Audio Sentiment Analysis

Project Vader

Workflow


Features Extraction
Feature Analysis
Feature Selection
State-of-Art Research
Datasets Gathering
Data and Bias Analysis
Evaluation Strategy Definition
Classification Models Testing
Evaluation Overview of Classifiers

State-of-Art Research

SOA Articles Related to Sentiment Analysis

With a Speech Component Overview

Article	Publication Date	Audio Components Analyzed	Model(s) Used	Datasets & Accuracies
Multimodal Sentiment Analysis Using Multi-tensor Fusion Network with Cross-modal Modeling	19 Nov 2021	12 MFCCs Other low-level and medium-level features	Multi-tensor Fusion Network with Cross-modal Modeling	MOSI (2): 80.9% MOSI (7): 38.92% MOSEI (2): 80.8% MOSEI (5): 49.3% MOSEI (7): 50.1%
Speech Emotion Recognition Using Quaternion CNN	31 October 2021	Mel-spectrogram features encoded in an RGB quaternion domain	Quaternion CNN	RAVDESS (8): 77.87% IEMOCAP (4): 70.46% EMO-DB (7): 88.78%
A cognitive brain model for multimodal sentiment analysis based on attention NN	21 March 2021	12 MFCCs Other low-level acoustic features	Hierarchical Attention-LSTM based on Cognitive Brain	MOSI (5): 47.8% MOSI (7): 43.33% MOSEI (5): 47.4% MOSEI (7): 47.1%
Sentiment analysis in non-fixed length audios using a Fully CNN	8 July 2021	Mel spec-trogram MFCC	Fully CNN	RAVDESS (8): 75.28% EMO-DB (7): 92.71% TESS: 99.03%
Classification of Emotive Expression Using Verbal and Non-Verbal Components of Speech	11 June 2021	MFCC Zero Crossing Rate Chroma Energy Normalised	Ensemble of CNN and Logistic Regression	RAVDESS (8): 52.00% SAVEE: 76.00%
Speech Sentiment Analysis via Pre-Trained Features from End-to-End ASR Models	14 May 2020	–	RNN with attention and specAug	IEMOCAP (4): 71.7% SWBD-sentiment: 70.10%
Speech emotion recognition with deep CNN	15 February 2020	MFCC Chroma-gram Mel-scale spectrogram Tonnetz representation Spectral contrast features	One-dimensional CNN	RAVDESS (8): 71.61% IEMOCAP (4): 64.30% EMO-DB (7): 86.10%
Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel NN	6 December 2018	MFCC Spectral centroid Spectral contrast features	LSTM and CNN	CMU-MOSI (2): 69.64% CMU-MOSI (5): 37.71% CMU-MOSI (7): 29.26% MOUD (2): 59.74%
Sentiment analysis on speaker specific speech data	22 March 2018	MFCC (Dynamic Time Wrapping(DTW) for feature matching)	VADER	Twitter: 95.2% Movie Reviews: 96%

Datasets Gathering

Prominent Speech Emotion Datasets Overview

Dataset	Year	Format	Language	Content	Emotions	Type
CMU-MOSEI	2018	Audio Video	English	65 hours of annotated video from more than 1000 speakers and 250 topics.	6 Emotions (happiness, sadness, anger,fear, disgust, surprise) + Likert scale.	Natural
MSP-Podcast	2020	Audio	English	100 hours by over 100 speakers	Activation, dominance and valence and anger, happiness, sadness, disgust, surprised, fear, contempt, neutral...	Natural
IEMOCAP	2007	Audio Video	English	Dyadic conversations among pairs of 10 speakers spanning 12 hours of dialogue scenarios.	Emotions such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation and dominance.	Acted
eNTERFACE05	2005	Audio Video	English	Videos by 42 subjects, coming from 14 different nationalities.	6 emotions: anger, fear, surprise, happiness, sadness and disgust.	Elicited
MELD	2019	Audio Video	English	1400 dialogues and 14000 utterances from **Friends TV series** by multiple speakers.	7 emotions: Anger, disgust, sadness, joy, neutral, surprise and fear. Also positive, negative and neutral.	Acted
CMU-MOSI	2017	Audio Video	English	2199 **movie reviews** with annotated sentiment.	Very negative to very positive in seven Likert steps.	Natural
TESS	2010	Audio	English	2800 recording by **2 actresses**.	7 emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.	Acted
EMO-DB	2005	Audio	**German**	800 recording spoken by 10 actors (5 males and 5 females).	7 emotions: anger, neutral, fear, boredom, happiness, sadness, disgust.	Acted
MOUD	2017	Audio Video	English	80 **product reviews** youtube videos.	Positive, negative or neutral sentiment.	Natural
RAVDESS	2018	Audio Video	English	7356 recordings by 24 actors with **2 statements only**.	7 emotions: calm, happy, sad, angry, fearful, surprise, and disgust.	Acted
MSP-Improv	2017	Audio Video	English	20 sentences by 12 actors.	**4 emotions**: angry, sad, happy, neutral.	Elicited
CREMA-D	2017	Audio Video	English	7442 clips of **12 sentences** spoken by 91 actors (gender balanced).	**6 emotions**: angry, disgusted, fearful, happy, neutral, and sad.	Acted

Datasets

eNTERFACE'05 (2005)

Elicited and multi-speaker speech database

35 male and 7 female subjects with 14 nationalities

1 GB of data

Used for the **feature selection and analysis** for being a small and well-rounded database

IEMOCAP (2007)

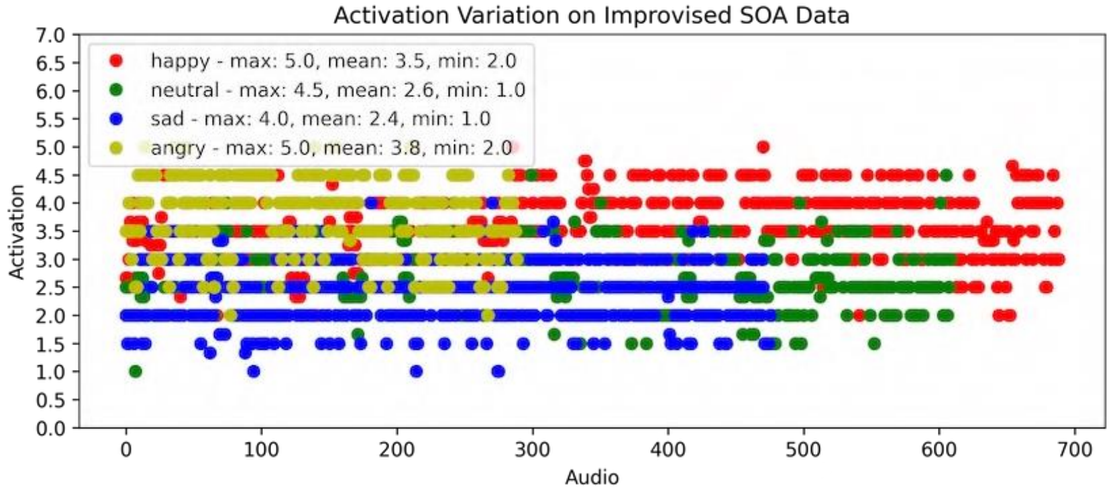
Acted and elicited multimodal and multi-speaker database

