time overhead introduced by the hybrid approach. Unlike the base unsupervised SER pipelines, our method includes a transcription step followed by the extraction of lexical features using small language models (SLMs), which incurs significant additional processing.

While our experiments demonstrate performance improvements in clustering quality, we did not quantify the trade-offs in terms of training time or computational cost. A comprehensive evaluation should include these factors to assess the true cost-effectiveness of the hybrid method. Due to resource constraints, we focused solely on performance gains, leaving overhead analysis as an area for future investigation in the future.

6. Conclusion

Our investigations into the incorporation of SLMs into unsupervised SER has yielded valuable insights. The experimental results demonstrate that a hybrid approach combining SLM generated lexical features with paralinguistic features outperforms a baseline model that relies solely on acoustic information. This enhanced performance highlights the potential of leveraging SLMs which in the case of SER, was able to help distinguish between emotions that may present similar acoustic patterns.

Despite the promising result, our work was greatly restricted due to several limitations and assumptions. In particular, restricted dataset, limited variety and diversity, and absence of extensive hyper parameter tuning prevent us from fully exploring the trade-offs between computational resources and model performance. Future works could explore cross domain

efficacy, larger and more diverse dataset, deeper model optimization, and true cost trade-offs of applying this approach.

Ultimately, this report advocates for a strategic use of SLMs in the domain of SER systems. By harnessing the use of SLMs, we can mitigate the reliance on labeled data and achieve competitive performance. One must carefully weigh the additional computational overhead against the benefit of the semantic understanding. With a balanced approach, the lack of labeled data can be effectively compensated, paving the way for more resource efficient and scalable solutions.

References & Citations

Thirupathi Kandadi, & G. Shankarlingam. (2025). DRAWBACKS OF LSTM ALGORITHM: A CASE STUDY. https://doi.org/10.2139/ssrn.5080605

Kavlakoglu, E., & Winland, V. (2024, June 26). K-Means Clustering. IBM. https://www.ibm.com/think/topics/k-means-clustering

Aouani, H., & Ayed, Y. B. (2020). Speech Emotion Recognition with deep learning. Procedia Computer Science, 176, 251–260. https://doi.org/10.1016/j.procs.2020.08.027

How HDBSCAN Works — hdbscan 0.8.1 documentation. (n.d.). Hdbscan.readthedocs.io.https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. Journal of Communication Disorders, 45(6), 393–402. https://doi.org/10.1016/j.jcomdis.2012.06.004

Small Language Models (SLM): A Comprehensive Overview. (2025, February 25). Huggingface.co. https://huggingface.co/blog/jjokah/small-language-model

Livingstone, S. R. (2018). RAVDESS Emotional speech audio. Kaggle.com. https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio/data

CheyneyComputerScience/CREMA-D. (2023, January 20). GitHub. https://github.com/CheyneyComputerScience/CREMA-D

