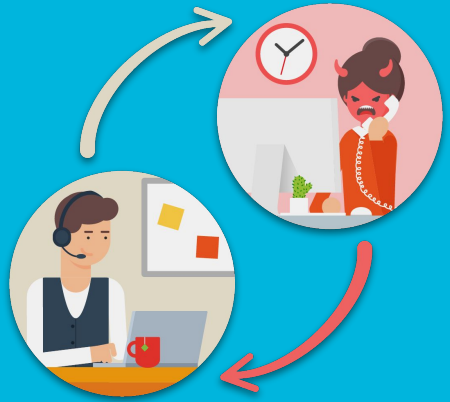
Models development for audio evaluation in affective computing

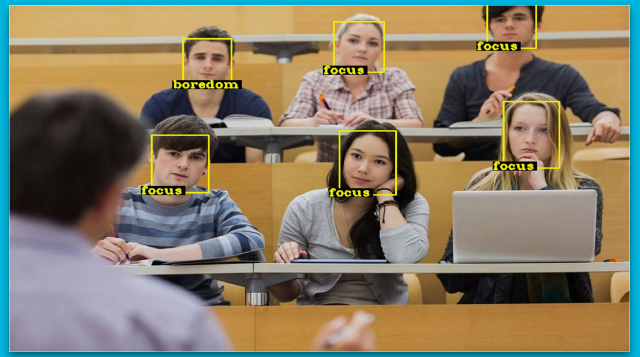
29/06/2023

Mário Francisco Costa Silva
Master in Informatics Engineering
Department of Electronics, Telecommunications and Informatics

Just like we can understand speech and machines can communicate in speech, we also understand and communicate with humor and other kinds of emotions. And machines that can speak the language of emotions are going to have better, more effective interactions with us.

- MIT Sloan professor Erik Brynjolfsson





Objectives

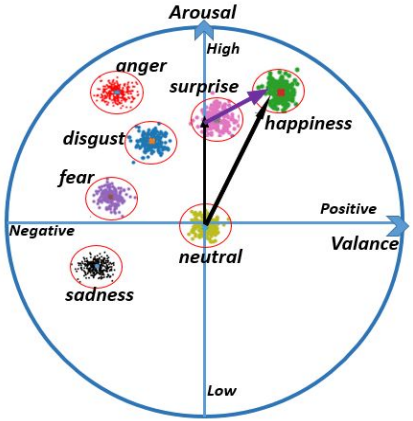
- Develop a **Speech Emotion Recognition (SER)** system.
- Develop a unique set of audio features and propose different **SER** models.
- Provide an offline and in real-time emotional evaluation.

Speech Emotion Recognition State-Of-The-Art

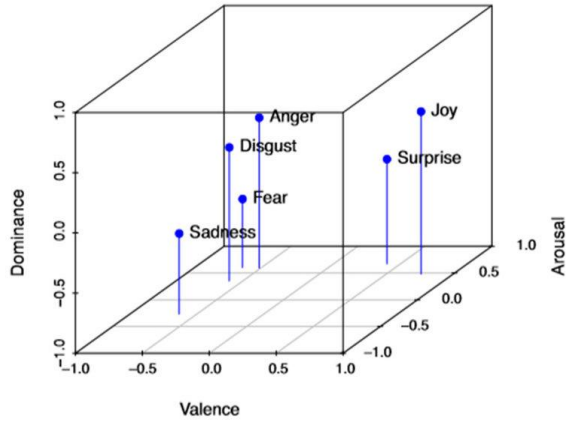
Emotion

- There is still no consensus of an accurate representation of the emotion concept due to its **natural subjectivity;**
- However, the most accepted and utilized are the **discrete** and **dimensional** models.

Discrete emotions in a 2-D valence-arousal plane.



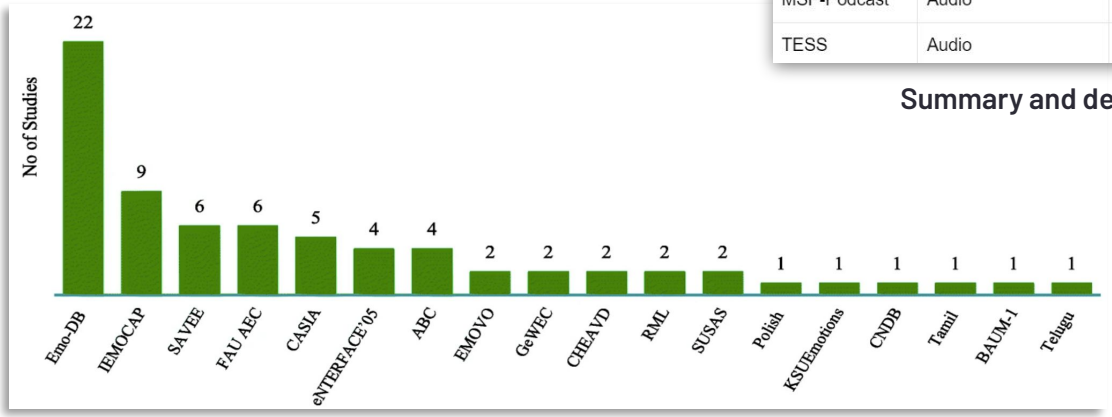
Discrete emotions in a 3-D valence-arousal-dominance plane.



Emotion Recognition Datasets

Dataset	Format	Language	Content	Labels	Type
EMO-DB	Audio	German	800 recordings by 10 actors	7 discrete	Acted
eNTERFACE'05	Audio and Video	English	42 subjects 14 with nationalities	6 discrete	Elicited
IEMOCAP	Audio, Video and Text	English	12 hours by 10 speakers	10 discrete and VAD	Acted
MOUD	Audio and Video	Spanish	80 product reviews videos	Polarity	Natural
CREMA-D	Audio,Video and Text	English	7442 clips by 91 actors	6 discrete	Acted
CMU-MOSI	Audio and Video	English	2199 movie reviews	Polarity (Likert scale)	Natural
MSP-Improv	Audio, Video and Text	English	8438 recordings by 12 actors	4 discrete	Elicited
MOSEI	Audio, Video and Text	English	65 hours of YouTube videos	6 discrete and polarity	Natural
MELD	Audio, Video and Text	English	1400 dialogues from Friends TV	7 discrete and polarity	Acted
RAVDESS	Audio and Video	English	7356 recordings by 24 actors	7 discrete	Acted
MSP-Podcast	Audio	English	100 hours by over 100 speakers	9 discrete and VAD	Natural
TESS	Audio	English	2800 recordings by 2 actresses	7 discrete emotions	Acted

Summary and description of emotion recognition datasets.



Databases used in articles found in SER survey (frequency distribution)

Speech Features

Speech is a vital tool for human communication and social interaction. It is a continuous audio signal that can convey information, emotions and meaning.

Prosodic Features

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

- Energy
- Pitch
- Duration

Spectral Features

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

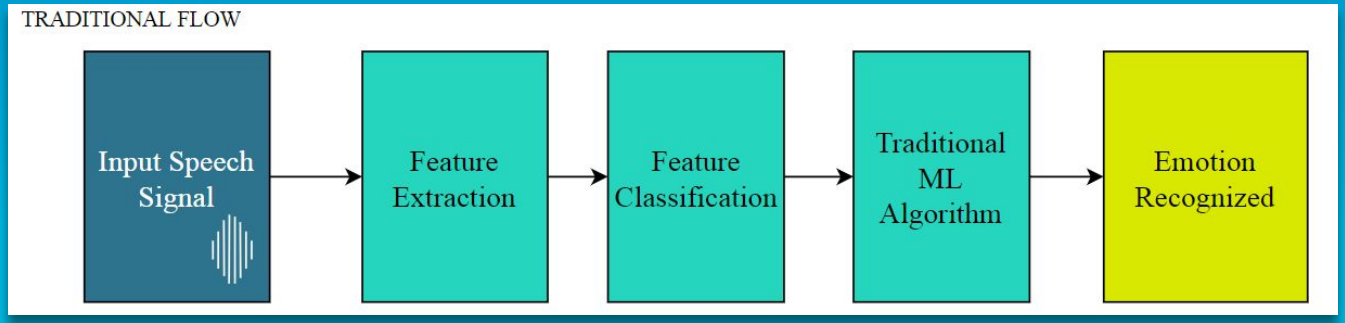
- Mel-frequency cepstral coefficients (MFCC)
- Mel spectrogram
- Spectral centroid
- Spectral bandwidth
- Spectral contrast
- Roll-off frequency

