Universidade de Aveiro
2023

Mário Francisco
Costa Silva

Avaliação Multimodal de Dados Afetivos em Sessões de Videoconferência

Multimodal Evaluation of Affective Data in Videoconference Sessions

Mário Francisco
Costa Silva

Avaliação Multimodal de Dados Afetivos em Sessões de Videoconferência

Multimodal Evaluation of Affective Data in Videoconference Sessions

"*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

Universidade de Aveiro
2023

Mário Francisco
Costa Silva

Avaliação Multimodal de Dados Afetivos em Sessões de Videoconferência

Multimodal Evaluation of Affective Data in Videoconference Sessions

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica da Doutora Susana Manuela Martinho dos Santos Baía Brás, Professora Investigadora no Instituto de Engenharia Eletrónica e Telemática de Aveiro do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Doutor Ilídio Castro Oliveira, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática
.

Palavras Chave

Computação Afetiva, Processamento de Voz, Reconhecimento de Emoções, Multimodalidade, Aprendizagem Automática

Resumo

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Keywords Affective Computing, Voice Processing, Emotion Recognition, Multimodality, Machine Learning

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

List of Figures

List of Tables

Glossary

SER	Speech Emotion Recognition
VAD	Voice Activity Detection
MFCC	Mel-frequency Cepstral Coefficients
VAD	Valence-Arousal-Dominance
VAD	Voice Activity Detection
CNN	Convolutional Neural Network
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LSTM	Long-Short Term Memory
DL	Deep Learning
SOTA	State-Of-The-Art
RNN	Recurrent Neural Network
MCC	Matthews Correlation Coefficient

CHAPTER 1

SER Development

1.1 Datasets

In this section, we will detail the two utilized datasets that were used for the development of our SER system. By utilizing two distinct datasets for our analysis, we are able to make the models more robust and effective, making the results less prone to overfitting.

The first dataset was used as a development set, which we used to explore and select the best features for the traditional approach models.

The second dataset was used as a training and test set for evaluating the performance of our predictive models and determining the most effective strategies.

1.1.1 eNTERFACE'05

The eNTERFACE'05 emotion database [24] was designed and collected during the eNTERFACE'05 workshop in 2006. The dataset contains audio and visual data from 42 subjects, coming from 14 different nationalities. Among the subjects, a percentage of 35 are men, while the remaining 7 are women, and, all the experiments were driven in English.

Table 1.1: eNTERFACE'05 subjects nationalities

Country	Number of Subjects	Country	Number of Subjects
Belgium	9	Cuba	1
Turkey	7	Slovakia	1
France	7	Brazil	1
Spain	6	U.S.A.	1
Greece	4	Croatia	1
Italy	1	Canada	1
Austria	1	Russia	1

This dataset contains six discrete annotated emotions: 1. anger 2. fear 3. surprise 4. happiness 5. sadness 6. disgust. Each subject was asked to listen to six successive short stories, each eliciting a particular emotion. If two human experts judged the reaction expressing the emotion unambiguously, then the sample was added to the database. Afterward, they were recorded saying five different sentences for each emotion, and, in total, there are 212 video and audio sequences per annotated emotion, recorded with a sample rate of 16000 Hertz.

The selection of this dataset for feature analysis and selection was based on several factors. Firstly, the controlled environment of the dataset ensured that the data was collected under controlled conditions, which minimized the impact of external factors that could have influenced our analysis. Moreover, the diversity of the subjects included in this dataset made it possible to identify and select features that are representative of several groups of people.

Another key factor in choosing this data was its size. Due to the limited size of this dataset, we are able to utilize computationally expensive methods, such as feature selection algorithms, that would have been prohibitively expensive with larger datasets.

Finally, the elicited nature of the data in this dataset was considered an essential aspect of our selection process. Elicited obtained data tends to be more genuine than acted, therefore, it provides a more accurate representation of video conferences' natural contexts.

1.1.2 IEMOCAP

The IEMOCAP database [25], created in 2008, is an acted and elicited multimodal and multi-speaker database. It consists of 12 hours of audiovisual data, including video, speech, motion capture of face, and text transcriptions.

Sessions were manually segmented into utterances, spoken by 10 (5 female and 5 male) professional actors in fluent English. Each utterance was annotated by at least 3 human annotators in 9 categorical attributes, and, in addition, it was annotated with 3-dimensional attributes using the Valence-Arousal-Dominance (VAD) emotion model. Similar to the development dataset, this data was collected using emotion elicitation techniques such as improvisations and scripts. Figure 1.1 from the research article **??**, demonstrates that there is a similar amount of annotated labels on scripted and spontaneous sessions on this dataset.

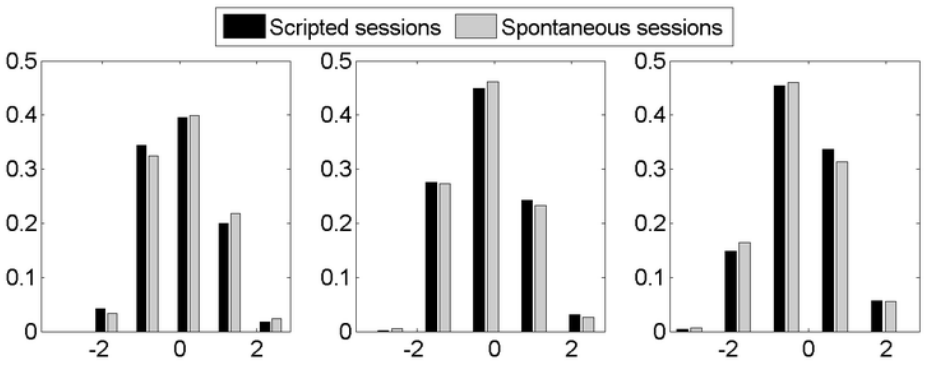


Figure 1.1: Distribution of the emotional content of the IEMOCAP corpus in terms of (a) valence, (b) activation, and (c) dominance. The results are separately displayed for scripted (black) and spontaneous (gray) sessions.

Several researchers when utilizing the IEMOCAP dataset tend to only perform 4 class sentiment analysis, and also, consider the emotion of excitement as happiness to even out the distribution of files per emotion, as shown in the table 1.2, ending up with a total of 5531 audio files, recorded with a sample rate of 16000 Hertz.

Table 1.2: Number of Audio Files Used per Emotion from the IEMOCAP dataset.

Emotion	Number of Audio Files
Anger	1103
Happiness	1636
Neutral	1708
Sadness	1084

Overall, this second dataset is a well-suited resource for our study, as the multimodal data, annotated using both discrete and dimensional models, allows us to perform a wide range of investigations, researchers have also noted the high quality of this dataset, being frequently used in the literature for evaluating emotion recognition models. This enables us to make well-founded comparisons of our own developed models, which is why we utilized IEMOCAP as a training and testing dataset for our SER models and to explore strategies and biases by stratifying the data.

1.2 Audio Preprocessing

TODO: rewrite this with more details.

Audio preprocessing is an essential step in preparing collected audio data for use in machine learning models. The preprocessing techniques involve noise reduction, feature extraction, analysis, and selection. This section discusses important aspects of audio preprocessing, including noise reduction and audio trimming.

1.2.1 Noise Reduction

One common strategy for denoising music is spectral gating, which involves gating the signal only on high-level sounds. Non-stationary noise reduction is an extension of stationary noise reduction that allows the noise gate to change over time. In this method, a spectrogram is calculated over the signal, and a time-smoothed version of the spectrogram is computed using an IIR filter applied forward and backward on each frequency channel. A mask is computed based on the time-smoothed spectrogram, which is then smoothed with a filter over frequency and time. The mask is applied to the spectrogram of the signal and then inverted.

To implement these noise reduction techniques, we recurred to the *noisereduce* library. This algorithm relies on the spectral gating method and estimates a noise threshold for each frequency band of the signal/noise. This threshold is used to compute a mask, which gates noise below the frequency-varying threshold. The Code Snippet 1 shows how we implemented the noise reduction technique.

```
noisereduce.reduce_noise(
        y=y, sr=16000, n_fft=2048, hop_length=512, prop_decrease=.75, time_constant_s=1
)
```

Code Snippet 1: Python code for applying noise reduction using the *noisereduce* library.

1.2.2 Audio Trim

Trimming silence at the beginning and end of audio files is another important preprocessing step. By removing the silence from audio files, the resulting audio data will be more focused on the relevant audio content and will be easier to process by machine learning algorithms.

Upon visually observing several wave plots we ended up considering 30 decibels or lower as silence and trimmed every audio file at the beginning and end of each audio.

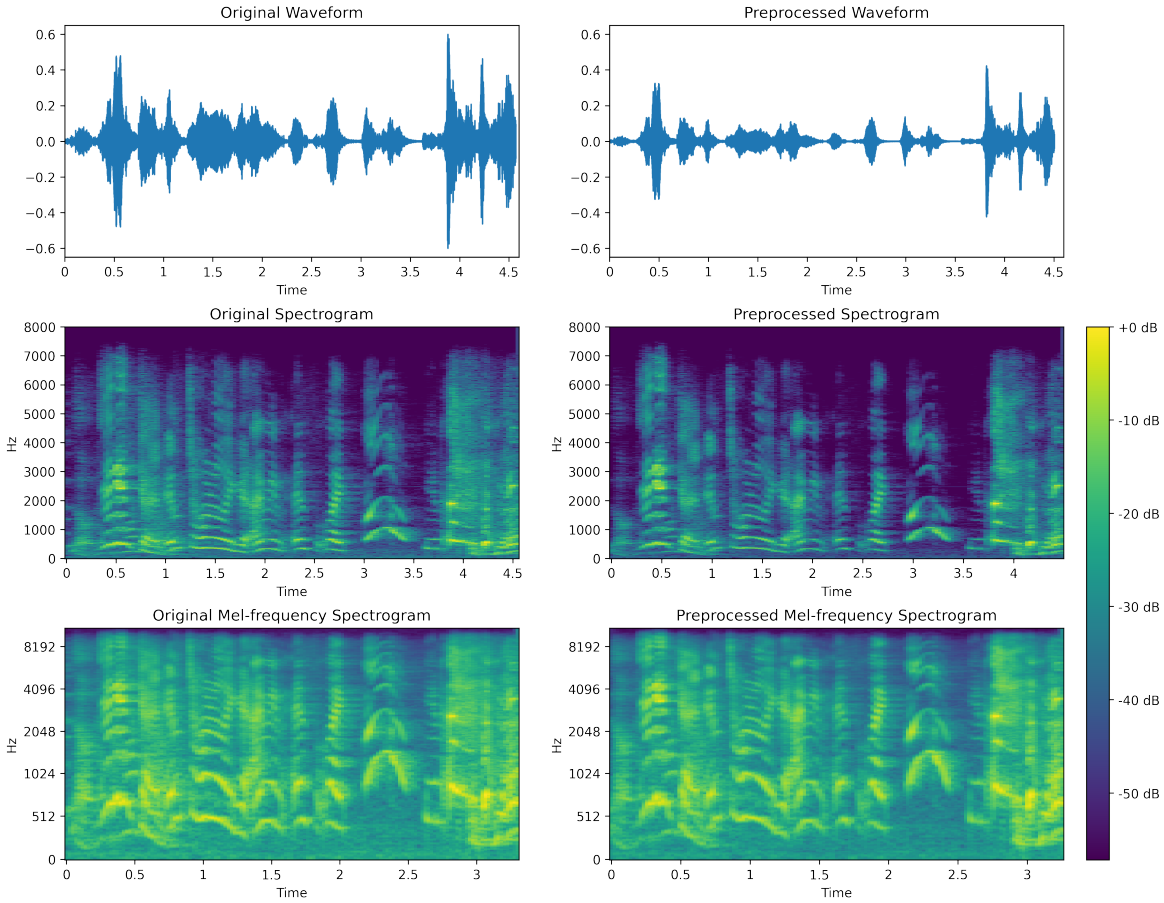


Figure 1.2: Visual representations of audio features before and after preprocessing.