Appendix

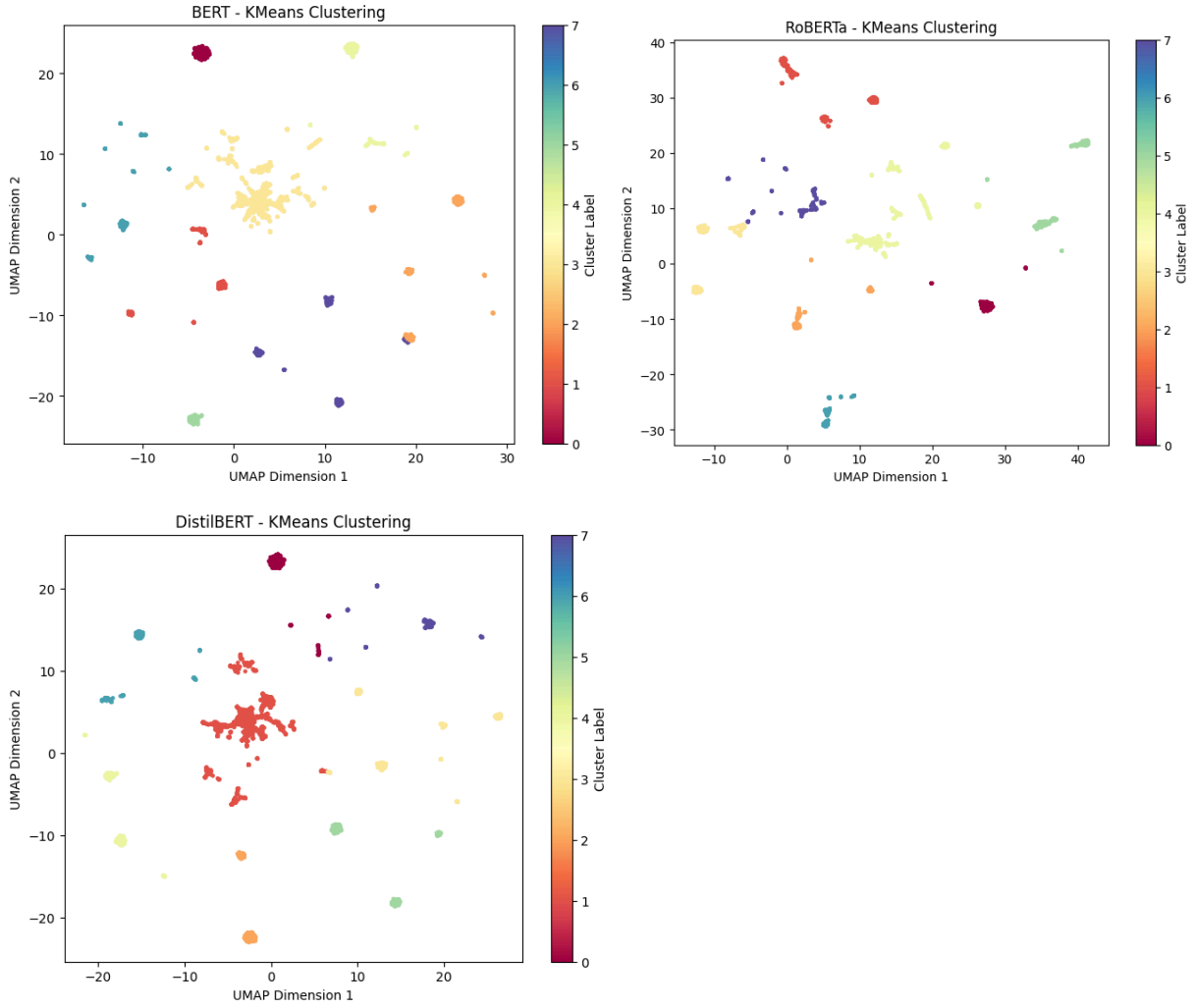


Figure 7 : Comparison between the SLM models (BERT, DistilBERT and ROBERTA) on paralinguistic features of openSMILE.