Voice Quality Features

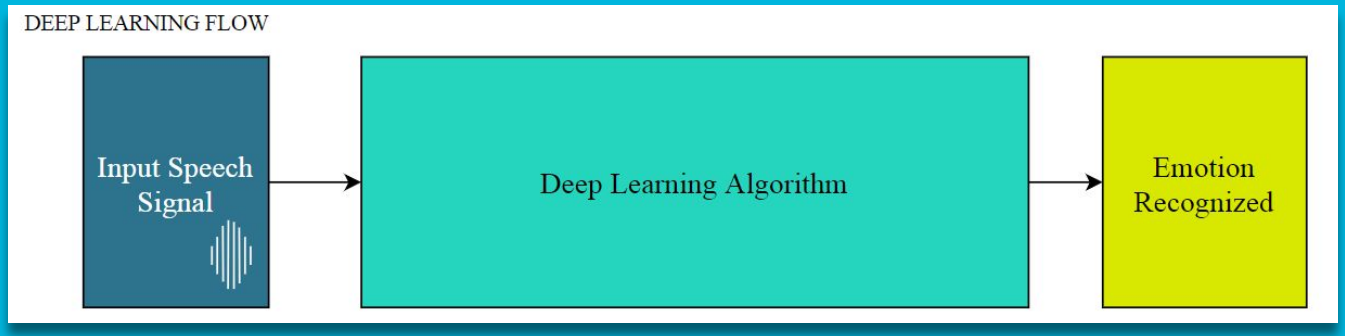
Distinctive characteristics of a person's voice and measure of the overall voice quality.

- Jitter
- Shimmer
- Harmonics-to-noise ratio

SER Strategies



Traditional Feature-Based SER Flow



Deep Learning-Based SER Flow

SER SOTA Articles

Audio classification articles with traditional feature-based approaches

Paper Title (Publication Date)	Audio Features	Classifier	Datasets (№ of Labels) & Accuracies (%)
Spoken emotion recognition using hierarchical classifiers (Jul. 2011)	MFCC Log-spectrum Pitch	HMM	EMO-DB (7): 71.75
Emotion recognition using a hierarchical binary decision tree approach (Nov. 2011)	Zero crossing rate Root-mean-square energy Pitch MFCC Harmonics-to-noise ratio	Multi-level binary decision trees	(UARs) AIBO (5): 41.57 IEMOCAP (4): 58.46
Multimodal SER and Ambiguity Resolution (2019)	Pitch Harmonics Pause	Random forest, extreme gradient boosting and multilayer perceptron	IEMOCAP (4): 56.6
SER with deep CNN (May 2020)	MFCC Chromagram Mel spectrogram Spectral contrast	One-dimensional CNN	RAVDESS (8): 71.61 IEMOCAP (4): 64.30 EMO-DB (7): 86.10

SER classification articles with deep learning-based approaches

Paper Title (Publication Date)	Audio Features	Classifier	Datasets (№ of Labels) & Accuracies (%)
Sentiment analysis in non-fixed length audios using a Fully CNN (2021)	Mel-Spectrogram MFCC	Fully CNN	RAVDESS (8): 75.28 EMO-DB (7): 92.71 TESS: 99.03
Emotion Recognition from Variable-Length Speech Segments Using DL on Spectrograms 2018	Log-Spectrograms	CNN and RNN	IEMOCAP (4): 64.22
Exploring Deep Spectrum Representations via Attention-Based RNN and CNN for SER (2019)	Mel-Spectrogram	Attention-based bidirectional LSTM and fully CNN	IEMOCAP (4): 67.0
Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network (2018)	MFCC Spectral centroid Spectral contrast Chromagram	LSTM and CNN	MOSI (4): 68.74 MOUD (2): 69.64
3-D Convolutional RNN With Attention Model for SER (2018)	Mel spectrogram	3-D attention-based convolutional RNN	(UARs) IEMOCAP (4): 64.74 EMO-DB (7): 82.82
SER Using Quaternion CNN (Jun. 2021)	Mel spectrogram encoded in an RGB quaternion domain	Quaternion CNN	RAVDESS (8): 77.87 IEMOCAP (4): 70.46 EMO-DB (7): 88.78
Rethinking CNN Models for Audio Classification (2020)	Mel spectrogram	Ensemble of ImageNet pre-trained DenseNets	ESC-50 (50): 92.89 UrbanSound8K (10): 87.42
SER Using Deep CNN and Discriminant Temporal Pyramid Matching (2018)	Mel spectrogram	Deep CNN (Fine-tuned the AlexNet pre-trained on ImageNet)	EMO-DB (7): 87.31 RML (6): 69.70 eNTERFACE05 (6): 76.56 BAUM-1s (6): 44.61

Key Takeaways

I. Emotion recognition systems face challenges due to the **subjective nature of emotions**, lack of consensus on labeling, and the need for diverse emotional data.

II. **Audio feature engineering** is crucial for the traditional feature-based SER, requiring **focus on feature selection** of a wide range of features, including **prosodic, spectral, and voice quality**.

III. **Deep Learning models**, such as CNN, RNN, and LSTM, have **outperformed** the **traditional approach**, using **Mel spectrogram** as input.

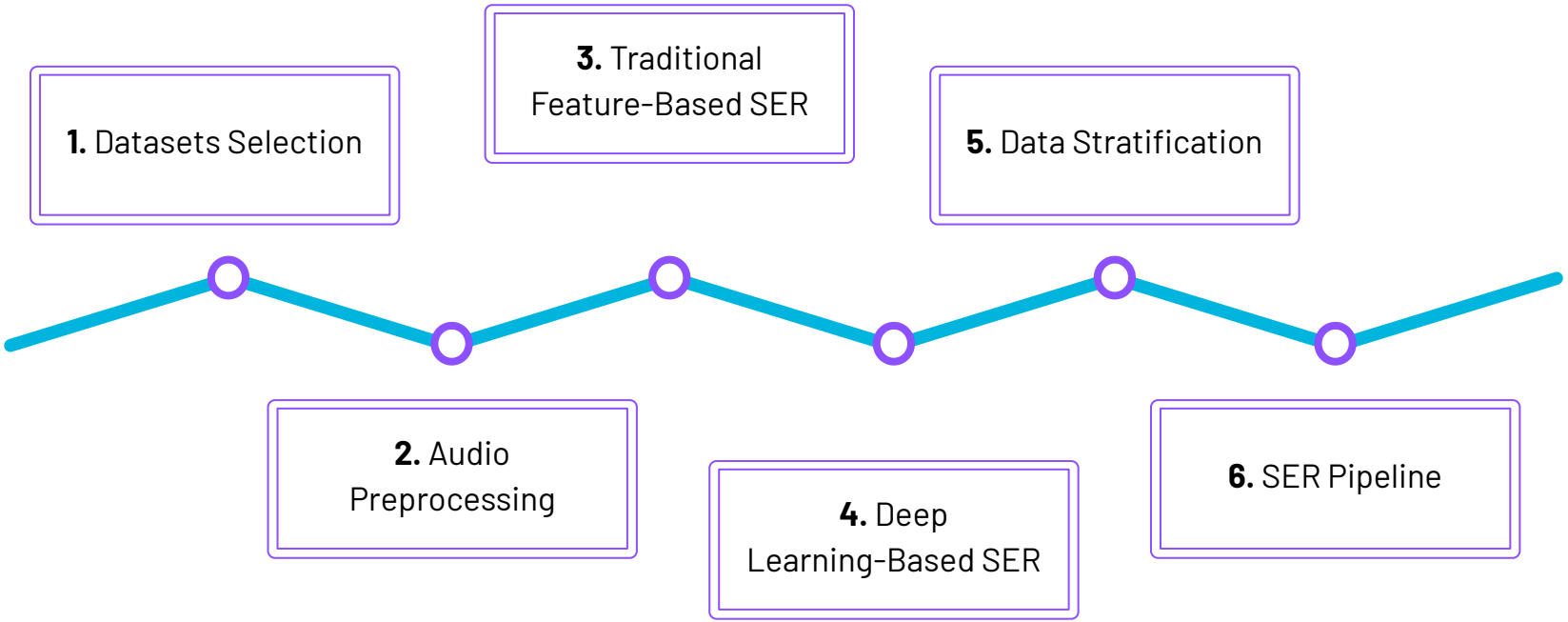
IV. Affective computing has shifted towards leveraging **multimodal** information, beyond facial cues, leading to improved emotion recognition accuracy in current market solutions.