1.3 Traditional Feature-Based SER

1.3.1 Feature Extraction

Feature extraction is an essential component in audio analysis tasks, as it allows the transformation of raw audio data into a set of informative features that can capture key characteristics of the signal.

In this regard, the widely-used Librosa toolkit was employed to extract various audio features, from the audio files of the IEMOCAP dataset, which were subsequently processed using statistical metrics. The extracted features and associated metrics are summarized in Table 1.3, having in total, extracted 327 features.

Table 1.3: Extracted Audio Features and Statistical Functions Applied to Them

Audio Features	Statistical Functions
	Minimum
Mel-frequency Cepstral Coefficientss (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 detailed on the Feature Selection section.
[2] Only for the MFCCs.

1.3.2 Feature Analysis

One important task following feature extraction is to analyze and interpret the extracted data to gain a deeper understanding of the audio signals and the features that describe them.

Audio Signal Study

In this process, we visually analyzed and interpreted the features' data by graphically representing each feature from an audio segment. The figures in Section **??** of the appendix demonstrate some of the graphics we used to visualize the data.

Spikes Metric

Initially, wave plots were observed, and we noted consistency in the number of high values. For this reason, we created a custom metric that calculates those high values, which we called "spikes", from the features' data.

In Figure 1.3, it is possible to visualize the zero crossing rates' wave plots in different emotions. The horizontal line represents the threshold that we considered, any value above was considered to be a spike, which is annotated with red dots in the graphic. The threshold used was manually tested and obtained decent consistency of the number of spikes, within an emotion, by using the mean value of the feature plus 2% of the standard deviation. To account for different-length audio signals, it was also divided the number of spikes to the total length of the data, as the Code Snippet 2 demonstrates. Consequently, this metric was also tested and applied to every other audio feature.

```python
def spikes(data):
        num_spikes = 0
        mean = np.mean(data)
        std = np.std(data)

        threshold = mean + np.abs(std) * 2 / 100
        for value in data:
                if value >= threshold:
                        num_spikes += 1

        return num_spikes / len(data)
```

Code Snippet 2: Python code for calculating the spikes metric.

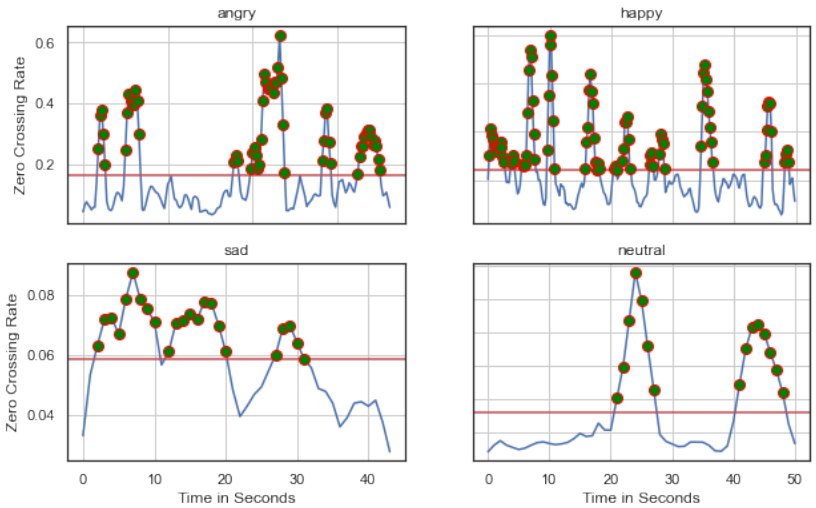


Figure 1.3: Zero crossing rate wave plot annotated with spikes.

Bar Plots

Furthermore, bar plots were useful for viewing the overall extracted features' data plainly and quickly, and to understand the numeric values of each feature and metric used on it.

For example, figure 1.4 shows clear differences in the mean values for some metrics used on the Mel Spectrogram.

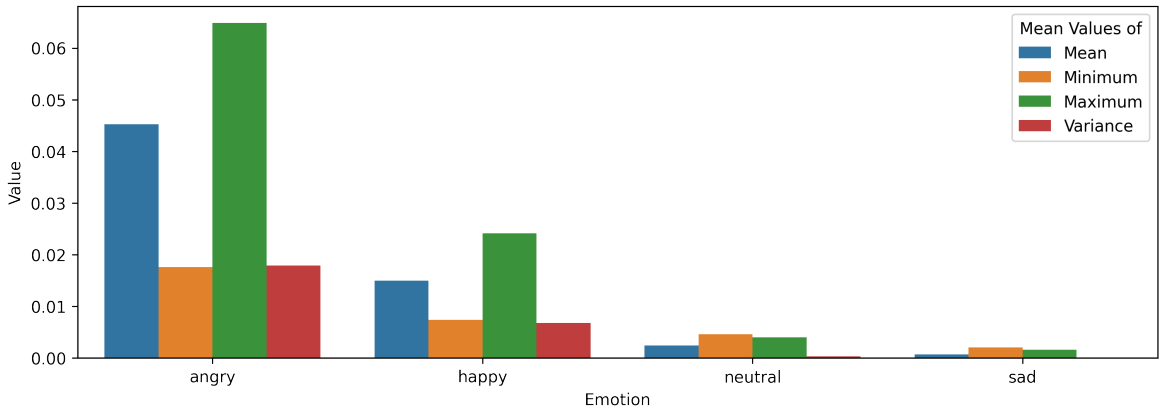


Figure 1.4: Bar plots mean for metrics used on the mel-scaled spectrogram feature

Wave Plots with Surrounding Areas

During the feature study process, it was observed the wave plots of some features surrounded by a small area above and below the original wave (defined through a selected threshold). This was done to corroborate how well the feature describes different emotions. A high degree of overlap between surrounding areas of a feature on a given emotion for different subjects could indicate that the feature is relevant for representing that emotion.

Figure 1.5 is an excerpt of the figure **??** in the appendix, and it demonstrates an example of this analysis for the zero crossing rate with 5 different subjects on the same sentence for the anger emotions.

From this graphic, it was observed that there is a sufficient amount of overlap between the surrounding areas for each emotion to conclude that the feature has some utility for describing each emotion. However, due to the different lengths of each audio segment, it is ambitious to guarantee this conclusion.

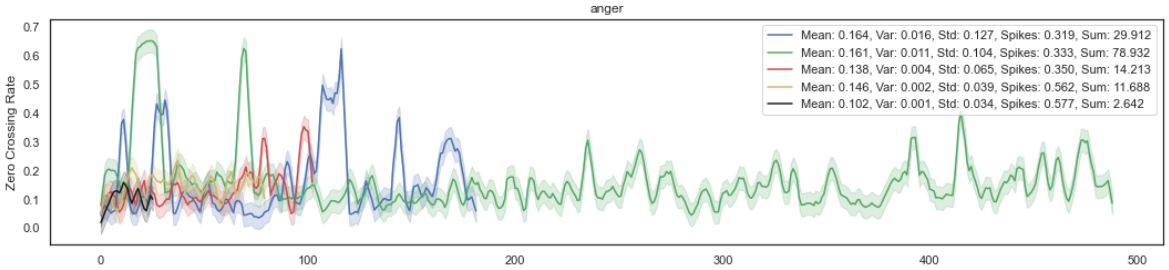


Figure 1.5: Zero crossing rate wave plot with a surrounding area of five male subjects for the same utterance with the anger emotion.

This same idea can also be used to determine whether a feature is favorable for creating a distinction between different emotions, which is naturally useful for the problem of classifying emotions. The conclusion can be drawn by observing the opposite of the previous example. If the areas around the zero crossing rate do not coincide too heavily, it is an indicator that the feature could be adequate for distinguishing different emotions.

Figure 1.6 displays six zero crossing rates of one subject saying the same sentence but expressing different emotions. As previously mentioned, since audio lengths are different, it is difficult to draw a direct and well-founded conclusion.

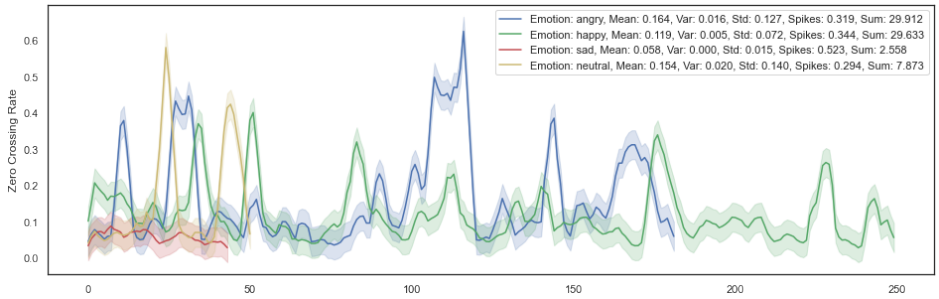


Figure 1.6: Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions.

Overall, this approach of surrounding wave plots with areas provided us valuable insight into the ability of a feature to describe and distinguish emotions, though it is a little limited by the varying lengths of audio segments.

Variation Plots

Another graph made was a variation plot, to perceive the differences in the features' values, across several audios for the same emotion. Figure 1.7 shows an example of this type of plot for the mean zero crossing rate value across 50 speech utterances for all emotions.

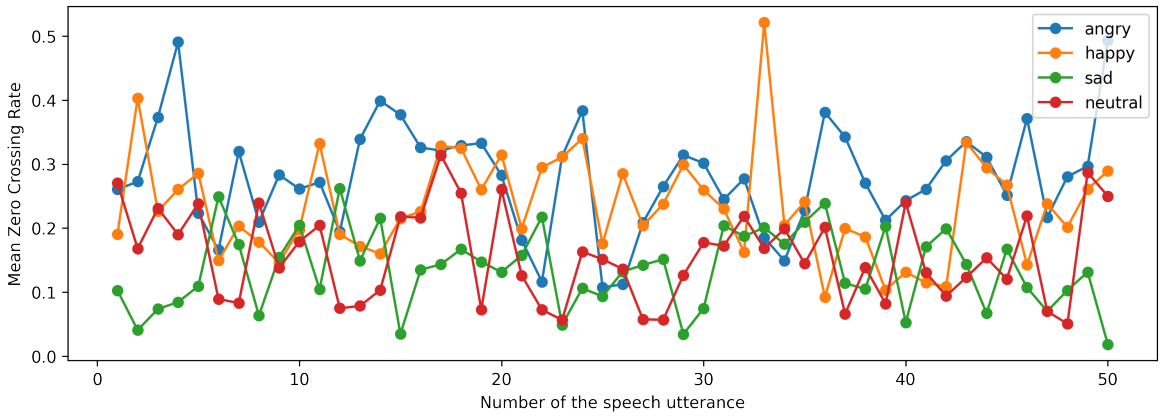


Figure 1.7: Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions

A common observation for most extracted feature plots was that the values were not consistent across multiple audio segments for the same emotion. However, the number of audio segments used in this study was relatively low (only 50) to observe big variability changes, but increasing the number of audio segments would also make it more challenging to observe such variability through a simple visual inspection.