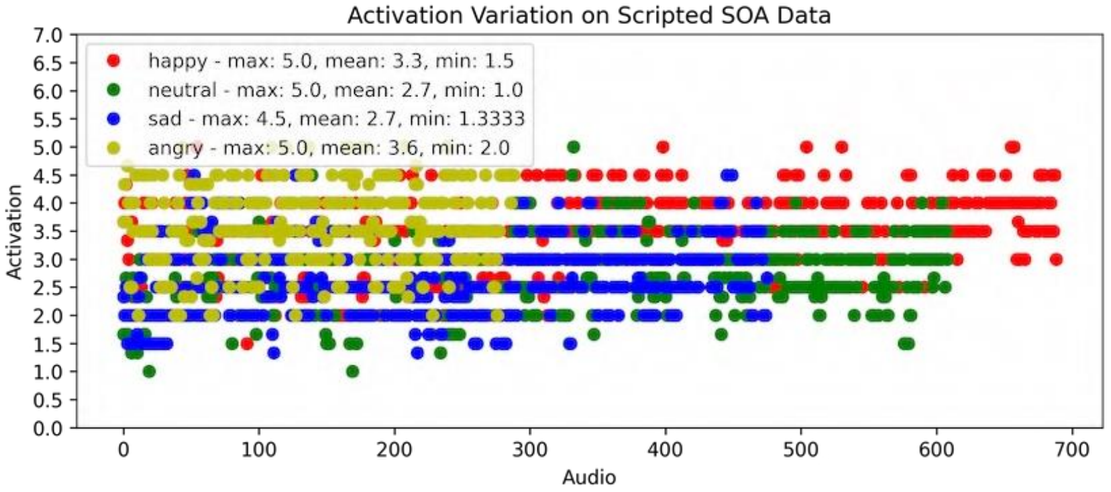
12 hours of audiovisual data (5 female and 5 male actors)

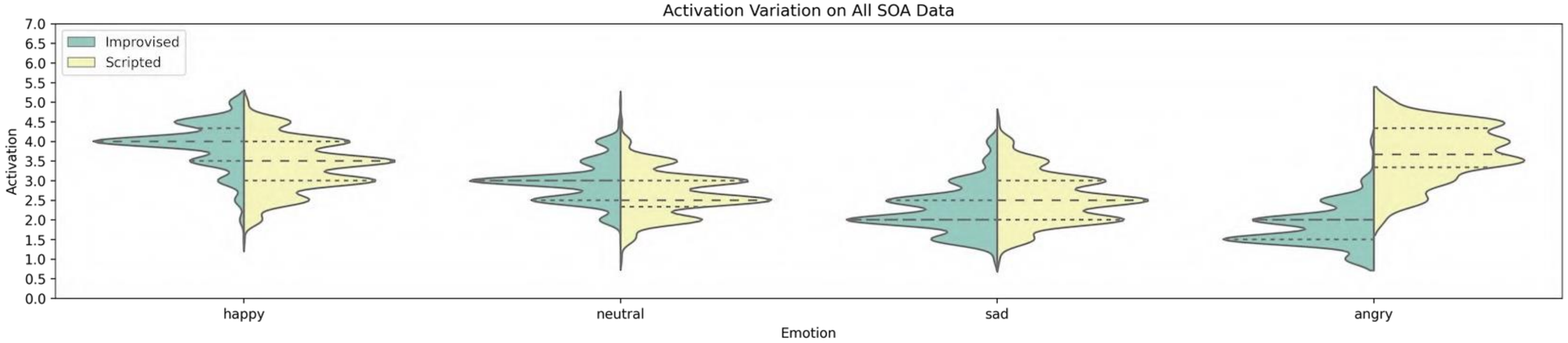
31 GB of data

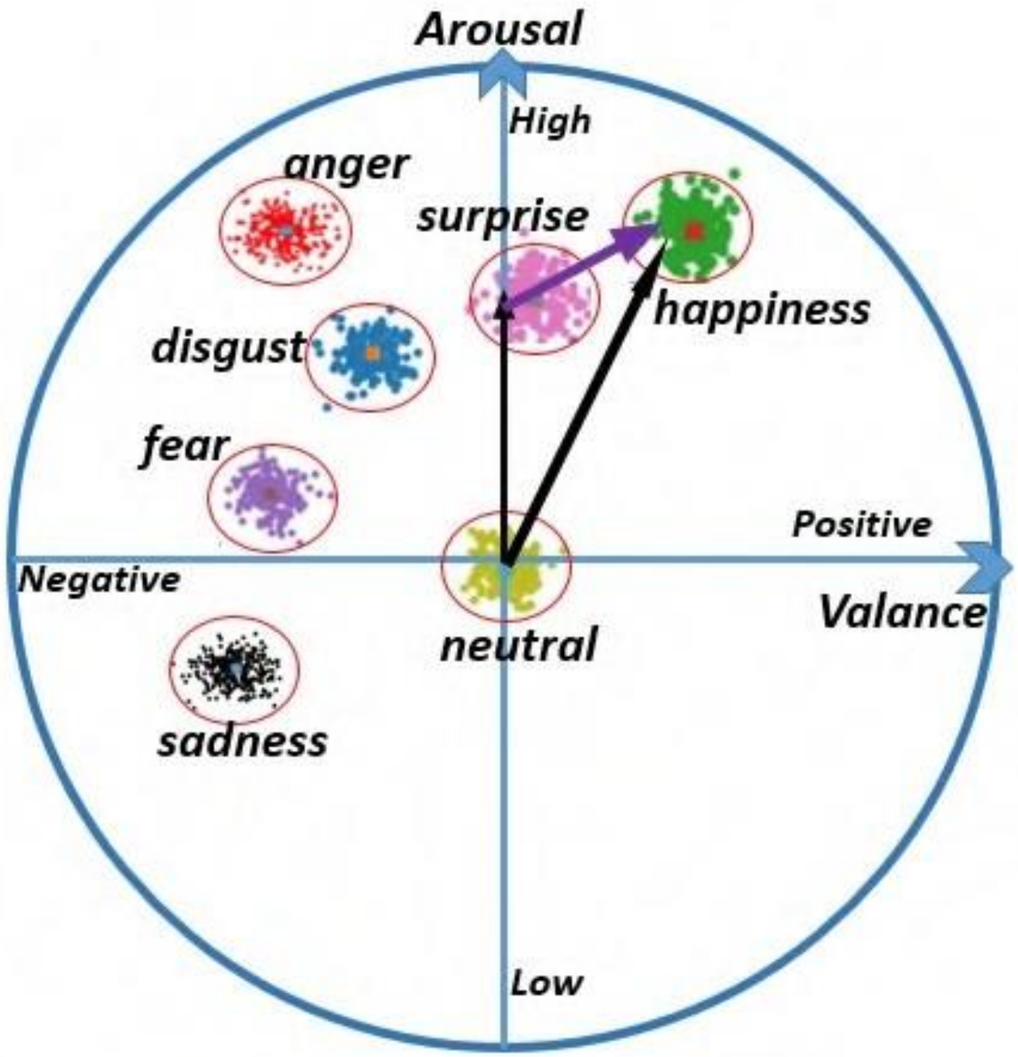
Used on **models' research, evaluation, and selection** due to being larger, making the processes more unbiased