Ethical Considerations

CONCERN	PROCEDURES
System may not be accurate in recognizing emotions for people with certain characteristics (age, gender, accents, speech disabilities, races, etc.).	It was **addressed and quantified potential biases** present in our proposed models.
System may be used to monitor or control participants without their consent.	Our models require **participants consent** to submit their data for emotional evaluation purposes.
Personal data collected being vulnerable to hacks or used for unintended purposes.	The SER pipelines only utilize the participant's streaming data for the models input. They **do not store or use sensitive data** for any other purposes than emotional analysis.
Revealing or identifying confidential data such as personal, financial, or business information.	The models **do not include speech transcriptions**, and therefore, do not capture the actual words spoken.

Speech Emotion Recognition Development

SER Development Methodology



1. Datasets Selection

The **IEMOCAP** dataset was utilized as the **main dataset**, which is for **feature engineering**, **models evaluation and selection**, and as **training data** for the final proposed models. For **cross-dataset validation** of these models, we selected three additional datasets: **eNTERFACE'05**, **CREMA-D**, and **EMO-DB**.

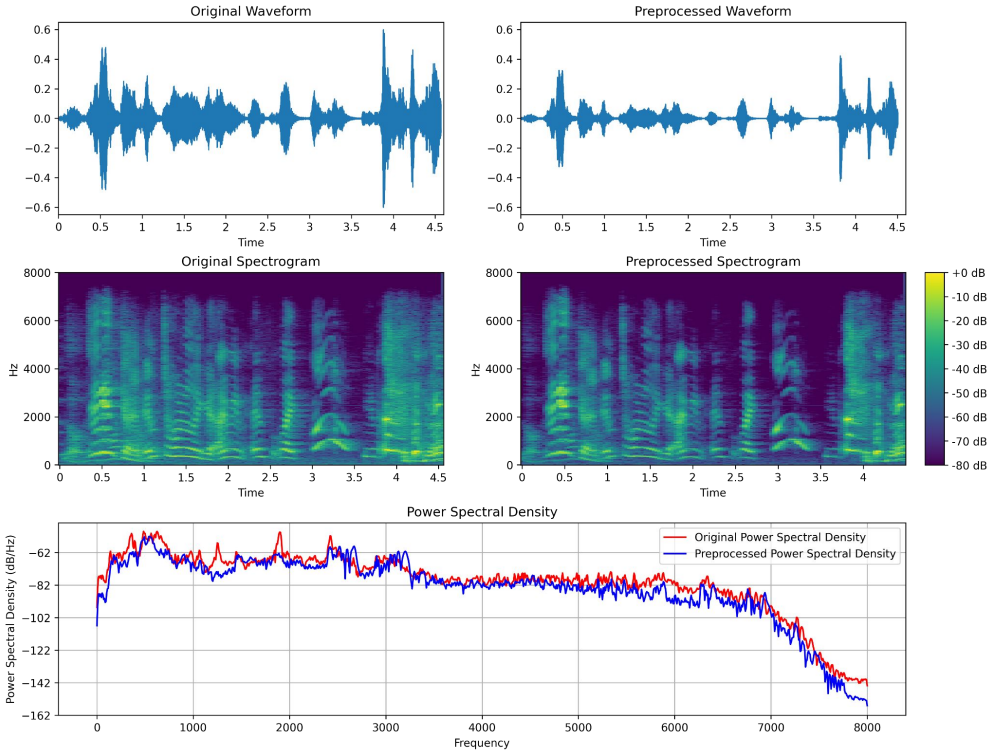
Number of audio files per emotion from the selected datasets.

	Number of Files			
Emotion	IEMOCAP	eNTERFACE'05	EMO-DB	CREMA-D
Anger	1103	210	127	1271
Happiness	1636	210	71	1271
Neutral	1708	0	79	1087
Sadness	1084	210	62	1269

2. Audio Preprocessing

- **Noise reduction**, using the spectral gating technique.

- **Audio trimming**, which removes silent frames at the beginning and end of an audio signal.



Graphics of audio features before and after preprocessing.

3. Traditional Feature-Based SER

3.1. Feature Extraction

Extracted audio features and the applied statistical functions.

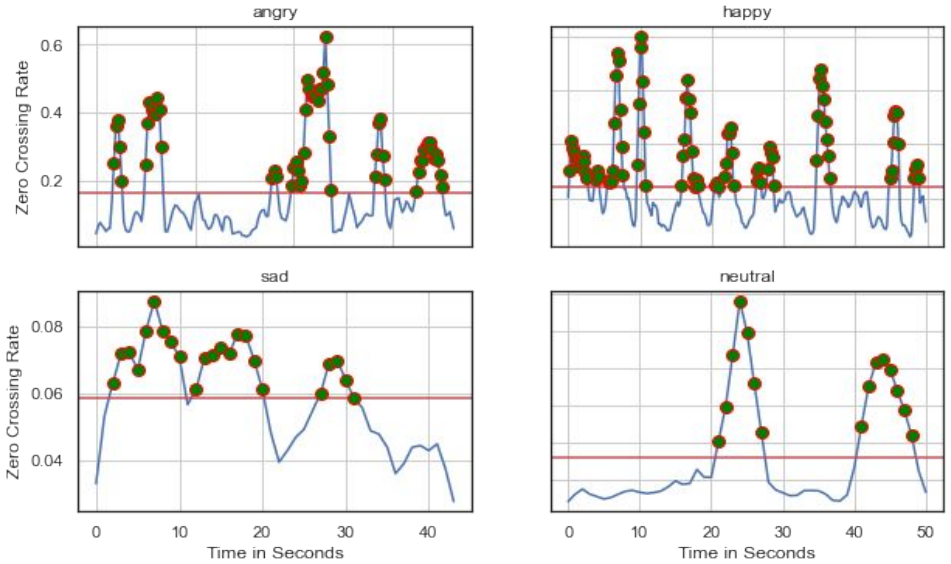
Audio Features	Statistical Functions
	Minimum
MFCCs 1 - 21	Mean
Mel Spectrogram	Maximum
Root-Mean-Square	Median
Chromagram	25th percentile
Spectral Centroid	75th percentile
Spectral Contrast	Spikes[1]
Spectral Bandwith	Variance
Roll-Off Frequency	Standard Deviation
Tonnetz	Sum
Zero-Crossing Rate	Kurtosis[2]
	Skew[2]

[1] Custom function.

[2] Only for the MFCCs.

3.2. Feature Analysis

Wave Plots and the “Spikes” Metric



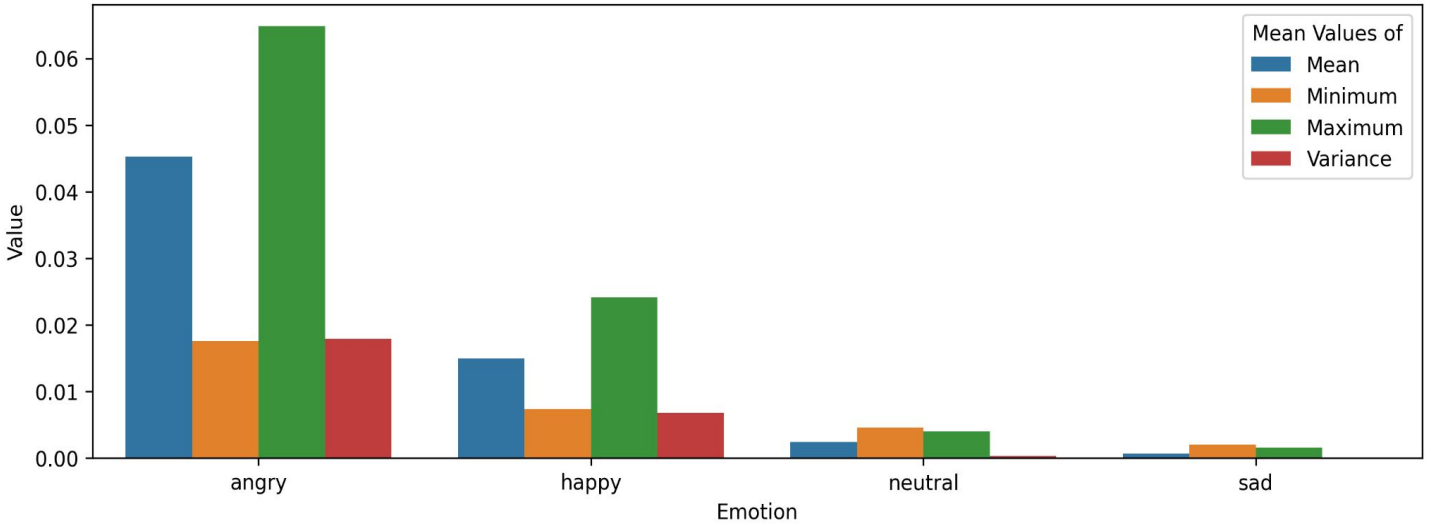
Zero crossing rate wave plot annotated with spikes.