Box Plots

Finally, we employed box plots to visualize the distribution of the features on different subjects, as well as to compare the values for each emotion. An example of this type of plot is shown in Figure 1.8, which displays the mean zero crossing rate feature for all emotions and different subjects.

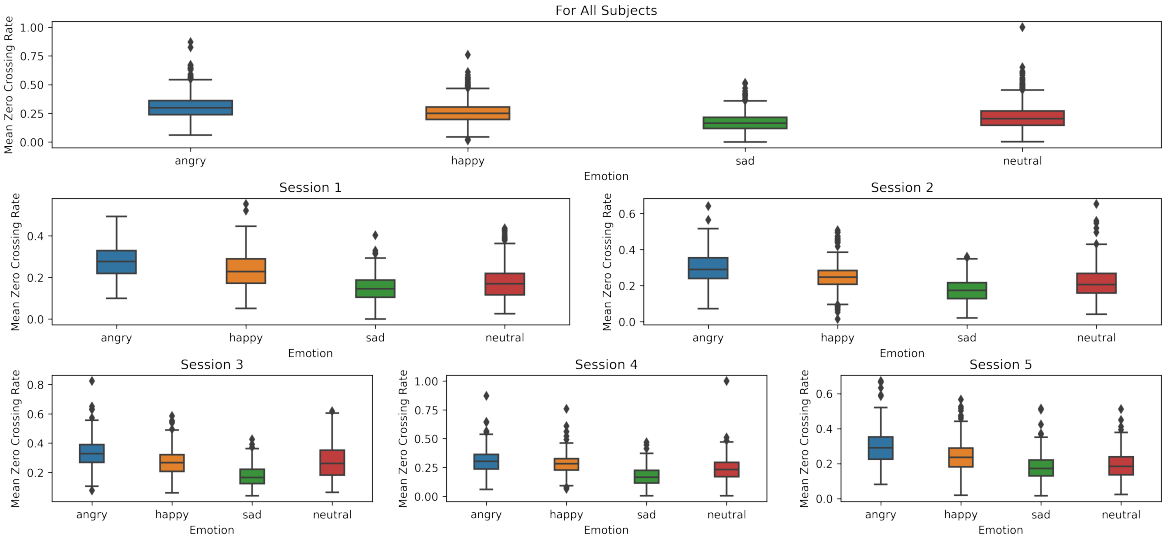


Figure 1.8: Zero crossing rate mean values box plot for all emotions and different subjects

The primary purpose of using these plots was to provide a simple and intuitive representation of

each feature. By comparing the values across all subjects or a selected few, any noticeable differences in feature values for each emotion could be easily perceived.

1.3.3 Feature Selection

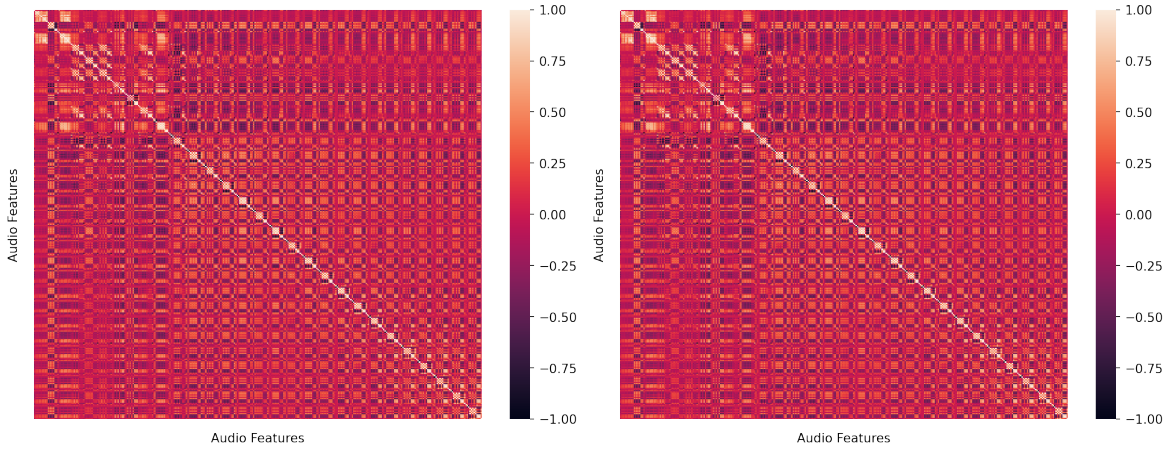
After the process of feature analysis, the next step in SER development is feature selection. Feature selection is a technique to choose a subset of the original set of features that are most relevant for the given task. The process of feature selection is aimed to improve the accuracy of the model and reducing the problem's complexity by removing redundant or irrelevant features.

The objective is to choose a smaller set of features that retain enough information for good classification performance while being computationally efficient. Hence, a smaller subset of features that can provide effective classification results is preferred over the larger set of features that may be computationally expensive and redundant.

High Correlation Elimination

Correlation among our extracted features is common since many of them use the same audio descriptor but with a different metric applied to them. Therefore, a correlation matrix for all 327 extracted features was calculated using the Pearson method, presented in figure **??**.

A high correlation elimination was performed by selecting every pair of features with a Pearson correlation coefficient absolute value of 0.6 or above, then it was removed the feature with the highest average correlation value with all the other features. This process resulted in the elimination of 229 features, leaving 98 features for subsequent analysis. The correlation matrix after the feature selection process is presented in figure 1.9b.



(a) Correlation Matrix of all the Features. **(b)** Correlation Matrix after High Correlation Elimination.

Figure 1.9: Audio Features' Pearson Correlation Matrices Before and After High Correlation Elimination.

Selecting an Initial Classifier

Along this process, it became necessary to choose a model to be used in computationally expensive feature selection methods. Consequently, several estimators were tested for their performance in classifying emotions.

To this end, we conducted 5-fold cross-validation and compared the mean and standard deviation accuracies of all folds, as well as the total execution time for various classifiers from the scikit-learn library [92], using the features obtained after the previous process, as shown in Table 1.4.

Table 1.4: Performance of various classifiers in 5-fold cross-validation using the features obtained after high correlation elimination.

Classifiers	Accuracy	Training Time (s)
XGBoost	0.617±0.013	17.628
Random Forest	0.578±0.010	7.451
Ridge	0.565±0.014	0.078
Extra Trees	0.561±0.005	1.831
AdaBoost	0.520±0.008	12.205
C-Support Vector	0.504±0.018	5.081
DecisionTree	0.450±0.022	1.886
Multi-layer Perceptron	0.446±0.027	4.821

Based on the evaluation results, the Random Forest classifier was chosen for further analysis. This model exhibited the second best average accuracy across the 5 folds, however it was more than twice as fast the XGBoost that had highest average accuracy. Therefore, it was deemed suitable for performing computationally expensive feature selection methods.

Backwards Selection

In the pursuit of completing the feature selection process, a sequential feature selection with backward propagation was employed. This method involves performing a 5-fold cross-validation with the previously selected Random Forest classifier, using all features except one, and then removing one feature based on the lowest mean accuracy of the 5 folds. This iterative process continues until only one feature remains.

A method was then developed to select the furthest and highest accuracy. This method resembles the standard maximum, but it multiplies the maximum value by a threshold value of 0.99, so that it can find close to the maximum values that have more features removed, creating a balance between accuracy and number of features. Figure 1.10 displays the mean accuracies obtained at each step and the chosen furthest highest accuracy.

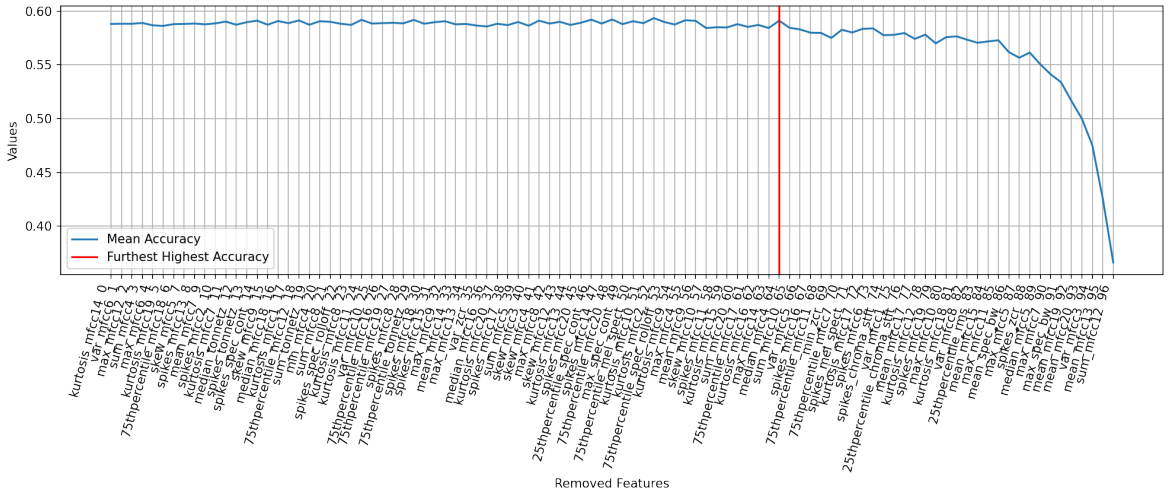


Figure 1.10: Sequential Feature Selection with Backward Propagation using the Mean Accuracy as the Selection Method.

This process led to the elimination of 65 features from the initial set of 98 obtained after the high correlation elimination, leaving a total of 33 features, as shown in Table 1.5.

Table 1.5: 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

Feature Selection Evaluation

To assess the feature selection quality on the development dataset, we trained and evaluated the predictions of a Random Forest Classifier model with different sets of features, using accuracy as the evaluation metrics. The results are presented in Table 1.6. Additionally, we plotted the confusion matrix of the predictions, which can be found in Appendix **??**.

Table 1.6: Evaluation Metrics of Random Forest Predictions Using Different Sets of Features and 5-Fold Cross Validation.

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

The results of the feature selection techniques have shown that while there may be a small loss of accuracy between the initial set of extracted features and the set obtained after the high correlation

elimination, the model can achieve comparable accuracy while using only around 70% of the original set. Moreover, applying backward selection to the remaining 98 features yields optimal results by maintaining the accuracy and reducing the original feature set by approximately 90%.

By using these feature selection techniques, we managed to remove redundant and irrelevant features, which also decreases the models' complexity and size, providing several practical benefits for its implementation and interpretation.

1.3.4 Classifiers Evaluation and Selection

Evaluation Strategy

TODO: explain each metric maybe

Evaluating a model is an essential step, since a wrongful evaluation may lead to deception in terms of the results obtained. It should be uniform for every model, and, it should be as meaningful as possible to the classification objective.

As mentioned previously, for this part and the subsequent parts of the SER development, the second dataset IEMOCAP is the one under use.

For the reasons above, it was decided to utilize 5-fold cross-validation, and, in terms of metrics, we decided to calculate 5 folds averages accuracy, macro-f1 score, precision, recall, and Matthews Correlation Coefficient (MCC). A confusion matrix of the predicted and real labels was also plotted, since it provides helpful insights, not only into the errors being made by the classifier but also, the types of errors occurring.

These were also the most recurred methods in state-of-the-art research, which provides more fairness to model comparisons.

Classifiers Exploration

An Automatic Machine Learning technique, namely Auto-SKLearn ensemble model **feurerneurips15a**, was employed to search for the best models and ensembles by exploring a vast space of possible algorithms and hyperparameters using a meta-learning approach.

After training the Auto-SKLearn model, the most influential classifiers of the ensemble were identified, resulting in an ensemble with 9 classifiers, including Random Forests with different hyperparameters, Linear Discriminant Analysis, Histogram-Based Gradient Boosting, Multilayer Perceptron, and Linear Passive Aggressive. Each model present in the ensemble was then tested and explored.