Data and Bias Analysis









Emotion

The concept of emotion should be carefully defined.

There is still no consensus of an accurate representation due to its natural subjectivity.

Although, both the **discrete** and **dimensional** models are commonly accepted and utilized for the sentiment analysis task, it was decided to focus on the **discrete** (categorical) approach first.

Audio Features Explored

Speech is a continuous signal of varying length that carries both information and emotion.

Prosodic Features

Can be perceived by humans, such as intonation, rhythm, and loudness.

- Chromogram (Pitch)
- Root-mean-square (Energy)
- Zero crossing rate (Energy)

Spectral Features

Obtained by transforming the time domain signal into the frequency domain.

- Mel-frequency cepstral coefficients
- Mel-scaled spectrogram
- Spectral centroid
- Spectral bandwidth
- Spectral contrast
- Roll-off frequency

Metrics

For each one of the features, various metrics were applied.

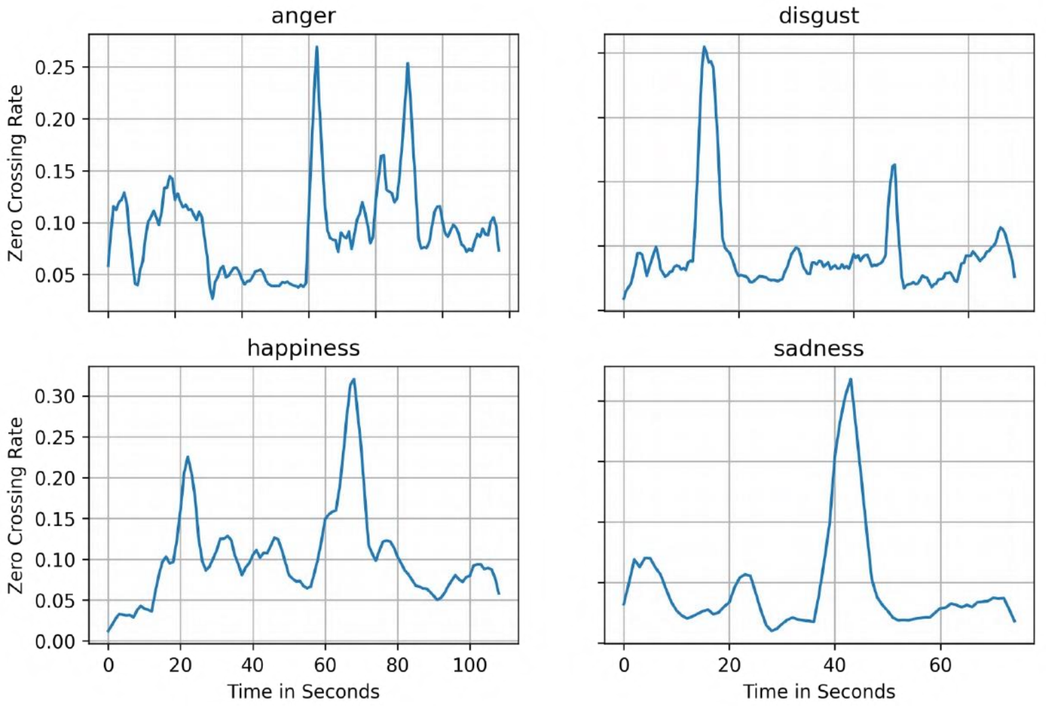
- Mean
- Variance
- Standard deviation
- Sum
- Minimum
- Maximum
- Spikes (explained later)