Python code for calculating the spikes metric:

```
def spikes(data):
        mean = np.mean(data)
        std = np.std(data)
        threshold = mean + np.abs(std) * 2 / 100
        num_spikes = 0
        for value in data:
                if value >= threshold:
                        num_spikes += 1
        return num_spikes / len(data)
```

3.2. Feature Analysis

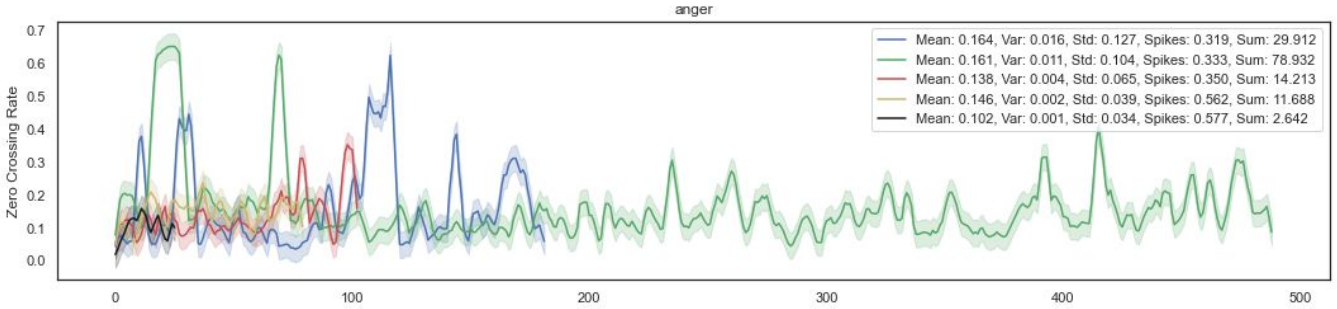
Bar Plots



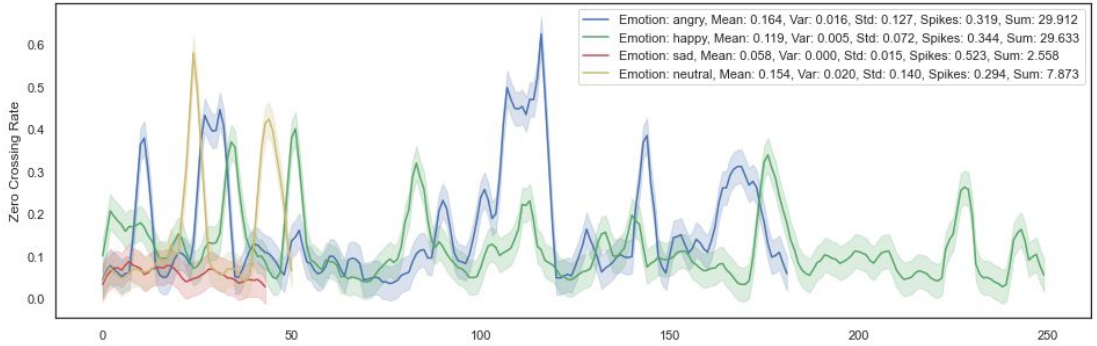
Bar plots mean for metrics used on the mel spectrogram feature.

3.2. Feature Analysis

Wave Plots with Surrounding Areas



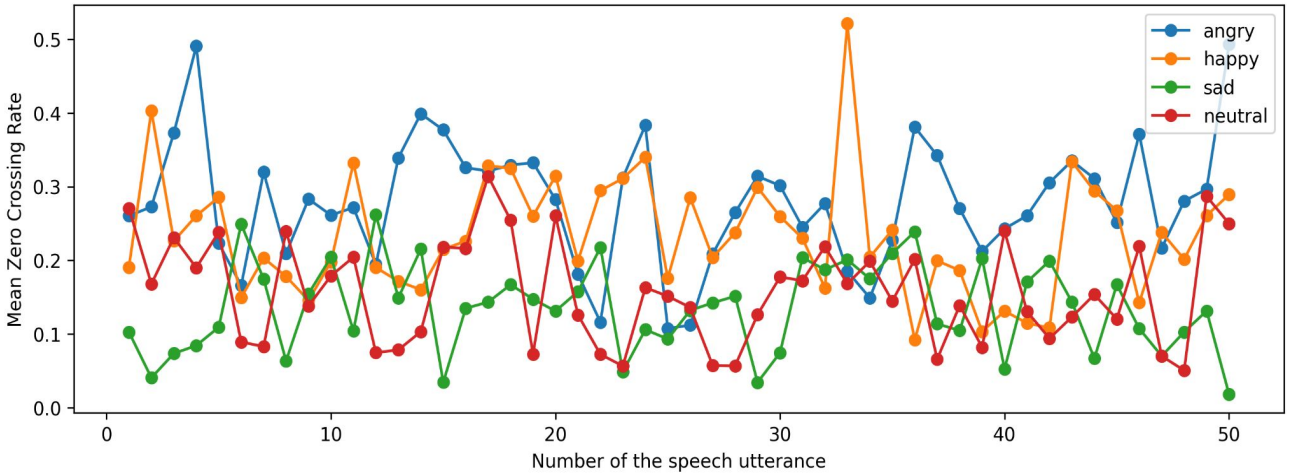
Zero crossing rate wave plots with a surrounding area of five male subjects for the same utterance with the anger emotion.



Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions.