Next, a simple Convolutional Neural Network was tested, and another AutoML technique, AutoKeras **jin2019auto**, was applied to create an optimized deep-learning model.

Results and Conclusions

The results obtained from the tested models were compiled and exhibited in Table 2.1. The highest-ranked Random Forest from the Auto-SKLearn ensemble model achieved the best results, leading to its selection as the base estimator of an AdaBoost classifier, improving slightly the overall Random Forest results.

Table 1.7: Tested Classification Models 5-Fold Cross-Validation Performance on IEMOCAP.

Model	Accuracy	Macro F1	Precision	Recall	MCC	Training Time
AdaBoost	60.04±0.95	60.76	61.29	60.59	0.459	8.45
Random Forest	59.77±0.72	60.43	60.97	60.30	0.456	50.62
Histogram Gradient Boosting	59.25±1.53	59.80	60.34	59.47	0.450	133.31
XGBoost	58.09±1.34	58.74	59.30	58.44	0.431	7.37
Balanced Random Forest	56.92±0.83	57.31	56.90	59.84	0.432	11.29
Ridge	53.28±0.98	54.14	53.94	54.44	0.369	0.02
Linear Discriminant Analysis	54.04±1.38	55.06	55.01	55.23	0.379	0.06
Long-Short Term Memory (LSTM)	51.0±0.54	51.42	51.35	51.87	0.339	775.95
Convolutional Neural Network (CNN)	50.41±0.95	51.25	51.61	52.69	0.340	340.74

Table 1.8: SOTA Traditional Classification Models Performance on IEMOCAP.

Model	Input	Accuracy
Dilated Residual Network **Li_2019**	Audio Features	67.4
Recurrent Neural Network (RNN) W/ Attention **Lu_2020**	Audio and Text Features	72.6
Deep CNN **Issa_2020**	193 Audio Features	64.30

The application of data preprocessing techniques to clean the audio data was performed to handle different technical settings. However, in some cases, noise may be part of the signal of interest, and removing it may cause the algorithm to misinterpret the signal.

The state-of-the-art results presented in Table 1.8 are not directly comparable, as authors use different data and evaluation methodologies. Additionally, some authors do not consider the emotion of excitement as happiness or perform different validation methods such as 10-fold cross-validation.

Upon analyzing the explored classifiers results, AdaBoost with Random Forest as the base estimator is the best candidate, reaching an average accuracy of 60.04% while utilizing only 33 audio features which makes it a model that is relatively simple, fast to train and make predictions, while also being generalized to different datasets due to the preprocessing techniques applied. The Python code for the model was implemented using the SK-Learn library, version 1.2.1, and is presented on the following Code Snippet 3.

```
AdaBoostClassifier(estimator=RandomForestClassifier(n_estimators=512))
```

Code Snippet 3: Python code for the selected AdaBoost classifier using the traditional-based SER approach.

1.4 Deep Learning-Based SER

This section presents the exploration of using deep learning classifiers for audio-based emotion recognition, focusing on the use of various features for the classification task.

1.4.1 Deep Learning Features

Initially, three different features were extracted from the raw audio signals, using the Librosa library. The numeric values of the extracted features were saved into a Pickle file, while the visual representation of the feature was saved as a Portable Network Graphic (PNG) file. The PNG file was generated using a Matplotlib figure with 100 dots per inch, without the axis and the frame. The color map used was *viridis_r*.

The 2D deep learning models employed in this study required the input data to possess consistent dimensions. To this end, the numeric data of every feature used only the first 6 seconds of every audio file, with shorter audio signals padded with trailing zeros to achieve the required length.

The first feature explored was the spectrogram. The Short-time Fourier transform (STFT) was used to calculate the spectrogram, using a windowed signal length of 2048, after padding with zeros. This resulted in matrices with a dimension of $1025x188$. The amplitude spectrogram was converted to a dB-scaled spectrogram, which was then used for the PNG file.

Another feature explored was the Mel Spectrogram. For this, the previously calculated spectrogram was mapped onto the mel scale, using 256 Mel bands. This resulted in matrices with a dimension of $256x188$. The dB-scaled Mel Spectrogram was also used for the PNG file.

The third feature explored was the MFCC as they are commonly used for audio signal processing tasks due to their ability to capture the spectral characteristics of audio signals. 40 MFCC were extracted from the previously calculated Mel Spectrogram, resulting in matrices with a dimension of $40x188$.

These three used features are displayed in figure 1.11.

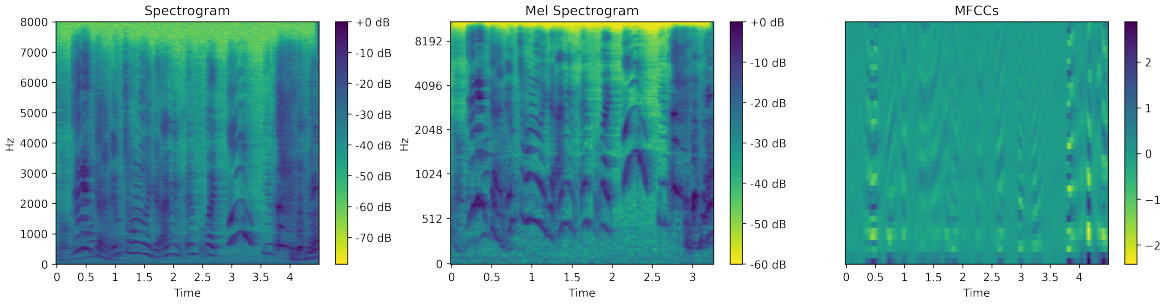


Figure 1.11: Graphical representations of the features used as input for the DL classifiers.

1.4.2 Classifiers Evaluation and Selection

Evaluation Strategy

The classifiers were trained using a Tesla P4 GPU on the Google Colab service. In order to evaluate the performance of the classifiers, 5-fold cross-validation was performed using the stratified K-fold strategy to preserve the percentage of samples for each class. The models were trained for 80 epochs, with a batch size of 128 and a learning rate decay of 10% every 10 epochs. The Adam optimizer was used for all models with a learning rate of 0.001.

The evaluation of the classifiers was done based on various metrics including accuracy, precision, recall, and macro F1-score. The training time was also annotated and the corresponding confusion

matrices were also plotted, present in the appendix **??**. These metrics were computed using the average across the 5 folds, without using class weights, to give an overall performance overview of the classifier.

Numeric Data Classification

Image Classification

In our study of deep learning-based SER using images, we utilized transfer learning techniques with three different pre-trained models: ResNet50, VGG16, and Xception.

ResNet50, VGG16, and Xception are popular deep convolutional neural networks (CNN) that have shown outstanding performance in various computer vision tasks, including image classification. We used the pre-trained versions of them on the large-scale ImageNet dataset, which contains millions of labeled images belonging to thousands of different classes. This means their weights and biases have already been adjusted for the ImageNet dataset, and they have learned how to extract meaningful features from images, which helps improving the accuracy and generalization ability of the models, as it allows them to recognize patterns and shapes that are common across a wide range of images.

To prepare the data for these models, we loaded the images with a dimension of $224x224x3$ using the TensorFlow Keras *load_img* function from the *preprocessing.image* module. The images were then converted into arrays using the *img_to_array* function from the same module. In addition, before inputting the data into the models, we applied the respective preprocessing technique for each classifier. For example, we used the *preprocess_input* function from the Tensorflow Keras *applications.resnet50* module for the ResNet50 classifier.

In the transfer learning technique, all layers of the chosen classifier were frozen, and a new Dense layer with 64 units with *relu* activation was added to the model. A Dropout layer with a 0.5 rate was then included to avoid overfitting, followed by a Dense layer with 4 units with *softmax* activation to output the predicted emotion.

Through the implementation of these pre-trained models with transfer learning, we aim to harness their robust feature extraction abilities and significantly reduce the training duration required for the inherently computationally intensive 3D classification task at hand.

Results and Conclusions

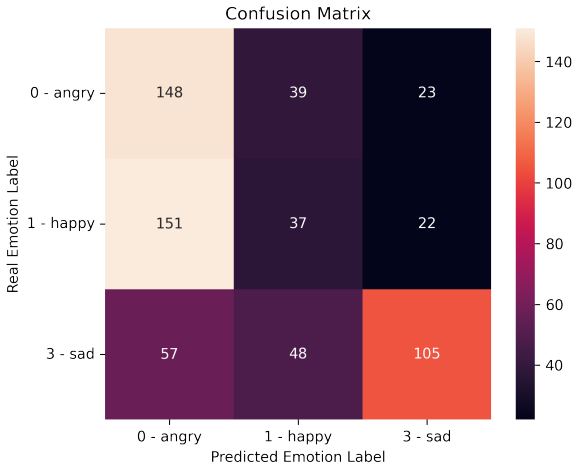
Table 1.9 displays the results obtained. The outcomes of the experiments illustrate the efficacy of employing transfer learning with pre-trained models for the task of SER. While the spectrogram image feature attained the highest average accuracy when utilizing the Resnet50 model, the mel spectrogram feature achieved the best overall performance across all models. Furthermore, the image of the mel spectrogram obtained the second-highest accuracy with the Resnet50 model and displayed a smaller standard deviation, indicating that it performed comparably well in all cross-validation folds.

Table 1.9: DL Classification Models Performance on IEMOCAP.

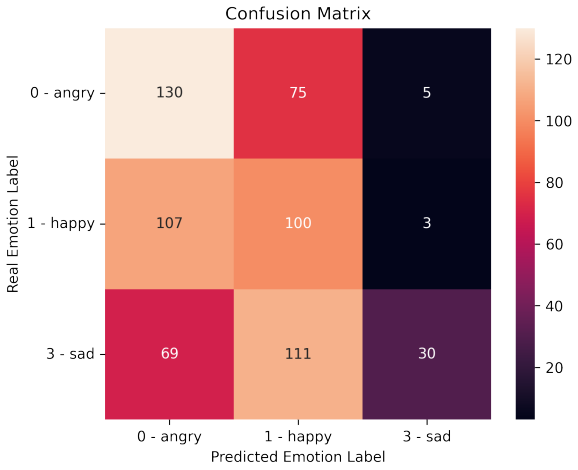
Feature	Model	Accuracy	Macro F1	Precision	Recall	MCC	Training Time
Spectrogram Image	Resnet50	58.24±2.20	58.97	59.38	59.00	0.436	1047.72
Mel Spectrogram Image	Resnet50	57.95±1.36	58.71	59.27	58.49	0.430	1133.66
MFCCs Image	Resnet50	56.59±0.45	57.29	58.59	56.67	0.410	1044.96
Mel Spectrogram Image	VGG16	55.07±2.23	55.82	56.77	55.29	0.389	1027.24
MFCCs Image	VGG16	54.73±1.47	55.51	56.32	55.14	0.386	1032.96
Spectrogram Image	VGG16	54.28±0.90	55.21	55.85	54.87	0.379	1147.05
Mel Spectrogram Image	Xception	53.10±1.42	53.84	54.27	53.68	0.364	1171.14
MFCCs Image	Xception	52.78±0.96	53.47	54.10	53.22	0.359	1181.58
Spectrogram Image	Xception	52.78±1.54	53.51	53.48	53.62	0.361	1190.07
Spectrogram	2D-CNN	50.12±0.91	50.04	52.98	49.65	0.320	3167.14
Mel Spectrogram	2D-CNN & LSTM	48.02±1.14	47.93	48.6	48.47	0.298	1427.02
MFCCs	2D-CNN	46.70±0.85	47.13	49.53	46.75	0.275	303.49
Spectrogram	2D-CNN & LSTM	46.01±1.77	47.09	47.37	46.87	0.269	5413.99
MFCCs	2D-CNN & LSTM	45.56±1.15	46.26	46.29	46.25	0.263	298.06
Mel Spectrogram	2D-CNN	32.51±1.13	21.34	20.38	30.26	0.102	1166.9