Feature Analysis

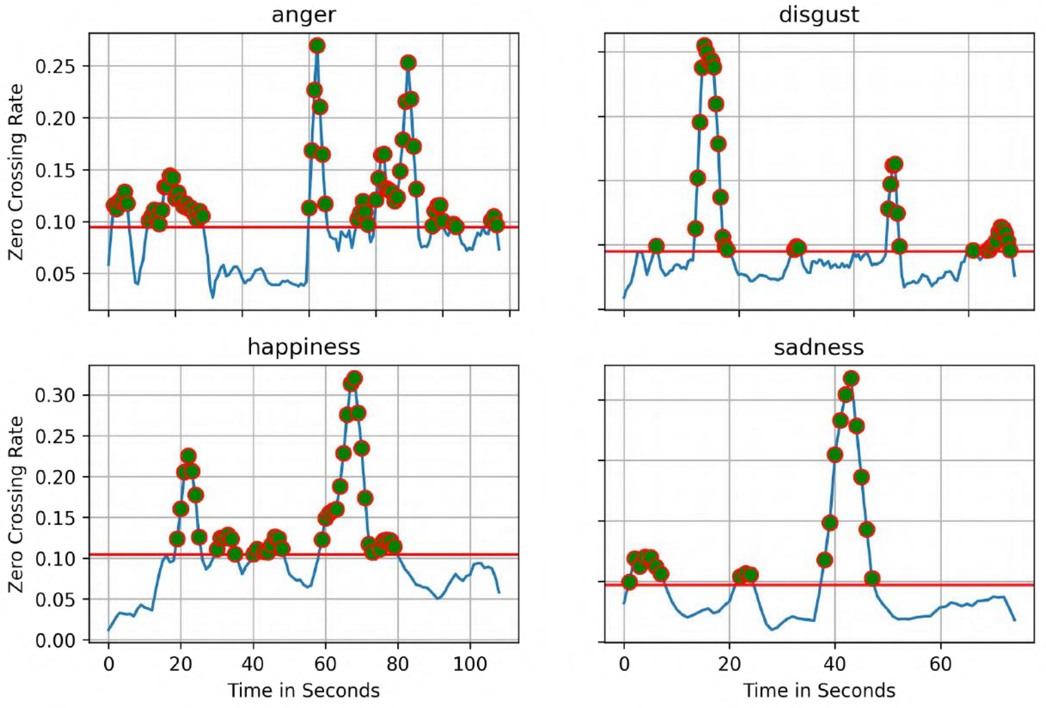
The features were analyzed and interpreted through graphical visual observations.

Initially, some wave plots were observed, and, noted consistency in the number of high values. For this reason, it was created a custom metric that calculates those high values, which were called “spikes”.

- **WAVE PLOTS**



Zero crossing rate wave plot



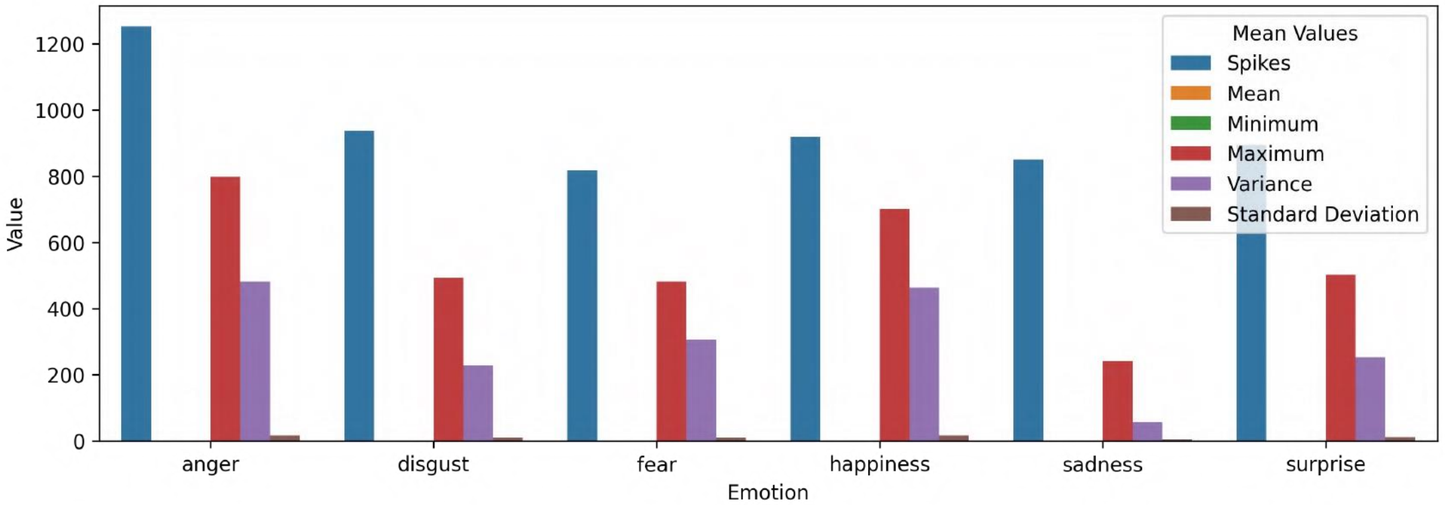
Zero crossing rate wave plot annotated with spikes

Feature Analysis(Continuation)

Bar plots were useful for viewing the overall extracted features plainly, and to understand their numeric values .