3.2. Feature Analysis

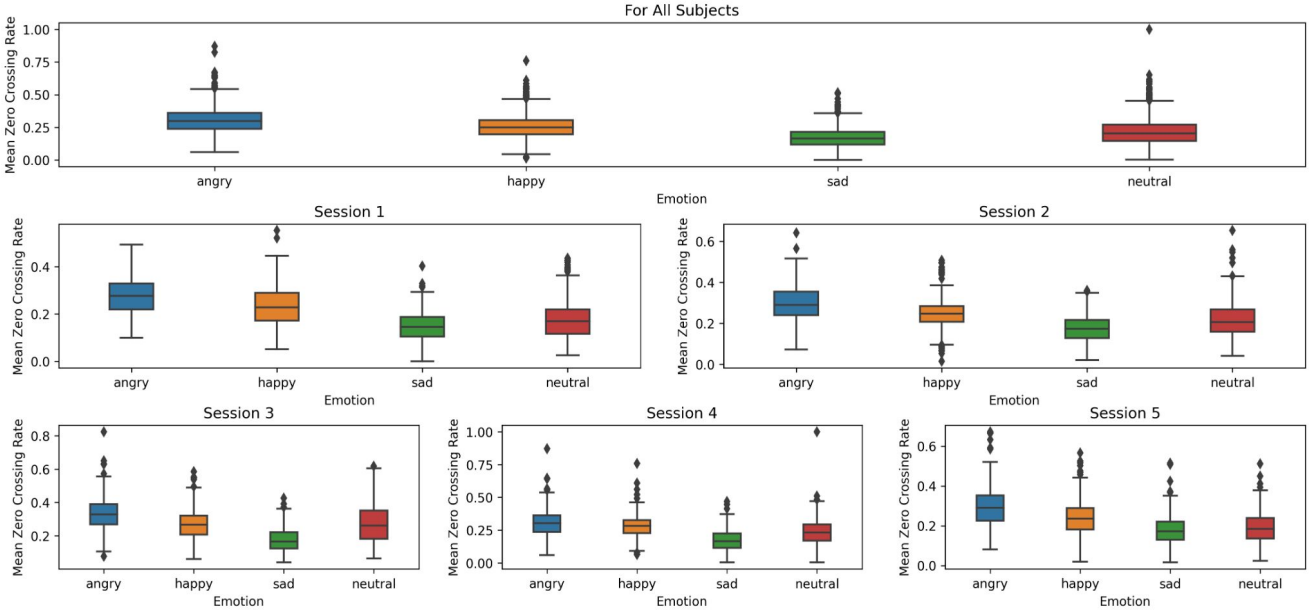
Variation Plots



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

3.2. Feature Analysis

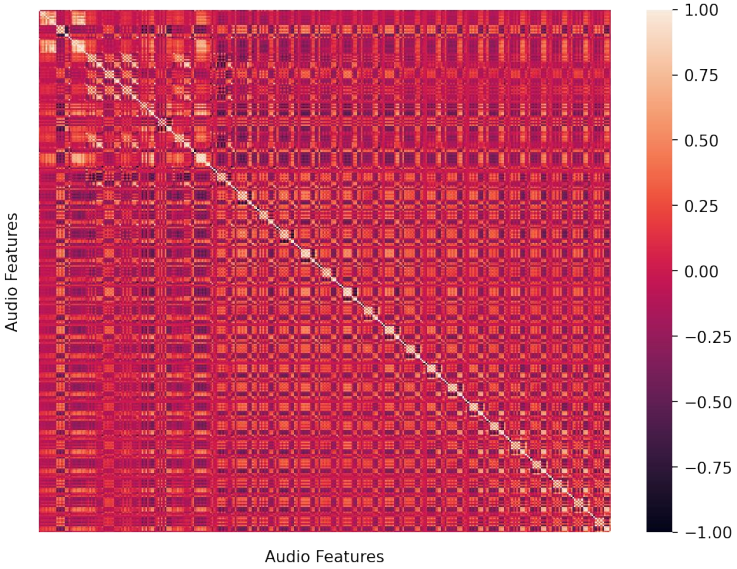
Box Plots



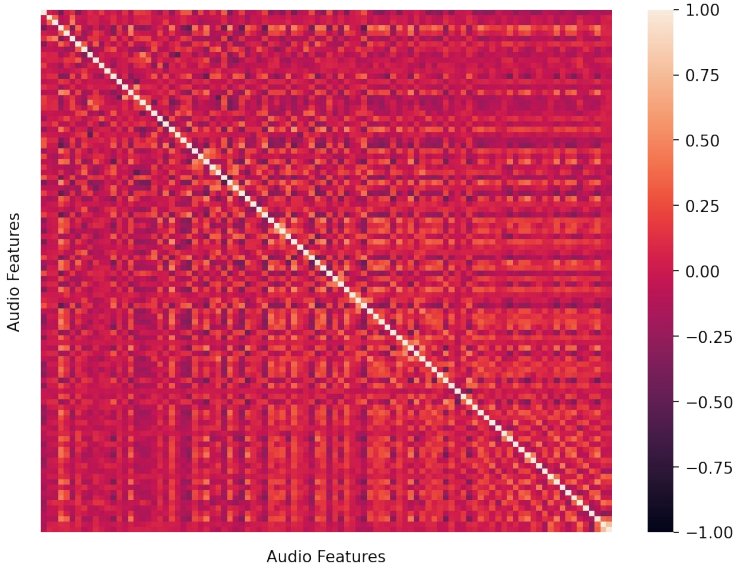
Zero crossing rate mean values box plot for all emotions and different subjects.

3.3. Feature Selection

Correlation-Based Feature Selection (CFS)



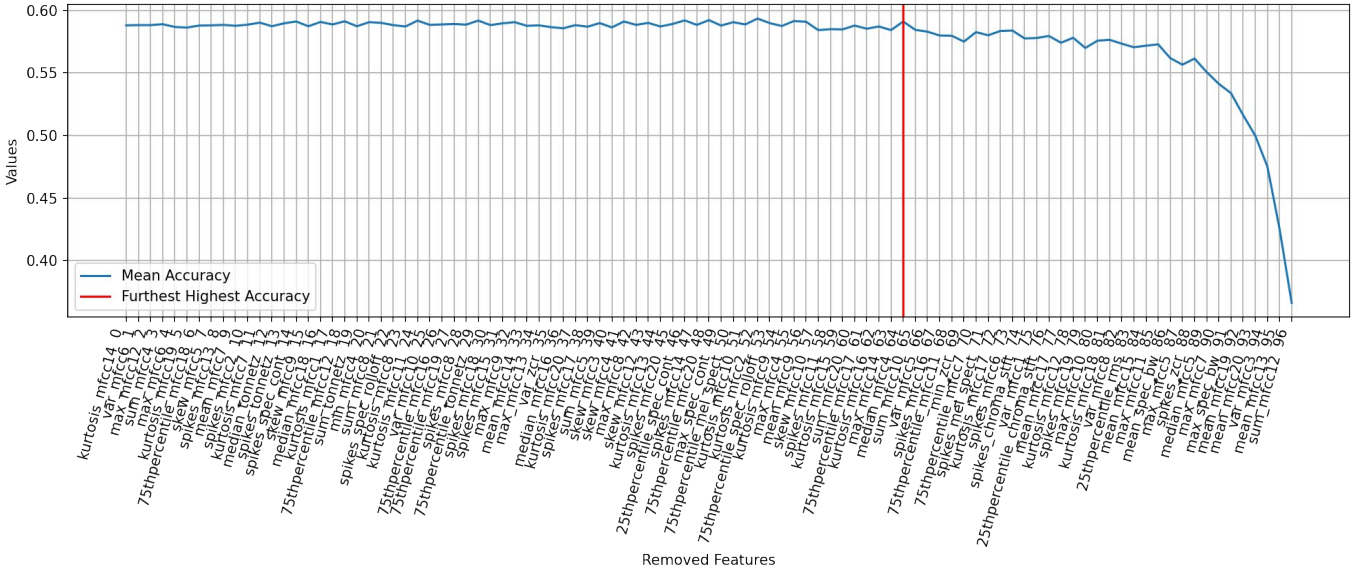
Correlation matrix of all the audio features.



Correlation matrix of the selected audio features after CFS.

3.3. Feature Selection

Backwards Selection



Backwards Selection using the mean accuracy as the selection criteria.

3.3. Feature Selection

Random Forest 5-fold cross-validation results using different sets of features.

Feature Selection Method	N.º of Features	Accuracy	Training Time (s)
None	327	59.14±0.68	15.34
CFS	98	57.87±1.07	7.89
CFS & Backward Selection	33	59.12±1.05	4.29

Final set of the 33 selected features.

Metric	Audio Features
Spikes	Mel-Spectrogram, Chromagram, Zero Crossing Rate, MFCC-6, MFCC-16, MFCC-19
Mean	Spectral Bandwidth, MFCC-13, MFCC-15, MFCC-17, MFCC-19, MFCC-20
Maximum	Spectral Bandwidth, MFCC-5, MFCC-7, MFCC-10, MFCC-11
Variance	Mel-Spectrogram, MFCC-1, MFCC-3, MFCC-5, MFCC-8
Kurtosis	MFCC-12, MFCC-17, MFCC-18
25th Percentile	Chromagram, Root Mean Square
75th Percentile	MFCC-7, MFCC-11
Sum	MFCC-10, MFCC-12
Median	MFCC-5
Min	Zero Crossing Rate

3.4. Classifiers Evaluation and Results

- ➔ Sampling Technique: **5-Fold Stratified Cross-Validation**
- ➔ Metrics: **Accuracy**, **Recall**, **Precision**, **Macro-F1 Score**, **Matthews Correlation Coefficient**

Tested traditional models' evaluation results on IEMOCAP.

Model	Accuracy	Macro F1	Precision	Recall	MCC	Prediction Time
XGBoost	60.69±1.17	61.32	61.66	61.19	0.468	0.07
AdaBoost	60.04±0.95	60.76	61.29	60.59	0.459	0.41
Balanced RF	59.99±0.5	60.87	61.41	60.57	0.458	0.62
RF	59.77±0.72	60.43	60.97	60.30	0.456	0.38
Histogram Gradient Boosting	59.25±1.53	59.80	60.34	59.47	0.450	0.55
SVM	54.28±0.57	54.96	55.51	54.78	0.380	1.27
Linear Discriminant Analysis	54.04±1.38	55.06	55.01	55.23	0.379	0.01
Ridge	53.28±0.98	54.14	53.94	54.44	0.369	0.01
LSTM	51.96±1.02	52.87	54.0	52.54	0.349	1.18
CNN	50.41±0.95	51.25	51.61	52.69	0.340	0.81

3.5. Proposed Classifier

eXtreme Gradient Boosting (XGBoost)

XGBoost is a renowned and highly accurate model that uses **gradient-boosted decision tree ensembles**. It effectively mitigates overfitting by employing a combination of **L1 and L2 regularization** techniques. The model **sequentially** builds decision trees using **boosting**, where each subsequent tree **corrects errors made by the previous one**, with a focus on **addressing misclassified data points**.