Table 1.10

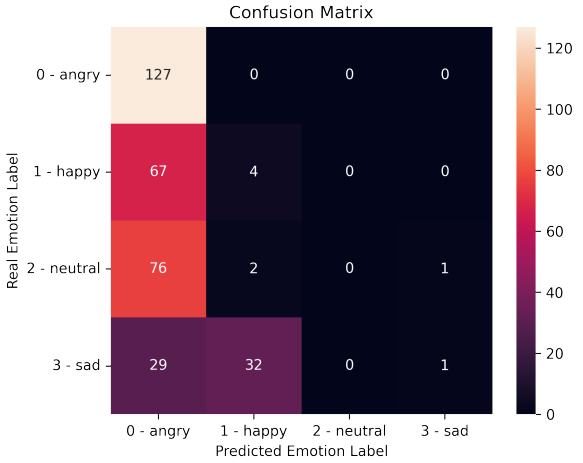
Dataset	Model	Accuracy	Macro F1	Precision	Recall	MCC	Time
eNTERFACE'05	Traditional	46.03	44.26	47.14	46.03	0.203	0.18
	Deep Learning	41.27	38.3	52.13	41.27	0.131	0.27
EMO-DB	Traditional	38.94	17.52	25.75	26.81	0.088	0.11
	Deep Learning	37.76	14.56	12.78	25.35	0.039	0.20
CREMA-D	Traditional	48.88	37.41	40.01	47.13	0.370	0.11
	Deep Learning	39.61	33.84	34.91	39.97	0.216	0.22



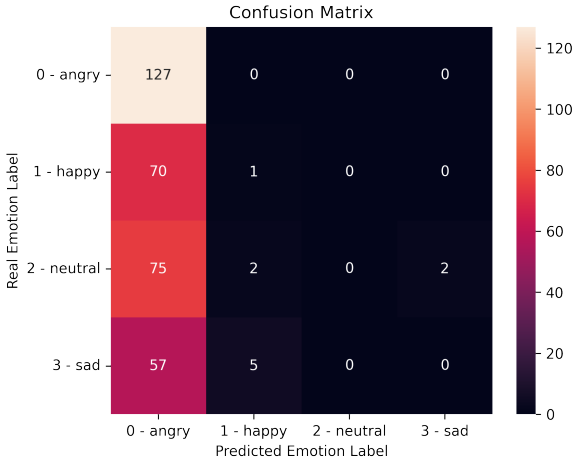
(a) eNTERFACE'05 Traditional Model Confusion Matrix.



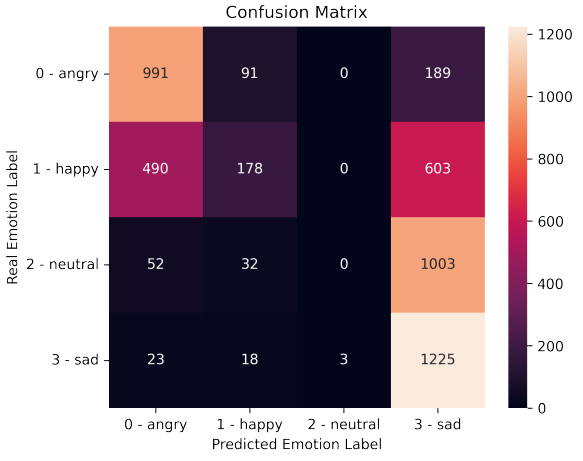
(b) eNTERFACE'05 DL Model Confusion Matrix.



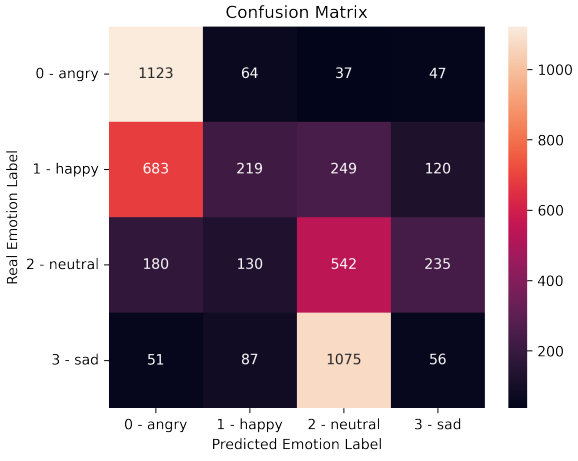
(c) EMO-DB Traditional Model Confusion Matrix.



(d) EMO-DB DL Model Confusion Matrix.



(e) CREMA-D Traditional Model Confusion Matrix.



(f) CREMA-D DL Model Confusion Matrix.

Figure 1.12: Final Models Confusion Matrices on the eNTERFACE'05, EMO-DB and CREMA-D Datasets.

CHAPTER 2

Data Visualization and Quality Analysis

Table 2.1: Random forest 5-fold cross-validation results with different classification labels

Classification Labels	Total Data	Accuracy	Macro F1	Precision	Recall	MCC.
Two Labels						
Angry, Sad	2187	92.04±1.18	92.04±1.18	92.05±1.12	92.04±1.18	0.84±0.02
Angry, Neutral	2811	85.63±1.18	84.42±1.35	86.18±1.15	83.5±1.4	0.7±0.03
Sad, Neutral	2792	78.08±0.86	76.34±1.08	77.3±0.87	75.8±1.16	0.53±0.02
Angry+Sad, Happy+Excited	2739	76.78±1.89	75.32±2.1	76.17±2.02	74.88±2.12	0.51±0.04
Sad, Happy+Excited	2720	83.93±1.66	83.27±1.73	83.22±1.76	83.33±1.72	0.67±0.03
Neutral, Happy+Excited	3344	72.52±0.75	72.32±0.83	72.85±0.65	72.37±0.79	0.45±0.01
Angry+Sad, Happy+Excited	3823	75.23±0.77	74.22±0.92	75.0±0.71	73.91±0.94	0.49±0.02
Angry+Sad, Neutral+Happy+Excited	5531	73.93±1.0	71.34±1.13	73.43±1.18	70.73±1.09	0.44±0.02
Three Labels						
Angry, Happy+Excited, Sad	3823	71.96±1.63	72.22±1.62	72.81±1.56	71.95±1.64	0.57±0.03
Angry, Happy+Excited, Neutral	4447	65.1±1.56	64.77±1.55	66.03±1.39	64.52±1.65	0.48±0.02
Sad, Happy+Excited, Neutral	4428	64.39±2.13	64.39±2.11	64.91±2.07	64.17±2.11	0.46±0.03
Angry, Sad, Neutral	3895	73.81±1.24	73.87±1.41	75.81±0.83	72.7±1.77	0.6±0.02
High Arousal, Neutral Arousal, Low Arousal	5531	68.27±1.48	65.43±1.89	65.73±1.93	65.19±1.91	0.49±0.02
High Valence, Neutral Valence, Low Valence	5531	61.0±1.11	60.1±1.25	60.53±1.06	60.12±1.22	0.41±0.02
Four Labels						
Angry, Sad, Neutral, Happy+Excited	5531	60.26±0.46	60.93±0.56	61.93±0.62	60.49±0.63	0.46±0.01

Table 2.2: Data duration analysis

	Duration							
Emotion	Count	Mean	Std	min	25%	50%	75%	max
angry	1103	4.51	3.00	0.76	2.43	3.61	5.66	26.77
excited	1041	4.78	3.46	0.58	2.34	3.82	6.21	34.13
happy	595	4.34	2.71	0.89	2.39	3.56	5.80	17.22
neutral	1708	3.90	2.58	0.73	2.10	3.13	4.91	20.29
sad	1084	5.49	4.04	0.76	2.68	4.14	7.01	31.91

Table 2.3: Juries arousal classification analysis

	Arousal							
Emotion	Count	Mean	Std	min	25%	50%	75%	max
angry	1103	3.63	0.67	1.5	3.0	3.5	4.0	5.0
excited	1041	3.57	0.60	2.0	3.0	3.5	4.0	5.0
happy	595	3.11	0.61	1.5	2.5	3.0	3.5	5.0
neutral	1708	2.72	0.54	1.0	2.5	2.6	3.0	5.0
sad	1084	2.56	0.62	1.0	2.0	2.5	3.0	4.5

Table 2.4: Juries valence classification analysis

	Valence							
Emotion	Count	Mean	Std	min	25%	50%	75%	max
angry	1103	1.90	0.52	1.0	1.5	2.0	2.0	4.0
excited	1041	3.94	0.62	1.5	3.5	4.0	4.5	5.5
happy	595	3.95	0.45	2.0	4.0	4.0	4.0	5.0
neutral	1708	2.97	0.51	1.5	2.5	3.0	3.0	5.0
sad	1084	2.25	0.58	1.0	2.0	2.0	2.5	4.0

Table 2.5: Juries dominance classification analysis

	Dominance							
Emotion	Count	Mean	Std	min	25%	50%	75%	max
angry	1103	3.94	0.64	1.0	3.5	4.0	4.5	5.0
excited	1041	3.40	0.76	1.0	3.0	3.5	4.0	5.0
happy	595	2.92	0.65	1.5	2.5	3.0	3.5	5.0
neutral	1708	2.83	0.60	0.5	2.5	3.0	3.0	4.5
sad	1084	2.82	0.81	1.0	2.3	3.0	3.5	5.0

Table 2.6: Juries dimensional emotion centroids classifications numeric visualization

	Centroids		
Emotion	Arousal	Valence	Dominance
Angry	3.63	1.90	3.94
Happy+Excited	3.40	3.94	3.23
Sad	2.56	2.25	2.82
Neutral	2.72	2.97	2.83