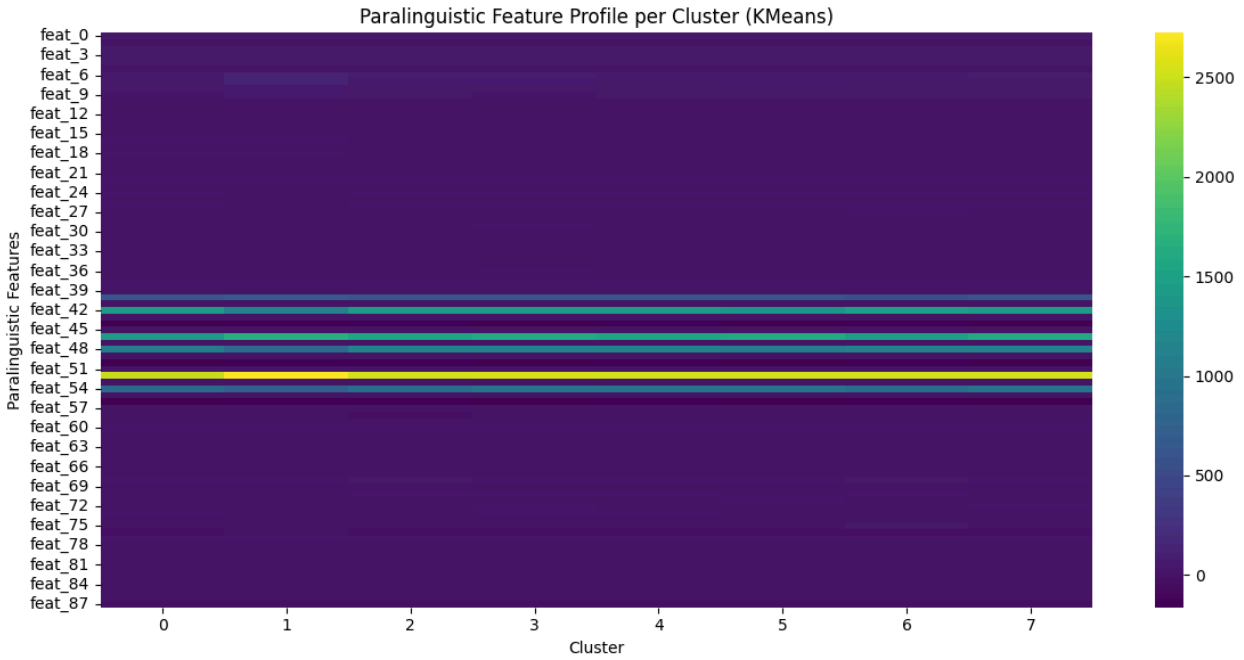


Figure 8 : Distribution of paralinguistic features (OpenSMILE has 88 Features) on our model.

cluster	feat_0	feat_1	feat_2	feat_3	feat_4	feat_5	\
0	32.056789	0.079926	30.308664	32.329399	33.958672	3.650009	
1	34.156948	0.121158	31.493690	34.253414	37.168217	5.674525	
2	31.883181	0.096303	29.778881	31.797693	34.216110	4.437226	
3	32.258999	0.105132	29.454798	32.574249	35.043789	5.588992	
4	31.600370	0.094469	29.419828	31.645306	33.974365	4.554535	

cluster	feat_6	feat_7	feat_8	feat_9	...	feat_78	\
					...		
0	60.886192	34.117424	29.200365	23.488823	...	-0.007843	
1	139.544388	130.579346	57.662884	56.507103	...	0.044697	
2	78.631378	59.626877	44.782928	45.225239	...	-0.002539	
3	78.128052	51.139618	40.948639	24.708805	...	-0.005291	
4	66.126053	44.564438	39.565762	30.316721	...	-0.006147	

cluster	feat_79	feat_80	feat_81	feat_82	feat_83	feat_84	feat_85	\
0	-0.004122	0.115246	2.907008	1.485894	0.171557	0.107434	0.481492	
1	0.008808	0.025742	2.269812	1.233378	0.328192	0.194737	0.474874	
2	-0.010532	0.112529	2.895634	2.106211	0.170177	0.111810	0.286269	
3	-0.007011	0.103016	2.996976	2.048981	0.117613	0.080778	0.356197	
4	-0.007033	0.097811	3.066076	2.014847	0.177862	0.101269	0.333949	

cluster	feat_86	feat_87
0	0.293948	-30.716372
1	0.433199	-39.651974
2	0.248626	-30.443975
3	0.278217	-32.536087
4	0.276675	-31.563303

[5 rows x 88 columns]

Figure : Top 5 samples of the paralinguistic features from each cluster on the dataset.

```
Cluster 0:
  Example 1: IT'S ELEVEN O'CLOCK
  Example 2: IT'S ELEVEN O'CLOCK
  Example 3: IT'S ELEVEN O'CLOCK

Cluster 1:
  Example 1: I WANT TOSABOUT
  Example 2: I THINK I HAVE A DACTOR'S APARTME
  Example 3: I THINK I'VE SEEN THIS BEFOER

Cluster 2:
  Example 1: DOGS ARE SITTING BY THE DOOR
  Example 2: DOGS ARE SITTING BY THE DOOR
  Example 3: AND DOGS ARE SITTING BY THE DOOR

Cluster 3:
  Example 1: I WONDER WHAT THIS IS ABOUT
  Example 2: I WONDER WHAT THIS IS ABOUT
  Example 3: I WONDER WHAT THIS IS

Cluster 4:
  Example 1: KIDS ARE TALKING BY THE DOOR
  Example 2: KIDS ARE TALKING BY THE DOOR
  Example 3: KIDS ARE DOGGING BY THE DOOR

Cluster 5:
  Example 1: THAT IS EXACTLY WHAT HAPPENED
  Example 2: THAT IS EXACTLY WHAT HAPPENED
  Example 3: THAT IS EXACTLY WHAT HAPPENED

Cluster 6:
  Example 1: DON'T FORGET A JACKET
  Example 2: DON'T FORGET A JACKET
  Example 3: DON'T FORGET A JACKET

Cluster 7:
  Example 1: I THINK I HAVE A DOCTOR'S APPOINTMENT
  Example 2: I THINK I HAVE A DOCTOR'S APPOINTMENT
  Example 3: I THINK I HAVE A DOCTOR'S APPOINTMENT
```

Figure 9 : The transcription of the represented data of each cluster taken nearest to the centroid.