Overall, XGBoost is a versatile and powerful algorithm that provides several advantages:

- Speed
- Performance
- Scalability
- Handling of Missing Valus
- Regularization
- Interpretability

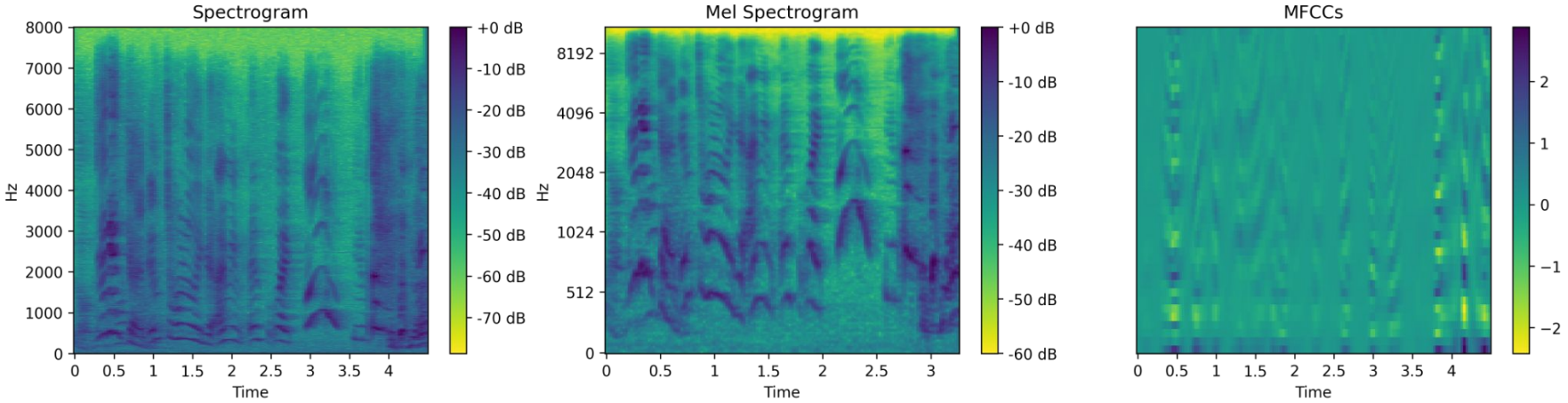
Python code for the proposed XGBoost (XGBoost Library):

```
XGBClassifier(
        max_depth=8,
        learning_rate=0.1,
        n_estimators=512,
        subsample=0.9,
        colsample_bytree=0.8,
        colsample_bylevel=0.8,
        n_jobs=-1
        )
```

4. Deep Learning-Based SER

4.1. Deep Learning Features

Features Dimensions:
For the 2D Models (Numeric Matrices): 1025x188 - Spectrogram, 256x188 - Mel Spectrogram, 40x188 - MFCC
For the 3D Models (PNGs): 224x224x3 - All Features



Graphical representations of the selected deep learning features.

4.2. Classifiers Evaluation and Results

- ➔ Sampling Technique: **5-Fold Stratified Cross-Validation**
- ➔ Metrics: **Accuracy**, **Recall**, **Precision**, **Macro-F1 Score**, **Matthews Correlation Coefficient**

Tested deep learning models' evaluation results on IEMOCAP.

Feature	Model	Accuracy	Macro F1	Precision	Recall	MCC	Prediction Time
Spectrogram Image	Resnet50	58.24±2.20	58.97	59.38	59.00	0.436	20.78
Mel Spectrogram Image	Resnet50	57.95±1.36	58.71	59.27	58.49	0.430	20.92
MFCC Image	Resnet50	56.59±0.45	57.29	58.59	56.67	0.410	23.19
Mel Spectrogram Image	VGG16	55.07±2.23	55.82	56.77	55.29	0.389	12.04
MFCC Image	VGG16	54.73±1.47	55.51	56.32	55.14	0.386	10.89
Spectrogram Image	VGG16	54.28±0.90	55.21	55.85	54.87	0.379	12.16
Mel Spectrogram Image	Xception	53.10±1.42	53.84	54.27	53.68	0.364	19.06
MFCC Image	Xception	52.78±0.96	53.47	54.10	53.22	0.359	18.33
Spectrogram Image	Xception	52.78±1.54	53.51	53.48	53.62	0.361	19.55
Spectrogram	2D-CNN	50.12±0.91	50.04	52.98	49.65	0.320	21.2
Mel Spectrogram	2D-CNN & RNN	48.02±1.14	47.93	48.60	48.47	0.298	20.05
MFCC	2D-CNN	46.70±0.85	47.13	49.53	46.75	0.275	5.18
Spectrogram	2D-CNN & RNN	46.01±1.77	47.09	47.37	46.87	0.269	32.07
MFCC	2D-CNN & RNN	45.56±1.15	46.26	46.29	46.25	0.263	12.29
Mel Spectrogram	2D-CNN	32.51±1.13	21.34	20.38	30.26	0.102	9.21

4.3. Proposed Classifier

ResNet50 using 3D Spectrogram Images

ResNet50, pre-trained on ImageNet, is a powerful deep neural network model with 50 convolutional layers, and with skip connections to enable the direct flow of information between layers, capturing complex representations of the input data.

Some advantages of ResNet50 are:

- **Addresses the vanishing gradient problem** with residual connections;
- Facilitates **transfer learning** with pre-trained models;
- **Fast training** for large-scale datasets;
- **High accuracy** in various image classification tasks.

Python code for the proposed Resnet50 (TensorFlow):

```
import tensorflow.keras as K
model = K.applications.resnet.ResNet50(weights='imagenet',
                input_shape=(224, 224, 3), include_top=False, pooling='avg')
model.trainable = False
inputs = K.Input(shape=(224, 224, 3))
x = model(inputs, training=False)
x = K.layers.Dense(64, activation='relu')(x)
x = K.layers.Dropout(0.5)(x)
outputs = K.layers.Dense(4, activation='softmax')(x)
model = K.Model(inputs, outputs)
model.compile(optimizer=K.optimizers.Adam(learning_rate=1e-3),
                        loss=K.losses.SparseCategoricalCrossentropy(),
                        metrics=['accuracy'])
```

SER Proposed Models

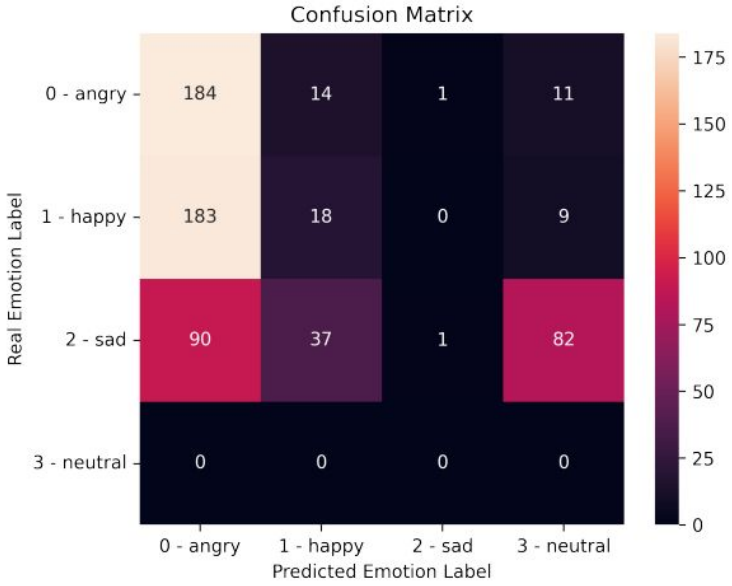
Cross-Dataset Validation

Proposed models trained on IEMOCAP and evaluated on different datasets.