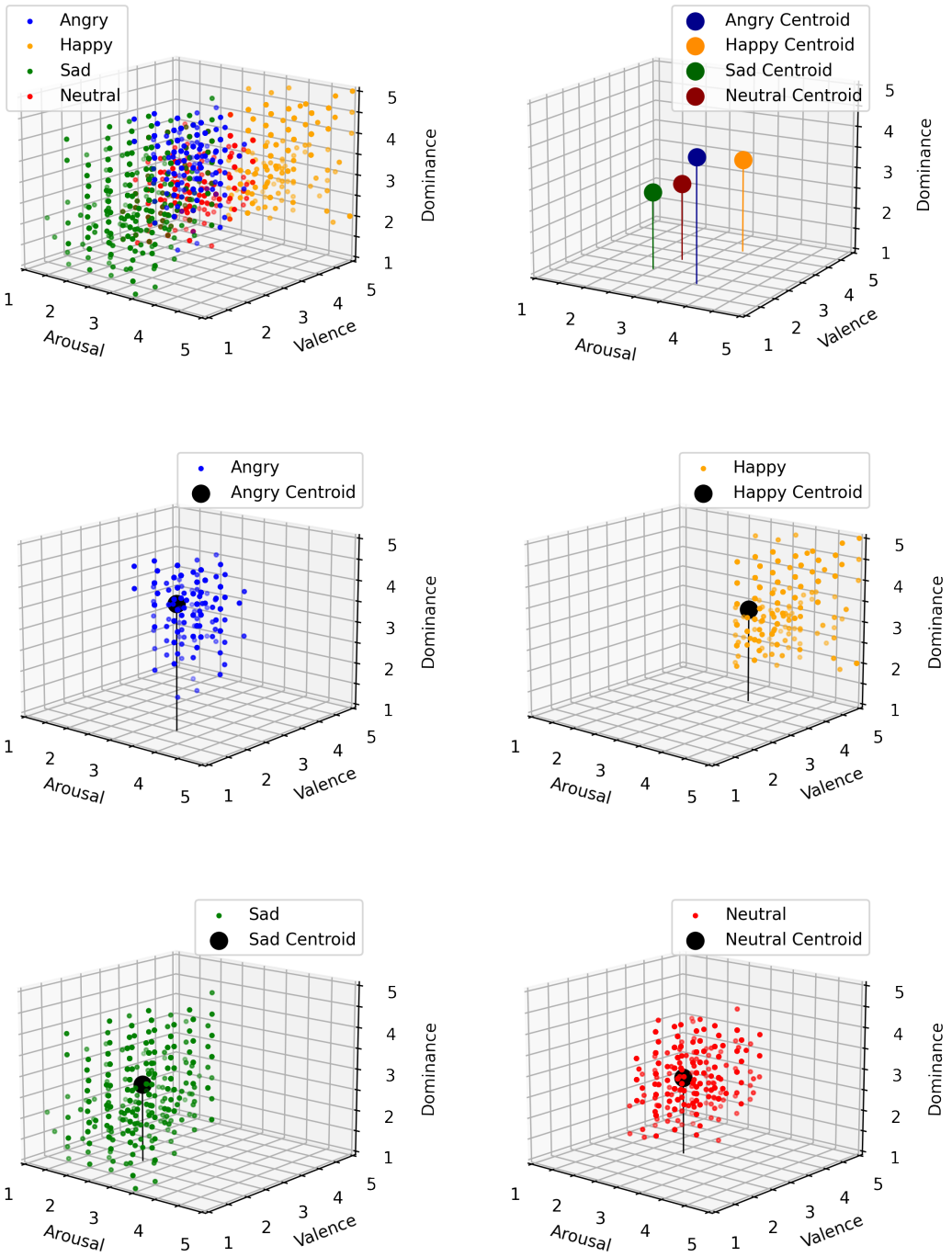


Figure 2.1: Juries dimensional emotion classifications 3D visualization

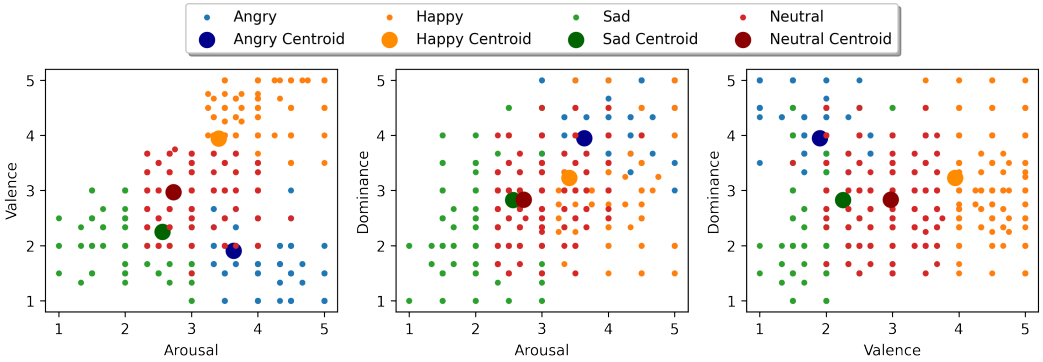


Figure 2.2: Juries dimensional emotion classifications 2D visualization

CITE CONFLICTS FROM THIS https://ieeexplore.ieee.org/document/9746930:

TODO: NAO MOSTRAR OS CONFLITOS RETIRADOS MAS SIM MOSTRAR OS RANGES MANTIDOS

Table 2.7: Conflicts between emotion's categories and primitives

	Conflicts		
Emotion Categories	Arousal	Valence	Dominance
Non-Strict Conflicts			
Angry	$[1, 2.5]$	$]3.5, 5]$	$[1, 2[$
Happy+Excited	$[1, 3[$	$[1, 3[$	$[1, 2[$
Sad	$[3.5, 5]$	$[3.5, 5]$	None
Neutral	$[1, 2[$	$]4, 5]$	$[1, 1.5[\cup]4.5, 5]$
Strict Conflicts			
Angry	$[1, 3]$	$[3, 5]$	$[1, 2[$
Happy+Excited	$[1, 3]$	$[1, 3]$	$[1, 2[$
Sad	$[4, 5]$	$[4, 5]$	None
Neutral	$[1, 2]$	$[4, 5]$	$[1, 2[\cup]4, 5]$

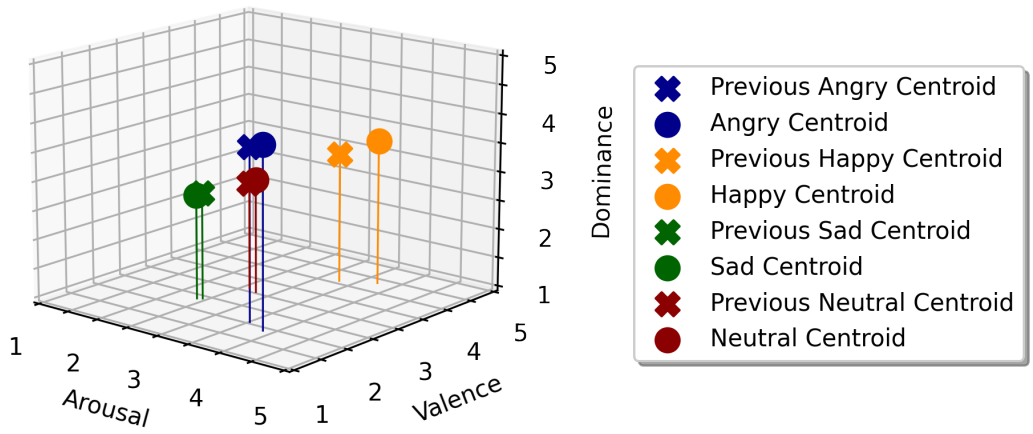


Figure 2.3: Data with and without conflicts between emotion's categories and primitives emotion centroids' 3D visualization

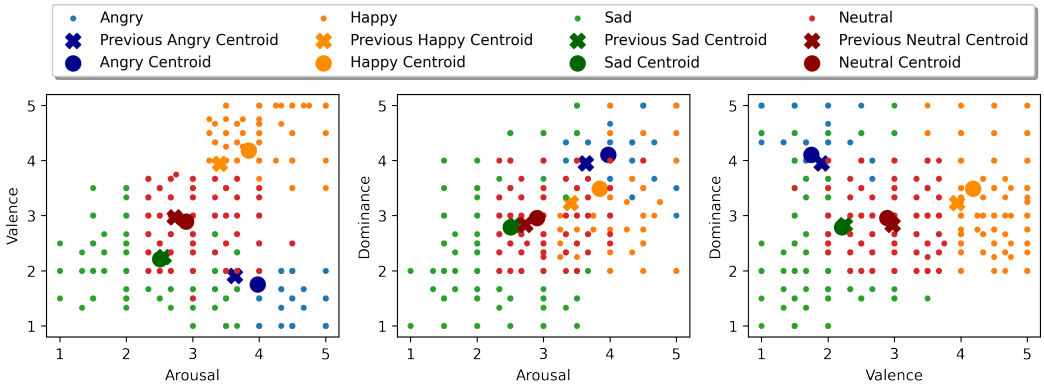


Figure 2.4: Data with and without conflicts between emotion's categories and primitives emotion centroids' 2D visualization

Table 2.8: Results obtained after eliminating emotions based on VAD conflicts with the categorical annotations.

Conflicts Removed	Total Data	Accuracy
None	5531	60.26+-0.46
Non-strict Arousal	4955	63.37+-1.09
Non-strict Valence	5382	60.46+-0.95
Non-strict Dominance	5491	60.5+-0.93
All Non-strict	4816	65.05+-0.85
Strict Arousal	4222	66.79+-1.85
Strict Valence	5128	62.07+-1.77
Strict Dominance	5432	60.53+-1.34
All Strict	3911	69.09+-1.17

TODO: UTILIZAR RANDOMS DADOS PARA CHEGAR AO MSM NUMERO DE DADOS

(VARIAS VEZES) PARA VER SE A RAZAO DE MELHORAR DADOS É A QUANTIDADE DE DAODS OU MESMO A MA ANOTACAO

CHAPTER 3

Speech Emotion Recognition on a video conference system

3.1 Audio Pipeline

3.1.1 Noise reduction

TODO: TESTAR DIFERENTES RESULTADOS DE PARAMETROS QUE MANTENHAM O SOM MINIMANTE ALTO SO PARA TER A CERTEZA QUE N VALE A PENA

https://pypi.org/project/noisereduce/

In real life, the noise present in the environment is captured along with the speech signal. This affects the recognition rate, hence some noise reduction techniques must be used to eliminate or reduce the noise. Minimum mean square error and log-spectral amplitude MMSE (LogMMSE) estimators are the most successfully applied methods for noise reduction.

3.1.2 Voice Activity Detection

https://thegradient.pub/one-voice-detector-to-rule-them-all/

https://github.com/snakers4/silero-vad

https://github.com/wiseman/py-webrtcvad

The detection of the presence of voiced speech among various unvoiced speech and silence is called endpoint detection, speech detection, or voice activity detection.

The performance of the detection algorithm could affect the accuracy of the system. The goal is to detect silent and noisy frames that are potentially irrelevant in terms of SER, this will also decrease the complexity and increase the accuracy of the model. The most widely used methods for voice activity detection are zero-crossing rate, short-time energy, and auto-correlation method.

Zero crossing rate is the rate at which a signal changes its sign from positive to negative or vice versa within a given time frame.

The voiced speech has high energy due to its periodicity, while low energy is observed in the unvoiced speech.

The auto-correlation method provides a measure of similarity between a signal and itself as a function of delay. It is used to find repeating patterns. Because of its periodic nature, voiced signals can be detected using the auto-correlation method.

3.1.3 Speech Segmentation

Speech segmentation, also known as framing, is the process in which continuous speech signals are partitioned into segments.

As mentioned previously, emotions are usually short-lived, and the speech remains invariant for a brief period. By segmenting this data, it is possible to obtain local features of emotions, hence, frames with a short range of length are suitable for classifiers while maintaining the emotional information in a continuous speech.

3.2 Results and Discussion

Bibliography

[1] R. . E. . Kaliouby. «This app knows how you feel – from the look on your face». (Jun. 15, 2015), [Online]. Available: https://www.ted.com/talks/rana_el_kaliouby_this_app_knows_how_you_feel_from_the_look_on_your_face (visited on 01/05/2023).

[2] S. B. Daily, M. T. James, D. Cherry, *et al.*, «Affective computing: Historical foundations, current applications, and future trends», in *Emotions and Affect in Human Factors and Human-Computer Interaction*, Elsevier, 2017, pp. 213–231. DOI: 10.1016/b978-0-12-801851-4.00009-4. [Online]. Available: https://doi.org/10.1016/b978-0-12-801851-4.00009-4.

[3] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, «A review of affective computing: From unimodal analysis to multimodal fusion», *Information Fusion*, vol. 37, pp. 98–125, 2017, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.02.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253517300738.

[4] H. Ai, D. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare, «Using system and user performance features to improve emotion detection in spoken tutoring dialogs», Jan. 2006.