- **BAR PLOTS**

Bar plots mean for metrics used on the mel-scaled spectrogram feature

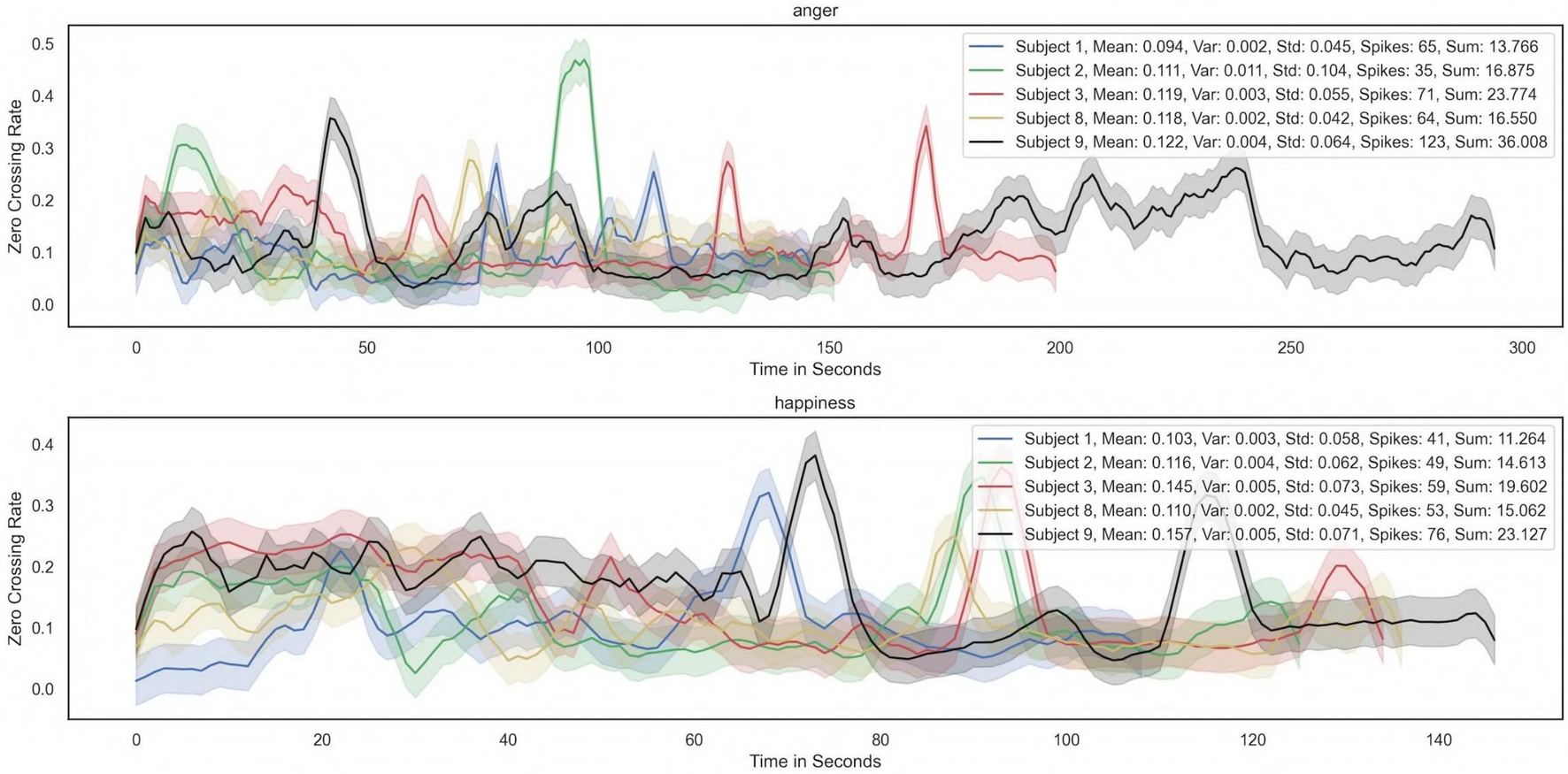


Bar plots mean of some spectral features

Feature Analysis(Continuation)

If the surrounding areas of a feature on a given emotion for different subjects are relatively overlapping, it could be an indicator that the feature is relevant for representing that emotion.

- **WAVE PLOTS WITH SURROUNDING AREAS**

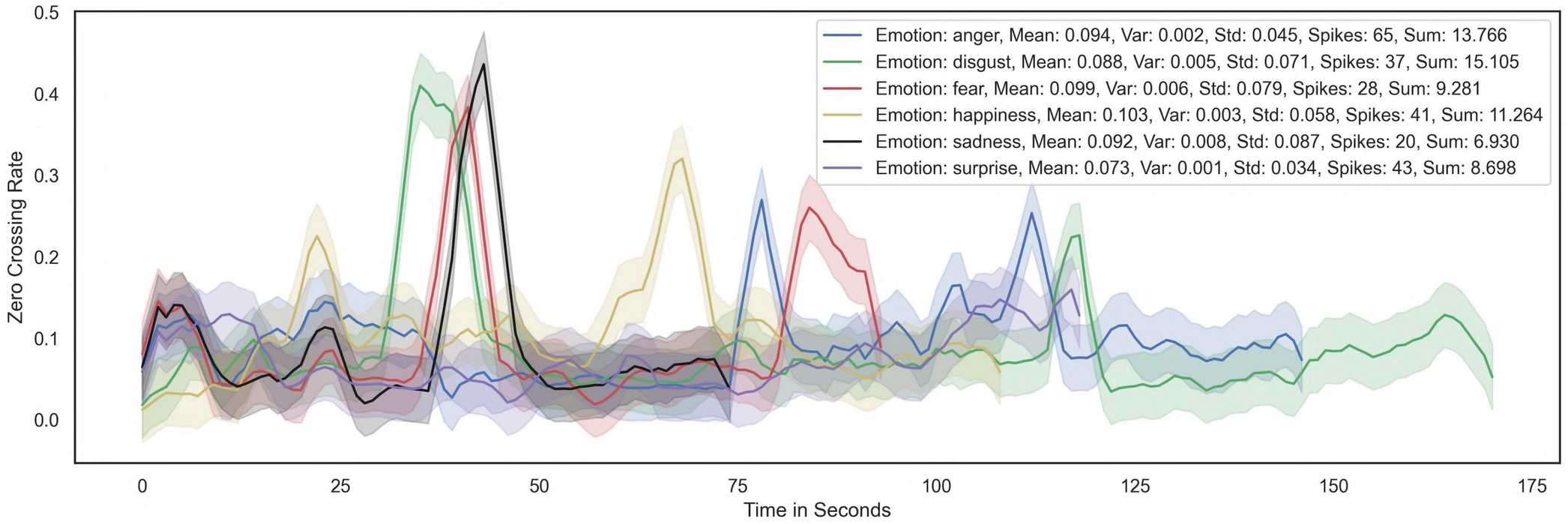


Zero crossing rate wave plots with a surrounding area of kve male subjects for anger and happiness

Feature Analysis(Continuation)

The previous idea can be “reshaped” to conceive if the feature is favorable to disguish different emotions.