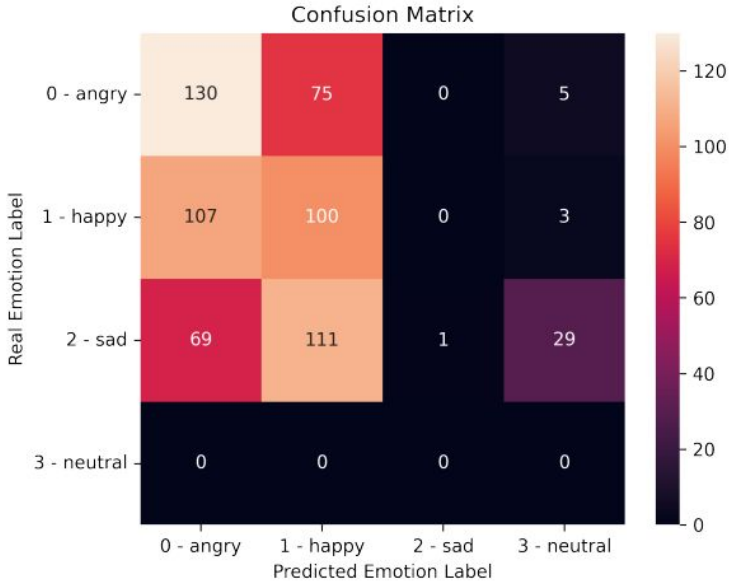
Dataset	Model	Accuracy	Macro F1	Precision	Recall	MCC	Time
eNTERFACE'05	Traditional	32.22	17.25	29.09	24.17	0.08	0.17
	Deep Learning	36.67	22.91	44.36	27.50	0.087	0.25
EMO-DB	Traditional	38.35	15.82	14.80	26.06	0.07	0.10
	Deep Learning	38.35	15.79	37.78	25.99	0.066	0.18
CREMA-D	Traditional	45.22	38.96	47.62	46.41	0.3151	0.10
	Deep Learning	54.14	47.71	51.68	52.98	0.407	0.30

SER Proposed Models

Cross-Dataset Validation



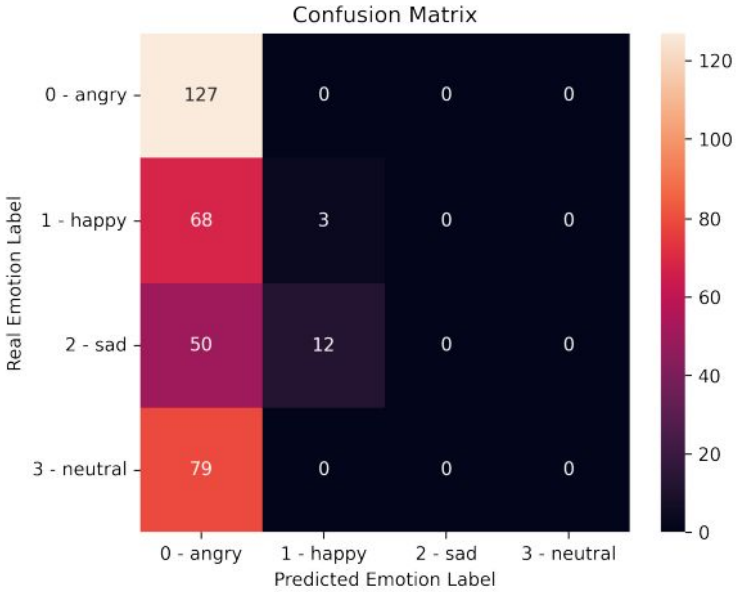
(a) eNTERFACE'05 traditional model confusion matrix.



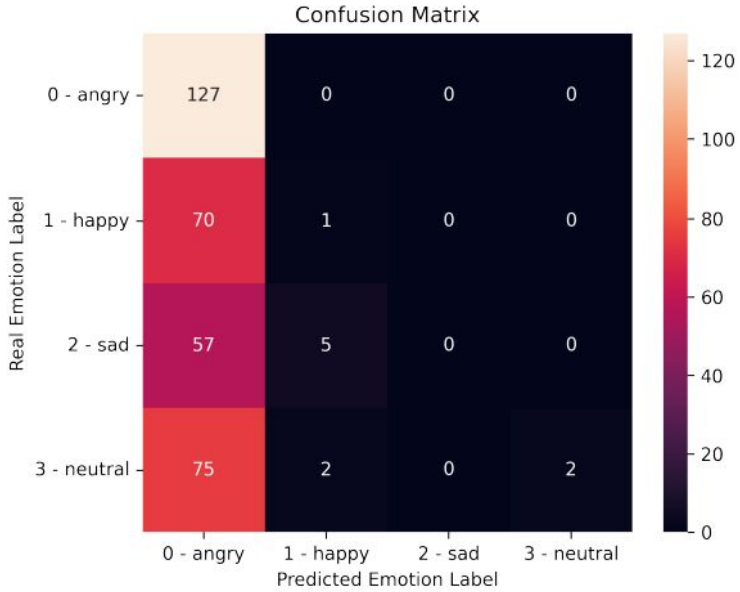
(b) eNTERFACE'05 DL model confusion matrix.

SER Proposed Models

Cross-Dataset Validation



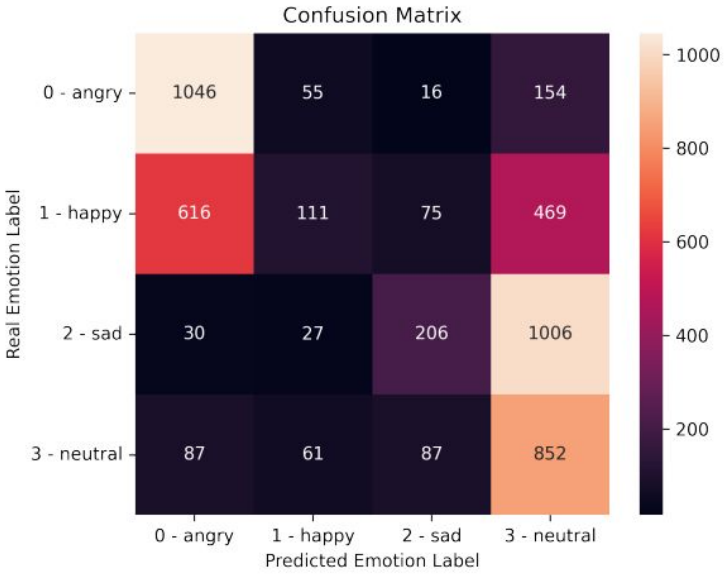
(c) EMO-DB Traditional model confusion matrix.



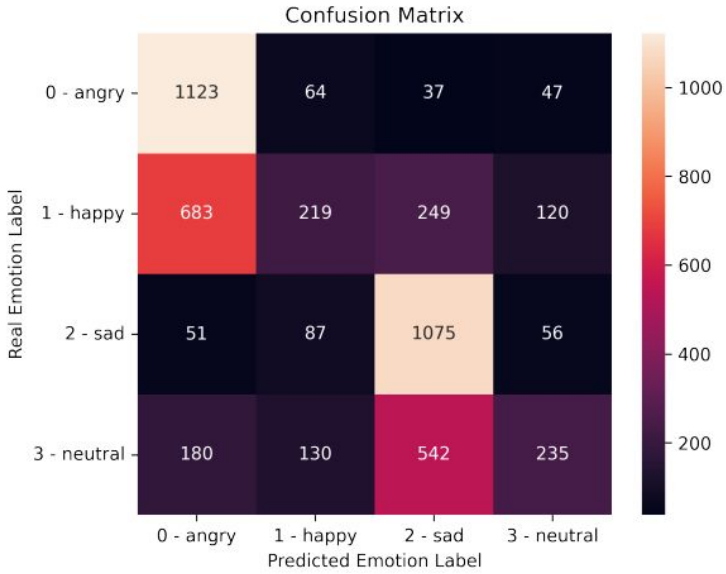
(d) EMO-DB DL model confusion matrix.

SER Proposed Models

Cross-Dataset Validation



(e) CREMA-D Traditional model confusion matrix.



(f) CREMA-D DL model confusion matrix.

SER Proposed Models

SOTA Comparison

SOTA SER classification models performance on IEMOCAP.

Model	Input	Evaluation Strategy	Accuracy (%)
Traditional Feature-Based SER Approaches			
Ensemble of RF, XGBoost and Multilayer Perceptron [53]	8-dimensional audio features vector	1 random 80:20 train-test split	56.00
Multi-level binary decision trees [52]	384 audio features vector	10 fold CV	58.46
XGBoost [Ours]	33 audio features vector	5-fold CV	60.69
CNN [54]	193 audio features vector	5-fold CV	64.30
Deep Learning-Based SER Approaches			
Resnet50 [Ours]	3-D Spectrogram Image	5-fold CV	58.24
CNN and RNN [58]	Log-Spectrogram	5-fold CV	64.22
3-D attention-based convolutional RNN [61]	3-D Mel-Spectrogram Image	10-fold CV	64.7
CNN and LSTM with attention [59]	Mel-Spectrogram	5-fold CV	67.0
Quaternion CNN [62]	Mel spectrogram encoded in an RGB quaternion domain	5-fold CV	70.46