[5] L. Devillers and L. Vidrascu, «Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs.», Jan. 2006.

[6] F. Burkhardt, M. van Ballegooy, and R. Englert, «An emotion-aware voice portal», Jan. 2005.

[7] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, «Detecting anger in automated voice portal dialogs.», Jan. 2006.

[8] T. Kanda, K. Iwase, M. Shiomi, and H. Ishiguro, «A tension-moderating mechanism for promoting speech-based human-robot interaction», in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2005. DOI: 10.1109/iros.2005.1545035. [Online]. Available: https://doi.org/10.1109/iros.2005.1545035.

[9] J. A. Balazs and J. D. Velásquez, «Opinion mining and information fusion: A survey», *Information Fusion*, vol. 27, pp. 95–110, 2016, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2015.06.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253515000536.

[10] L. Deng, *Dynamic Speech Models*. Springer International Publishing, 2006. DOI: 10.1007/978-3-031-02555-6. [Online]. Available: https://doi.org/10.1007/978-3-031-02555-6.

[11] A. . Hagerty and A. . Albert, *AI is increasingly being used to identify emotions – here's what's at stake*, Apr. 2021. [Online]. Available: https://theconversation.com/ai-is-increasingly-being-used-to-identify-emotions-heres-whats-at-stake-158809.

[12] E. Hudlicka, «Computational modeling of cognition–emotion interactions: Theoretical and practical relevance for behavioral healthcare», in *Emotions and Affect in Human Factors and Human-Computer Interaction*, Elsevier, 2017, pp. 383–436. DOI: 10.1016/b978-0-12-801851-4.00016-1. [Online]. Available: https://doi.org/10.1016/b978-0-12-801851-4.00016-1.

[13] V. Shuman and K. R. Scherer, «Emotions, psychological structure of», in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2015, pp. 526–533. DOI: 10.1016/b978-0-08-097086-8.25007-1. [Online]. Available: https://doi.org/10.1016/b978-0-08-097086-8.25007-1.

[14] X. Jin and Z. Wang, «An emotion space model for recognition of emotions in spoken chinese», in *Affective Computing and Intelligent Interaction*, J. Tao, T. Tan, and R. W. Picard, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 397–402, ISBN: 978-3-540-32273-3.

[15] O. Mitruț, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, «Emotion classification based on biophysical signals and machine learning techniques», *Symmetry*, vol. 12, p. 21, Dec. 2019. DOI: 10.3390/sym12010021.

[16] J. A. Russell and A. Mehrabian, «Evidence for a three-factor theory of emotions», *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 1977, ISSN: 0092-6566. DOI: https://doi.org/10.1016/0092-6566(77)90037-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/009265667790037X.

[17] K. R. Scherer, «A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology», in *Interspeech*, 2000.

[18] M. Slaney and G. McRoberts, «Baby ears: A recognition system for affective vocalizations», in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, IEEE. DOI: 10.1109/icassp.1998.675432. [Online]. Available: https://doi.org/10.1109/icassp.1998.675432.

[19] R. Rajoo and C. C. Aun, «Influences of languages in speech emotion recognition: A comparative study using malay, english and mandarin languages», in *2016 IEEE Symposium on Computer Applications &amp Industrial Electronics (ISCAIE)*, IEEE, May 2016. DOI: 10.1109/iscaie.2016.7575033. [Online]. Available: https://doi.org/10.1109/iscaie.2016.7575033.

[20] T. Vogt and E. Andre, «Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition», in *2005 IEEE International Conference on Multimedia and Expo*, IEEE. DOI: 10.1109/icme.2005.1521463. [Online]. Available: https://doi.org/10.1109/icme.2005.1521463.

[21] J. Wilting, E. Krahmer, and M. Swerts, «Real vs. acted emotional speech», English, in *Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006)*, ISCA, 2006.

[22] C.-H. Wu, J.-C. Lin, and W.-L. Wei, «Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies», *APSIPA Transactions on Signal and Information Processing*, vol. 3, no. 1, 2014. DOI: 10.1017/atsip.2014.11. [Online]. Available: https://doi.org/10.1017/atsip.2014.11.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, «A database of german emotional speech», in *Interspeech 2005*, ISCA, Sep. 2005. DOI: 10.21437/interspeech.2005-446. [Online]. Available: https://doi.org/10.21437/interspeech.2005-446.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas, «The eNTERFACE&ampamp#14605 audio-visual emotion database», in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, IEEE, 2006. DOI: 10.1109/icdew.2006.145. [Online]. Available: https://doi.org/10.1109/icdew.2006.145.

[25] C. Busso, M. Bulut, C.-C. Lee, *et al.*, «IEMOCAP: Interactive emotional dyadic motion capture database», *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Nov. 2008. DOI: 10.1007/s10579-008-9076-6. [Online]. Available: https://doi.org/10.1007/s10579-008-9076-6.

[26] M. Kosti, T. Pappas, and G. Potamianos, *Multimodal opinion and sentiment (moud) dataset*, 2013. [Online]. Available: http://multicomp.cs.cmu.edu/resources/moud-dataset/.

[27] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, «CREMA-d: Crowd-sourced emotional multimodal actors dataset», *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, Oct. 2014. DOI: 10.1109/taffc.2014.2336244. [Online]. Available: https://doi.org/10.1109/taffc.2014.2336244.

[28] Zadeh, A. and Morency, L.-P. and Yannakakis, G. and Poria, S. and Cambria, E. and Howard, N. and Pappas, T. and Morency, L. P., *Cmu multimodal opinion sentiment and emotion intensity (cmu-mosi)*, 2017. [Online]. Available: http://multicomp.cs.cmu.edu/resources/moud-dataset/.

[29] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, «MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception», *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, Jan. 2017. DOI: 10.1109/taffc.2016.2515617. [Online]. Available: https://doi.org/10.1109/taffc.2016.2515617.

[30] Zadeh, A. and Poria, S. and Cambria, E. and Howard, N. and Pappas, T. and Morency, L.-P., *Cmu multimodal opinion sentiment and emotion intensity (cmu-mosei)*, 2018. [Online]. Available: http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/.

[31] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, *Meld: A multimodal multi-party dataset for emotion recognition in conversations*, 2018. DOI: 10.48550/ARXIV.1810.02508. [Online]. Available: https://arxiv.org/abs/1810.02508.

[32] S. R. Livingstone and F. A. Russo, «The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english», *PLOS ONE*, vol. 13, no. 5, pp. 1–35, May 2018. DOI: 10.1371/journal.pone.0196391. [Online]. Available: https://doi.org/10.1371/journal.pone.0196391.

[33] R. Lotfian and C. Busso, «Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings», *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, Oct. 2019. DOI: 10.1109/taffc.2017.2736999. [Online]. Available: https://doi.org/10.1109/taffc.2017.2736999.

[34] M. K. Pichora-Fuller and K. Dupuis, *Toronto emotional speech set (tess)*, 2020. DOI: 10.5683/SP2/E8H2MF. [Online]. Available: https://borealisdata.ca/citation?persistentId=doi:10.5683/SP2/E8H2MF.

[35] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, «Deep learning approaches for speech emotion recognition: State of the art and research challenges», *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23 745–23 812, Jan. 2021. DOI: 10.1007/s11042-020-09874-7. [Online]. Available: https://doi.org/10.1007/s11042-020-09874-7.

[36] S. Narayanan and P. G. Georgiou, «Behavioral signal processing: Deriving human behavioral informatics from speech and language», *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013. DOI: 10.1109/jproc.2012.2236291. [Online]. Available: https://doi.org/10.1109/jproc.2012.2236291.

[37] B. Schuller, «Voice and speech analysis in search of states and traits», in *Computer Analysis of Human Behavior*, Springer London, 2011, pp. 227–253. DOI: 10.1007/978-0-85729-994-9_9. [Online]. Available: https://doi.org/10.1007/978-0-85729-994-9_9.

[38] X. A. Rathina, «Basic analysis on prosodic features in emotional speech», *International Journal of Computer Science, Engineering and Applications*, vol. 2, no. 4, pp. 99–107, Aug. 2012. DOI: 10.5121/ijcsea.2012.2410. [Online]. Available: https://doi.org/10.5121/ijcsea.2012.2410.

[39] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, «A survey of affect recognition methods: Audio, visual, and spontaneous expressions», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009. DOI: 10.1109/tpami.2008.52. [Online]. Available: https://doi.org/10.1109/tpami.2008.52.

[40] B. Schuller, G. Rigoll, and M. Lang, «Hidden markov model-based speech emotion recognition», in *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, vol. 1, Jul. 2003, pp. I–401. DOI: 10.1109/ICME.2003.1220939.

[41] H. Zhao, N. Ye, and R. Wang, «A survey on automatic emotion recognition using audio big data and deep learning architectures», in *2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS)*, IEEE, May 2018. DOI: 10.1109/bds/hpsc/ids18.2018.00039. [Online]. Available: https://doi.org/10.1109/bds/hpsc/ids18.2018.00039.

[42] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, «Speech emotion recognition using deep learning techniques: A review», *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019. DOI: 10.1109/access.2019.2936124. [Online]. Available: https://doi.org/10.1109/access.2019.2936124.

[43] B. Schuller, G. Rigoll, and M. Lang, «Hidden markov model-based speech emotion recognition», in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 2, 2003, pp. II–1. DOI: 10.1109/ICASSP.2003.1202279.

[44] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, «Emotion recognition from speech using global and local prosodic features», *International Journal of Speech Technology*, vol. 16, no. 2, pp. 143–160, Aug. 2012. DOI: 10.1007/s10772-012-9172-2. [Online]. Available: https://doi.org/10.1007/s10772-012-9172-2.

[45] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, «Automatic emotion recognition using prosodic parameters», in *Interspeech 2005*, ISCA, Sep. 2005. DOI: 10.21437/interspeech.2005-324. [Online]. Available: https://doi.org/10.21437/interspeech.2005-324.

[46] G. Gosztolya, «Conflict intensity estimation from speech using greedy forward-backward feature selection», in *Interspeech 2015*, ISCA, Sep. 2015. DOI: 10.21437/interspeech.2015-332. [Online]. Available: https://doi.org/10.21437/interspeech.2015-332.

[47] B. Schuller, «Recognizing affect from linguistic information in 3d continuous space», *IEEE Transactions on Affective Computing*, vol. 2, no. 4, pp. 192–205, Oct. 2011. DOI: 10.1109/t-affc.2011.17. [Online]. Available: https://doi.org/10.1109/t-affc.2011.17.

[48] F. Eyben, K. R. Scherer, B. W. Schuller, *et al.*, «The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing», *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, Apr. 2016. DOI: 10.1109/taffc.2015.2457417. [Online]. Available: https://doi.org/10.1109/taffc.2015.2457417.

[49] L. Tarantino, P. N. Garner, and A. Lazaridis, «Self-attention for speech emotion recognition», in *Interspeech 2019*, ISCA, Sep. 2019. DOI: 10.21437/interspeech.2019-2822. [Online]. Available: https://doi.org/10.21437/interspeech.2019-2822.

[50] S. Kuchibhotla, H. D. Vankayalapati, R. S. Vaddi, and K. R. Anne, «A comparative analysis of classifiers in emotion recognition through acoustic features», *International Journal of Speech Technology*, vol. 17, no. 4, pp. 401–408, Jun. 2014. DOI: 10.1007/s10772-014-9239-3. [Online]. Available: https://doi.org/10.1007/s10772-014-9239-3.