- **WAVE PLOTS WITH SURROUNDING AREAS**

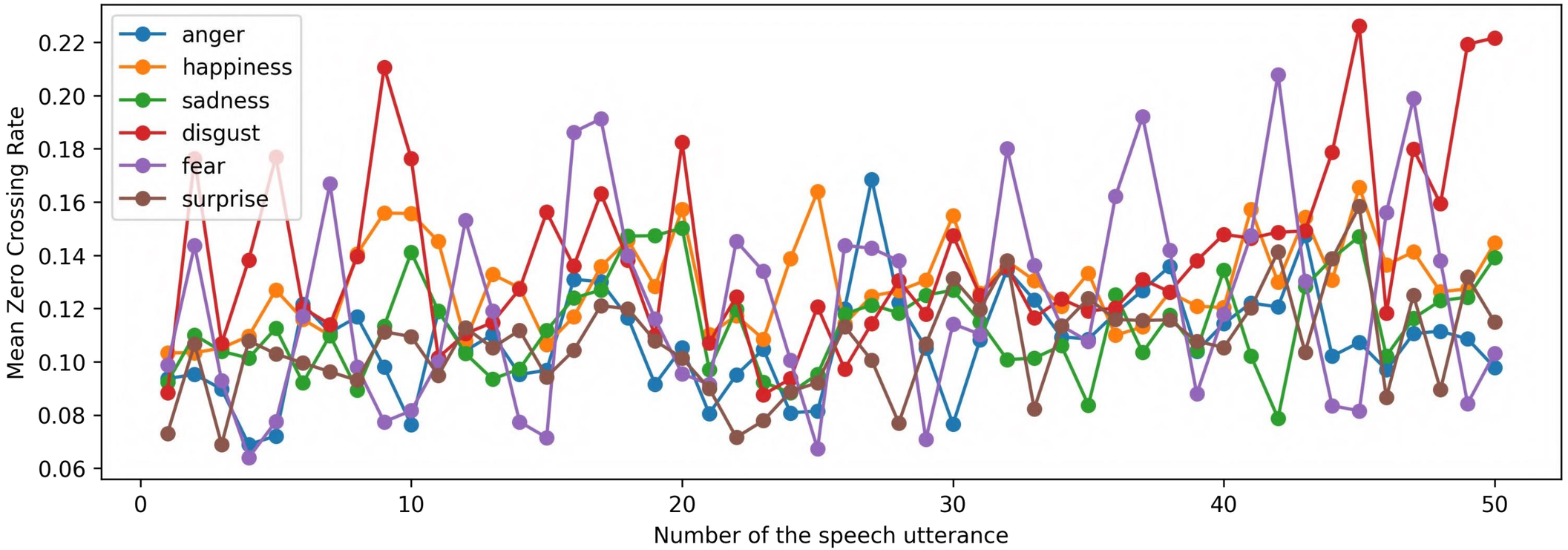


Zero crossing rate wave plots with a surrounding area of a male subject and the same sentence for all emotions

Feature Analysis(Continuation)

Variation plots help to perceive the differences of the features' values.

- **VARIATION PLOTS**

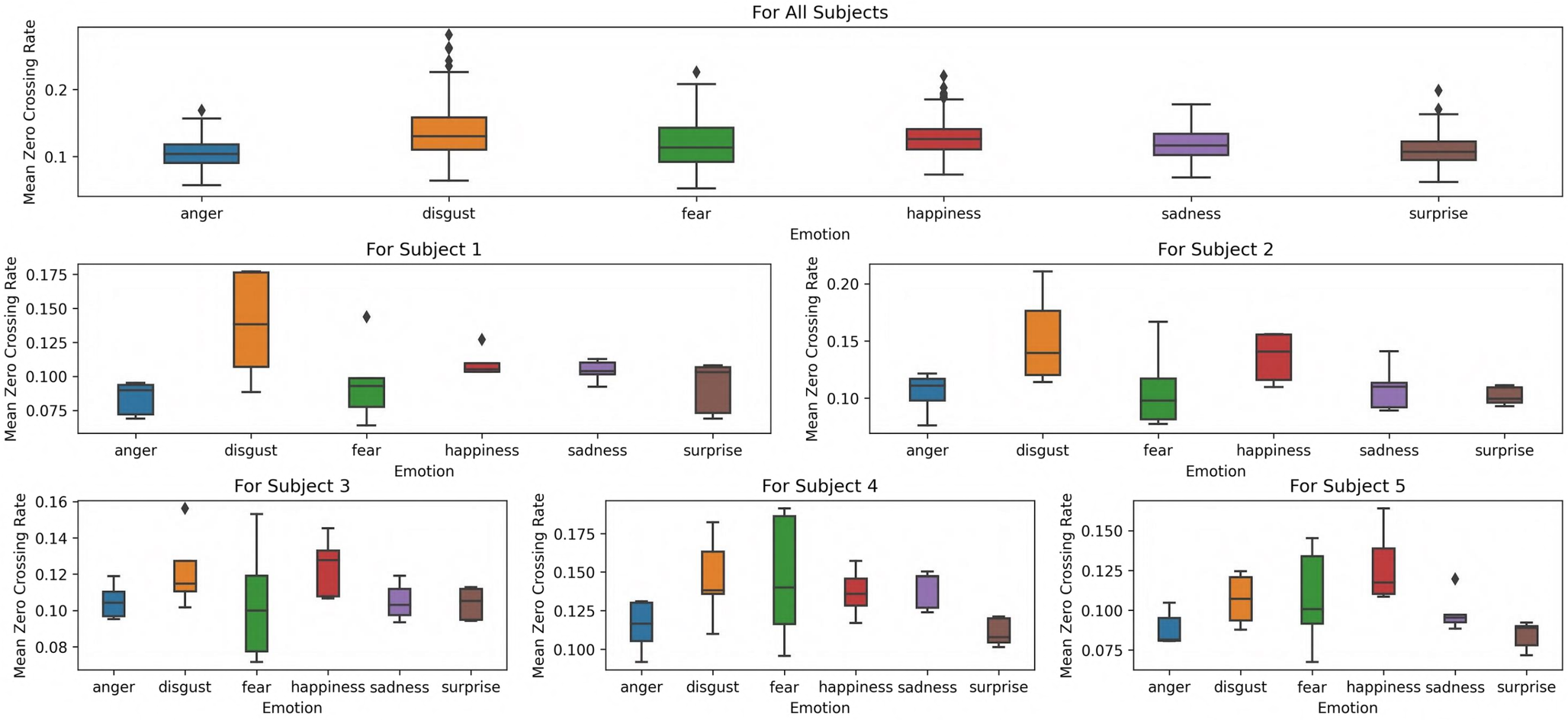


Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions

Feature Analysis(Continuation)

Box plots were used to get a simple visualization of each feature, on all subjects, and also, on a specific subjects, to seek for noticeable differences.