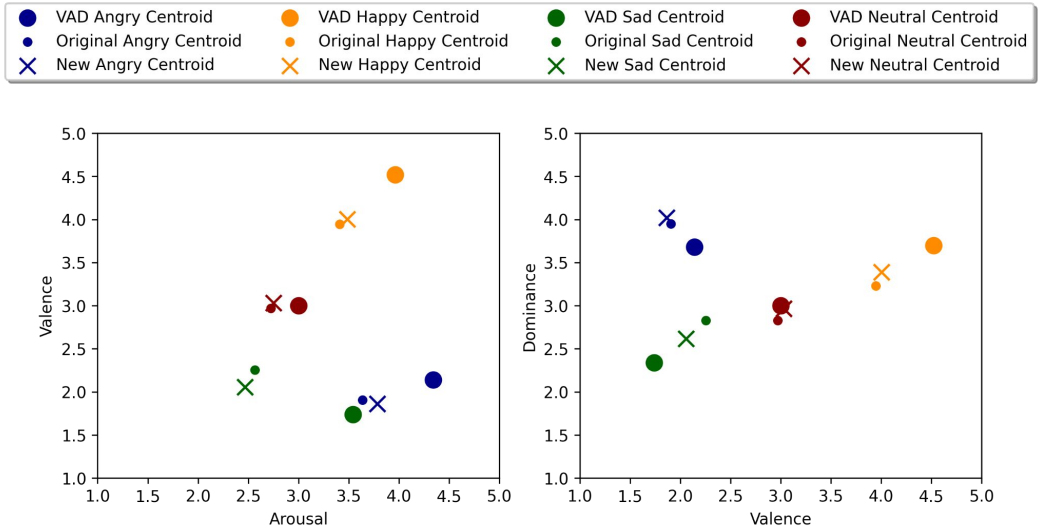
5. Data Stratification

To **uncover** the **shortcomings and biases** of the dataset and the proposed models, data was stratified based on: **recordings duration**, **speaker gender**, **discrete** and **dimensional labels**.

As a result of this process, a **set of conditions** was applied on the dataset, including a **minimum duration of 1 second**, and **dimensional annotations** must be **within a certain range** in relation to the discrete emotion. The resulting "stratified" data contained a total of 4200 audio files, with a nearly gender balanced distribution.

Maintained dimensional annotations range for each emotion category.

Emotion	Ranges Maintained		
	Arousal	Valence	Dominance
Angry	$]2, 5]$	$[1, 4.5]$	None
Happy	$[2.5, 5]$	$[3, 5[$	None
Sad	$]2, 5]$	$[1, 4]$	$[1, 4]$
Neutral	$]2, 4[$	$[2, 4]$	$]2, 4[$



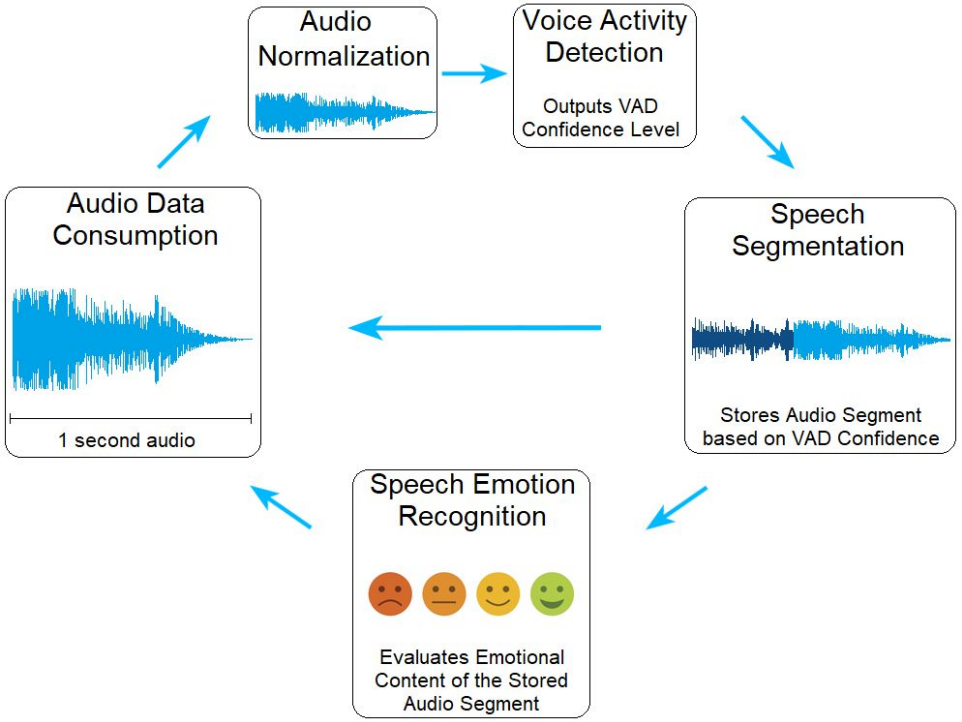
5.1. Validation

Proposed models trained on the entire and stratified IEMOCAP and evaluated on different datasets.

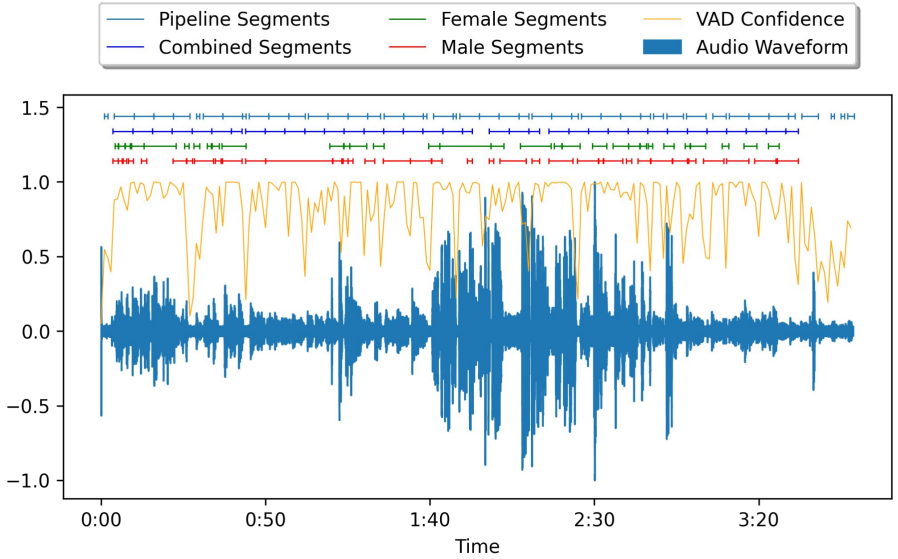
Dataset	Model	Accuracy	Macro F1	Precision	Recall	MCC
eNTERFACE'05	Traditional	32.22	17.25	29.09	24.17	0.080
	Stratified Traditional	32.38	16.11	29.23	24.29	0.077
	DL	36.67	22.91	44.36	27.50	0.087
	Stratified DL	37.14	22.39	43.51	27.86	0.073
EMO-DB	Traditional	38.35	15.82	14.80	26.06	0.065
	Stratified Traditional	38.64	16.82	35.42	26.48	0.077
	DL	38.35	15.79	37.78	25.99	0.066
	Stratified DL	38.05	15.22	34.71	25.63	0.052
CREMA-D	Traditional	45.22	38.96	47.62	46.41	0.315
	Stratified Traditional	46.06	39.90	47.85	46.79	0.313
	DL	54.14	47.71	51.68	52.98	0.407
	Stratified DL	55.29	50.06	54.11	54.05	0.417

6. SER Pipeline

Architecture



Pipeline Validation



A session from IEMOCAP segments and the pipeline's detected segments.

Annotated emotions of the entire IEMOCAP dataset and the pipeline's predicted segments.

Emotions	Dataset Labeled Segments	Pipeline Predicted Segments
Neutral	1708	1292
Happiness	1636	2930
Angry	1103	2144
Sadness	1084	1332
Non-Identified	2510	-
Frustration	1849	-
Surprise	107	-
Fear	40	-
Disgust	2	-

Conclusion

1. **XGBoost** and **ResNet50** models that have a **competitive performance** and potential for real-world application;

2. **Small set of hand-crafted features (33)** that achieves high results in comparison to the SOTA.

3. Application of **transfer-learning techniques** for image recognition in the SER area.

4. Identification of **high-quality data** and **biases** through **data stratification** processes;

5. **SER pipeline** for efficient **detection** and **construction of speech segments**;

Future work: Fine-tuning models, exploring additional datasets, and leveraging multimodal approaches (speech, facial expressions, text) to enhance emotion recognition accuracy.

Models development for audio evaluation in affective computing

29/06/2023

Mário Francisco Costa Silva
Master in Informatics Engineering
Department of Electronics, Telecommunications and Informatics