[51] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, «Spoken emotion recognition using hierarchical classifiers», *Computer Speech &amp Language*, vol. 25, no. 3, pp. 556–570, Jul. 2011. DOI: 10.1016/j.csl.2010.10.001. [Online]. Available: https://doi.org/10.1016/j.csl.2010.10.001.

[52] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, «Emotion recognition using a hierarchical binary decision tree approach», *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, Nov. 2011. DOI: 10.1016/j.specom.2011.06.004. [Online]. Available: https://doi.org/10.1016/j.specom.2011.06.004.

[53] G. Sahu, *Multimodal speech emotion recognition and ambiguity resolution*, 2019. DOI: 10.48550/ARXIV.1904.06022. [Online]. Available: https://arxiv.org/abs/1904.06022.

[54] J. Huang, B. Chen, B. Yao, and W. He, «Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network», *IEEE Access*, vol. 7, pp. 92 871–92 880, 2019. DOI: 10.1109/ACCESS.2019.2928017.

[55] G. Zhou, Y. Chen, and C. Chien, «On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks», *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, Aug. 2022. DOI: 10.1186/s12911-022-01942-2. [Online]. Available: https://doi.org/10.1186/s12911-022-01942-2.

[56] D. Issa, M. F. Demirci, and A. Yazici, «Speech emotion recognition with deep convolutional neural networks», *Biomedical Signal Processing and Control*, vol. 59, p. 101 894, May 2020. DOI: 10.1016/j.bspc.2020.101894. [Online]. Available: https://doi.org/10.1016/j.bspc.2020.101894.

[57] M. T. García-Ordás, H. Alaiz-Moretón, J. A. Benítez-Andrades, I. García-Rodríguez, O. García-Olalla, and C. Benavides, «Sentiment analysis in non-fixed length audios using a fully convolutional neural network», *Biomedical Signal Processing and Control*, vol. 69, p. 102 946, 2021, ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.102946. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809421005437.

[58] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, «Emotion Recognition from Variable-Length Speech Segments Using Deep Learning on Spectrograms», in *Proc. Interspeech 2018*, 2018, pp. 3683–3687. DOI: 10.21437/Interspeech.2018-2228.

[59] Z. Zhao, Z. Bao, Y. Zhao, *et al.*, «Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition», *IEEE Access*, vol. 7, pp. 97 515–97 525, 2019. DOI: 10.1109/access.2019.2928625. [Online]. Available: https://doi.org/10.1109/access.2019.2928625.

[60] Z. Luo, H. Xu, and F. Chen, *Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network*, Dec. 2018. DOI: 10.29007/7mhj. [Online]. Available: https://doi.org/10.29007/7mhj.

[61] M. Chen, X. He, J. Yang, and H. Zhang, «3-d convolutional recurrent neural networks with attention model for speech emotion recognition», *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018. DOI: 10.1109/LSP.2018.2860246.

[62] A. Muppidi and M. Radfar, «Speech emotion recognition using quaternion convolutional neural networks», in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021. DOI: 10.1109/icassp39728.2021.9414248. [Online]. Available: https://doi.org/10.1109/icassp39728.2021.9414248.

[63] K. Palanisamy, D. Singhania, and A. Yao, *Rethinking cnn models for audio classification*, 2020. DOI: 10.48550/ARXIV.2007.11154. [Online]. Available: https://arxiv.org/abs/2007.11154.

[64] S. Zhang, S. Zhang, T. Huang, and W. Gao, «Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching», *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018. DOI: 10.1109/TMM.2017.2766843.

[65] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz, «Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review», *Sensors*, vol. 21, no. 15, p. 5015, Jul. 2021. DOI: 10.3390/s21155015. [Online]. Available: https://doi.org/10.3390/s21155015.

[66] J. Bhaskar, K. Sruthi, and P. Nedungadi, «Hybrid approach for emotion classification of audio conversation based on text and speech mining», *Procedia Computer Science*, vol. 46, pp. 635–643, 2015, Proceedings of the International Conference on Information and Communication Technologies, ICICT 2014, 3-5 December 2014 at Bolgatty Palace & Island Resort, Kochi, India, ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.02.112. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915001763.

[67] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, *Deep learning based emotion recognition system using speech features and transcriptions*, 2019. DOI: 10.48550/ARXIV.1906.05681. [Online]. Available: https://arxiv.org/abs/1906.05681.

[68] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, «Speech sentiment analysis via pre-trained features from end-to-end ASR models», in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2020. DOI: 10.1109/icassp40776.2020.9052937. [Online]. Available: https://doi.org/10.1109/icassp40776.2020.9052937.

[69] A. Handa, R. Agarwal, and N. Kohli, «Audio-visual emotion recognition system using multi-modal features», *International Journal of Cognitive Informatics and Natural Intelligence*, vol. 15, no. 4, pp. 1–14, Oct. 2021. DOI: 10.4018/ijcini.20211001.oa34. [Online]. Available: https://doi.org/10.4018/ijcini.20211001.oa34.

[70] X. Yan, H. Xue, S. Jiang, and Z. Liu, «Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling», *Applied Artificial Intelligence*, vol. 36, no. 1, Nov. 2021. DOI: 10.1080/08839514.2021.2000688. [Online]. Available: https://doi.org/10.1080/08839514.2021.2000688.

[71] P. Buitelaar, I. D. Wood, S. Negi, *et al.*, «MixedEmotions: An open-source toolbox for multimodal emotion analysis», *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2454–2465, Sep. 2018. DOI: 10.1109/tmm.2018.2798287. [Online]. Available: https://doi.org/10.1109/tmm.2018.2798287.

[72] «Ibm watson». (), [Online]. Available: https://www.ibm.com/watson (visited on 01/03/2023).

[73] Bitext. We help AI understand humans. – chatbots that work. «Bitext. we help ai understand humans. - chatbots that work - synthetic data». (Sep. 16, 2022), [Online]. Available: https://www.bitext.com/ (visited on 01/03/2023).

[74] U. Krcadinac, J. Jovanovic, V. Devedzic, and P. Pasquier, «Textual affect communication and evocation using abstract generative visuals», *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 3, pp. 370–379, Jun. 2016. DOI: 10.1109/thms.2015.2504081. [Online]. Available: https://doi.org/10.1109/thms.2015.2504081.

[75] «Cognitive services—apis for ai solutions». (), [Online]. Available: https://azure.microsoft.com/en-us/products/cognitive-services/ (visited on 01/03/2023).

[76] S. . Kristensen. «Imotions - powering human insights». (Dec. 26, 2022), [Online]. Available: https://imotions.com/ (visited on 01/03/2023).

[77] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, and R. e. Kaliouby, «Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit», in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '16, San Jose, California, USA: Association for Computing Machinery, 2016, pp. 3723–3726, ISBN: 9781450340823. DOI: 10.1145/2851581.2890247. [Online]. Available: https://doi.org/10.1145/2851581.2890247.

[78] «Emovu by eyeris». (), [Online]. Available: https://www.emovu.com/ (visited on 01/03/2023).

[79] «Human behaviour ai technology». (), [Online]. Available: https://www.nviso.ai/en/technology (visited on 01/03/2023).

[80] «Skybiometry | cloud based biometrics api as a service». (Jan. 12, 2022), [Online]. Available: https://skybiometry.com/ (visited on 01/03/2023).

[81] «Technology». (Jul. 20, 2022), [Online]. Available: https://www.audeering.com/technology/ (visited on 01/03/2023).

[82] «Emotion recognition by voice by powerful ai voice algorithms - good vibrations company». (), [Online]. Available: https://goodvibrations.nl/ (visited on 01/03/2023).

[83] «Vokaturi - eyes on speech communication». (), [Online]. Available: https://vokaturi.com/ (visited on 01/03/2023).

[84] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, «The social signal interpretation (ssi) framework: Multimodal signal processing and recognition in real-time», in *Proceedings of the 21st ACM international conference on Multimedia*, ser. MM '13, Barcelona, Spain: ACM, 2013, pp. 831–834, ISBN: 978-1-4503-2404-5. DOI: 10.1145/2502081.2502223. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502223.

[85] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[86] «Python package index - pypi». (), [Online]. Available: https://pypi.org/ (visited on 03/28/2021).

[87] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, «Array programming with NumPy», *Nature*, vol. 585, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.

[88] W. McKinney *et al.*, «Data structures for statistical computing in python», in *Proceedings of the 9th Python in Science Conference*, Austin, TX, vol. 445, 2010, pp. 51–56.

[89] J. D. Hunter, «Matplotlib: A 2d graphics environment», *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[90] M. Waskom, O. Botvinnik, D. O'Kane, *et al.*, *Mwaskom/seaborn: V0.8.1 (september 2017)*, version v0.8.1, Sep. 2017. DOI: 10.5281/zenodo.883859. [Online]. Available: https://doi.org/10.5281/zenodo.883859.

[91] A. Gulli and S. Pal, *Deep learning with Keras*. Packt Publishing Ltd, 2017.

[92] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, «Scikit-learn: Machine learning in python», *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[93] B. McFee, A. Metsai, M. McVicar, *et al.*, *Librosa/librosa: 0.9.2*, 2022. DOI: 10.5281/ZENODO.6759664. [Online]. Available: https://zenodo.org/record/6759664.

[94] F. Eyben, M. Wöllmer, and B. Schuller, «Opensmile», in *Proceedings of the international conference on Multimedia - MM '10*, ACM Press, 2010. DOI: 10.1145/1873951.1874246. [Online]. Available: https://doi.org/10.1145/1873951.1874246.

[95] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, «COVAREP &amp#x2014 a collaborative voice analysis repository for speech technologies», in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2014. DOI: 10.1109/icassp.2014.6853739. [Online]. Available: https://doi.org/10.1109/icassp.2014.6853739.

[96] A. Malek, S. Borzì, and C. H. Nielsen, *Superkogito/spafe: V0.2.0*, en, 2022. DOI: 10.5281/ZENODO.6824667. [Online]. Available: https://zenodo.org/record/6824667.

[97] A. Malek, *Pydiogment/pydiogment: 0.1.0*, version 0.1.2, Apr. 2020. [Online]. Available: https://github.com/SuperKogito/spafe.

[98] T. Sainburg, *Timsainb/noisereduce: V1.0*, version db94fe2, Jun. 2019. DOI: 10.5281/zenodo.3243139. [Online]. Available: https://doi.org/10.5281/zenodo.3243139.

[99] T. Sainburg, M. Thielk, and T. Q. Gentner, «Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires», *PLoS computational biology*, vol. 16, no. 10, e1008228, 2020.

[100] S. Team, *Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier*, https://github.com/snakers4/silero-vad, 2022.

[101] J. Wiseman, *Python interface to the webrtc voice activity detector*, 2021. [Online]. Available: https://github.com/wiseman/py-webrtcvad.

[102] M. B. Akçay and K. Oğuz, «Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers», *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020. DOI: 10.1016/j.specom.2019.12.001. [Online]. Available: https://doi.org/10.1016/j.specom.2019.12.001.

[103] J. Pohjalainen, F. F. Ringeval, Z. Zhang, and B. Schuller, «Spectral and cepstral audio noise reduction techniques in speech emotion recognition», in *Proceedings of the 24th ACM international conference on Multimedia*, ACM, Oct. 2016. DOI: 10.1145/2964284.2967306. [Online]. Available: https://doi.org/10.1145/2964284.2967306.

[104] M. Milling, A. Baird, K. D. Bartl-Pokorny, *et al.*, «Evaluating the impact of voice activity detection on speech emotion recognition for autistic children», *Frontiers in Computer Science*, vol. 4, Feb. 2022. DOI: 10.3389/fcomp.2022.837269. [Online]. Available: https://doi.org/10.3389/fcomp.2022.837269.