- **BOX PLOTS**

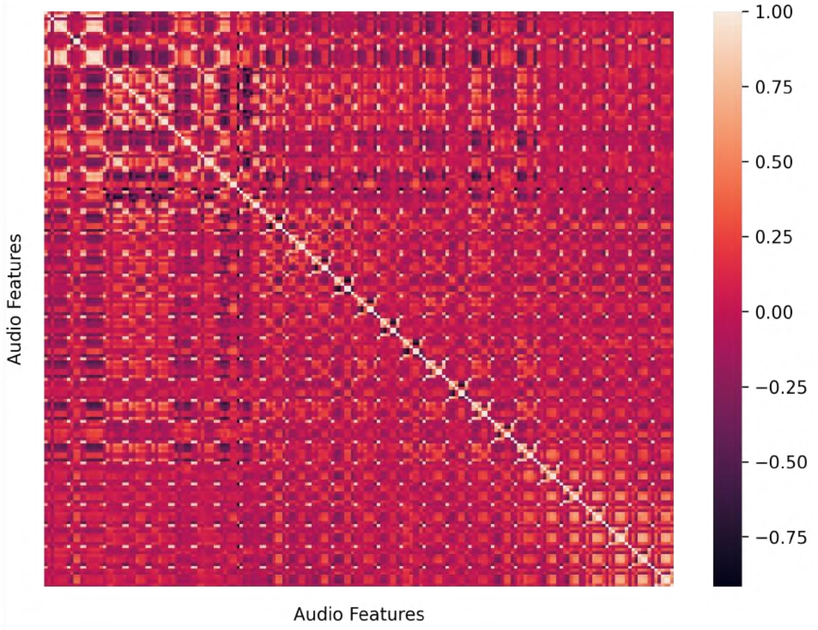


Zero crossing rate mean values box plot for all emotions and different subjects

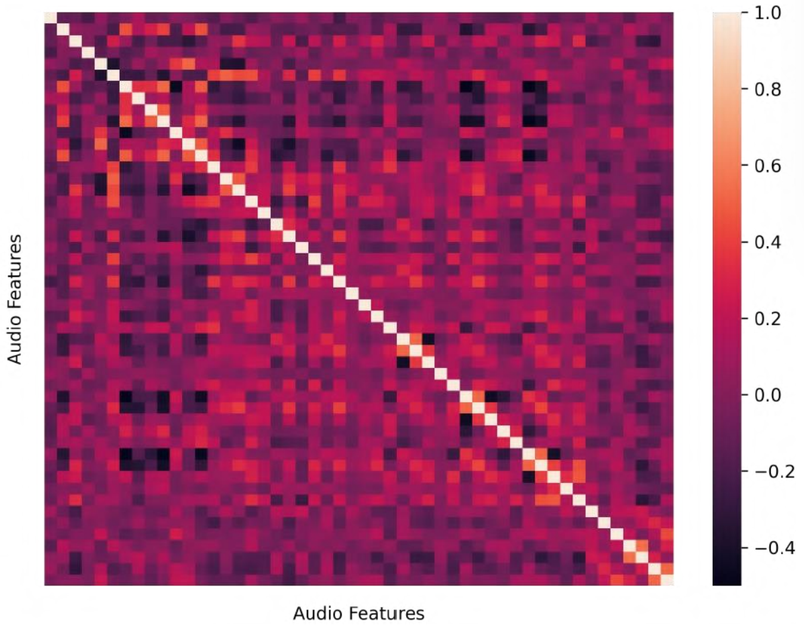
Feature Selection

Selecting the relevant input variables, not only reduces the computational cost but can also, improve the performance of the predictive models.

On this process, every pair of features with a **correlation higher than 0.6 (absolute)** is selected, and on each one of the pairs, the feature with the **highest average correlation value** between the two, is **eliminated**.



Correlation matrix of all features (194)



Correlation matrix of the lowest correlated features (50)

High Correlation Elimination

- **Eliminated 144 features** from the initial **194**, therefore remained **50 features.**

Feature Selection(Continuation)

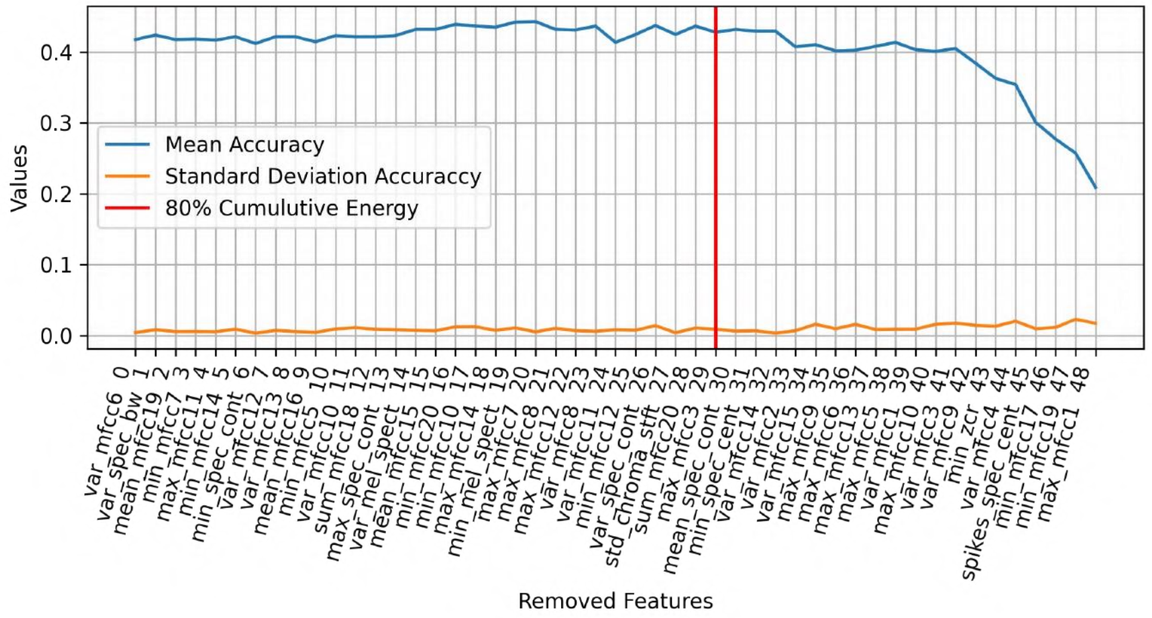
This method decides to remove a feature to remove based on evaluation scores of an estimator, in this case, a **5 k-fold cross-validation of a Random Forest**.

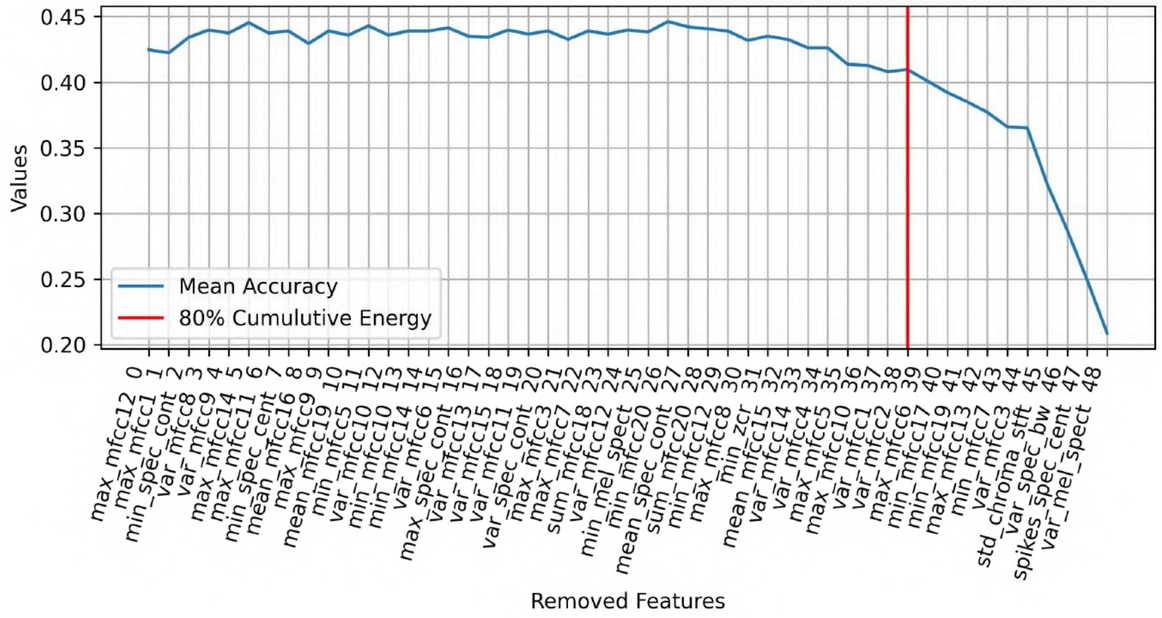
For deciding the feature to eliminate at each stage, two methods were used, one uses only the **mean accuracy**, and the other both the **mean and standard deviation**, each with **50% weight**.

The features that contain **80%** of the total **"energy"**, and that are **common to both methods**, were **eliminated**.

Feature Elimination with Backwards Propagation - Keeping 80% "Energy"

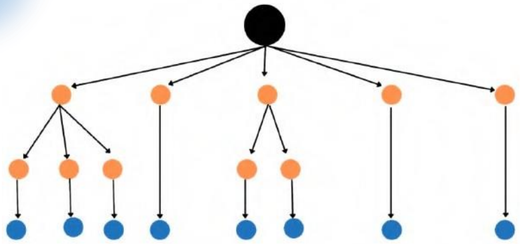
- Eliminated 26 features
- Reached the final **24 features**



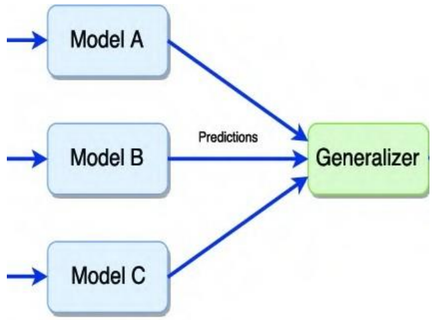


Classification Models

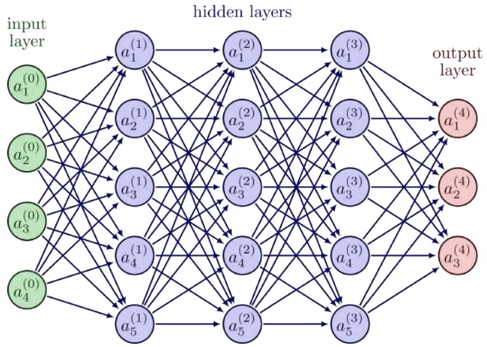
Approaches



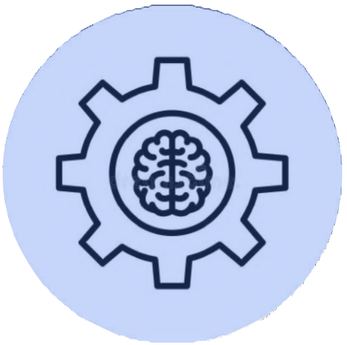
Decision Trees



Ensemble Methods

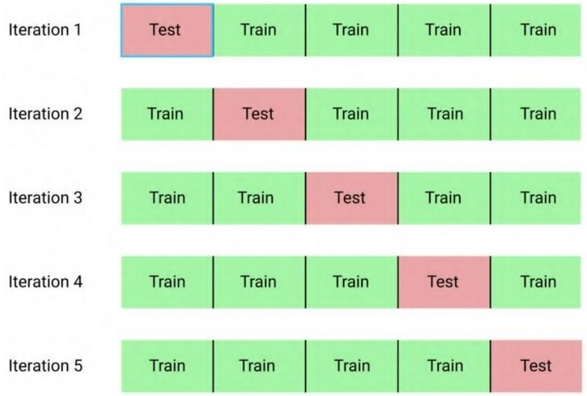


Neural Networks



Auto Machine Learning

Evaluation Strategy



5 Stratified Fold Cross-Validation

Accuracy
Macro F1
Precision
Recall
MCC

Metrics

Classification Models Performance on IEMOCAP

Model	Input	Accuracy	Macro F1	Precision	Recall
Random Forest	24 Audio Features	60.37	61.14	61.98	60.73
Ada Boost	24 Audio Features	60.15	60.87	61.90	60.42
Linear Discriminant Analysis	24 Audio Features	55.88	56.83	57.00	56.87
Histogram Gradient Boosting	24 Audio Features	60.17	60.91	61.40	60.66
Auto-SKLearn Ensemble	24 Audio Features	60.35	61.20	61.74	61.14
Simple CNN	24 Audio Features	45.62	46.22	48.56	46.13
AutoKeras NN	24 Audio Features	42.80	42.06	47.83	40.83
AutoKeras NN	Mel Spectrogram	46.81	46.80	47.24	46.96
AutoKeras NN	24 Audio Features + Mel Spectogram	50.50	50.85	50.78	51.92

Tested Models Results

Model	Input	Accuracy
Dilated Residual Network [Li et al. 2019]	Audio Features	67.4
RNN w/ Attention [Lu et al. 2020]	Audio + Text Features	72.6
Deep CNN [Issa et al. 2020]	193 Audio Features	64.30
CNN + LSTM [Chen et al. 2018]	Spectrogram	64.70
CNN [Kwon 2019]	Processed Spectrogram	81.75

State-of-Art Results

Areas to Explore

Hyperparameter Tuning

Audio Pre-Processing

Transfer Learning with Pre-Trained Models

Audio Spectrogram as Input for Deep Learning