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

Contents i

List of Figures iii

List of Tables v

Glossary vii

1 Speech Emotion Recognition (SER) Development 1

1.1 Datasets . . . 1

1.1.1 eNTERFACE'05 . . . 1

1.1.2 Interactive Emotional Dyadic Motion Capture (IEMOCAP) . . . 2

1.2 Audio Preprocessing . . . 3

1.2.1 Noise Reduction . . . 3

1.2.2 Audio Trim . . . 3

1.3 Traditional Feature-Based Classifier . . . 4

1.3.1 Feature Extraction . . . 4

1.3.2 Feature Analysis . . . 4

1.3.3 Feature Selection . . . 7

1.3.4 Conclusion . . . 11

1.4 Deep Learning Classifiers . . . 12

1.4.1 Raw Audio as a Feature . . . 12

1.4.2 MFCC as a Feature . . . 12

1.4.3 Mel-Spectrogram as a Feature . . . 12

1.5 Classifiers Discussion (pros & cons) . . . 12

Bibliography 13

List of Figures

1.1 Zero crossing rate wave plot annotated with spikes. 5
1.2 Bar plots mean for metrics used on the mel-scaled spectrogram feature 5
1.3 Zero crossing rate wave plot with a surrounding area of five male subjects for the same utterance with the anger emotion. 6
1.4 Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions. 6
1.5 Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions 7
1.6 Zero crossing rate mean values box plot for all emotions and different subjects 7
1.7 Audio Features' Pearson Correlation Matrices Before and After High Correlation Elimination. . . 8
1.8 Sequential Feature Selection with Backward Propagation using the Mean Accuracy as the Selection Method. 10

List of Tables

1.1 eNTERFACE'05 subjects nationalities 1
1.2 Extracted Audio Features and Statistical Functions Applied to Them 4
1.3 Performance of various classifiers in 5-fold cross-validation using the 97 features obtained after high correlation elimination. 9
1.4 Selected features. 10
1.5 Predictions' Evaluations Metrics with Different Strategies and Input Data 11

Glossary

SER Speech Emotion Recognition

MFCC Mel-frequency Cepstral Coefficients

VAD Valence-Arousal-Dominance

IEMOCAP Interactive Emotional Dyadic Motion Capture

CHAPTER 1

SER Development

TODO: mention recorded sample rates, display some data visualization graphics maybe

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 allowed us to explore and select the best features for our model.

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.

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:

· Anger · Happiness · Excitement · Sadness · Frustration · Fear · Surprise · Other · Neutral

In addition to the discrete emotions, it was also 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. The multimodal data, annotated using both discrete and dimensional models, allows us to perform a wide range of investigation. Researchers have also noted the high quality of the audiovisual data in this dataset, and it is frequently used in the literature for evaluating emotion recognition models. This enables us to make well-founded comparisons of our own developed models.

Overall, this second dataset is a well-suited resource for our study, as it allowed for a broad range of analysis, a comprehensive evaluation of our models' performance and facilitated benchmarking against existing models in the field.

1.2 Audio Preprocessing

To prepare the collected audio data for use in machine learning models, audio preprocessing techniques are required. In this section, we will explore important aspects of audio preprocessing: noise reduction, feature extraction, analysis, and selection.

1.2.1 Noise Reduction

Spectral Gating: Common strategy for denoising music by gating the signal only on high level sounds.

Noisereduce is a noise reduction algorithm in python that reduces noise in time-domain signals like speech, bioacoustics, and physiological signals. It relies on a method called "spectral gating" which is a form of Noise Gate. It works by computing a spectrogram of a signal (and optionally a noise signal) and estimating a noise threshold (or gate) for each frequency band of that signal/noise. That threshold is used to compute a mask, which gates noise below the frequency-varying threshold.

Non-stationary Noise Reduction The non-stationary noise reduction algorithm is an extension of the stationary noise reduction algorithm, but allowing the noise gate to change over time. When you know the timescale that your signal occurs on (e.g. a bird call can be a few hundred milliseconds), you can set your noise threshold based on the assumption that events occuring on longer timescales are noise. This algorithm was motivated by a recent method in bioacoustics called Per-Channel Energy Normalization. Steps of the Non-stationary Noise Reduction algorithm A spectrogram is calculated over the signal A time-smoothed version of the spectrogram is computed using an IIR filter aplied forward and backward on each frequency channel. A mask is computed based on that time-smoothed spectrogram The mask is smoothed with a filter over frequency and time The mask is appled to the spectrogram of the signal, and is inverted

$nr_x = nr.reduce_noise(y = x, sr = sr, n_fft = 2048, hop_length = 512, prop_decrease = .75, time_constant_s = 1)$

1.2.2 Audio Trim

3. Trim silence in the beginning and end. 30 decibels and lower are considered as silence

1.3 Traditional Feature-Based Classifier

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, which were subsequently processed using statistical metrics. The extracted features and associated metrics are summarized in Table 1.2, having in total, extracted 327 features.

Table 1.2: 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 1.1, 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. Consequently, this metric was also tested and applied to every other audio feature.

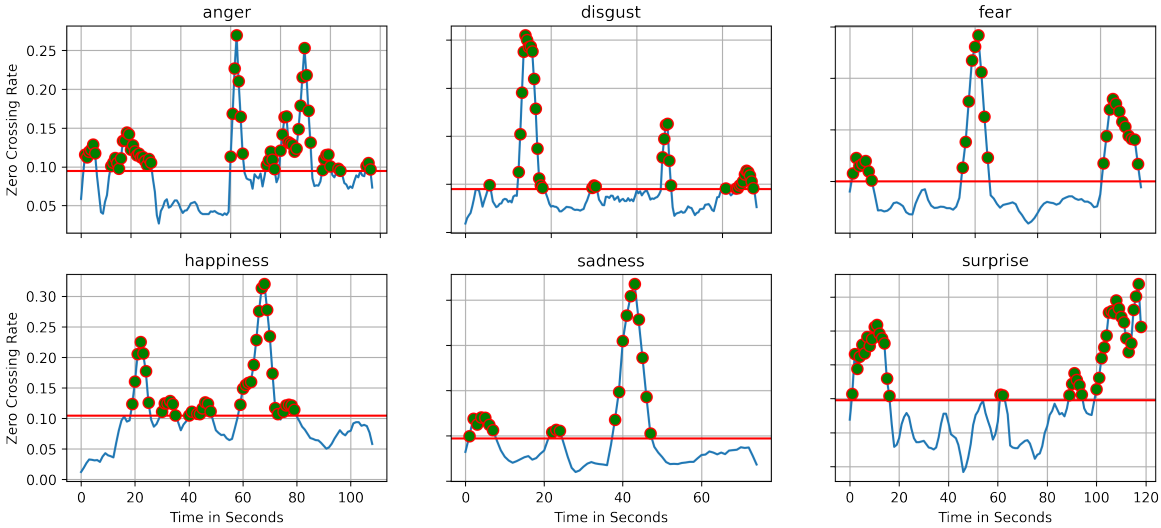


Figure 1.1: Zero crossing rate wave plot annotated with spikes.

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 1.2 shows clear differences in the mean values for some metrics used on the Mel Spectrogram.

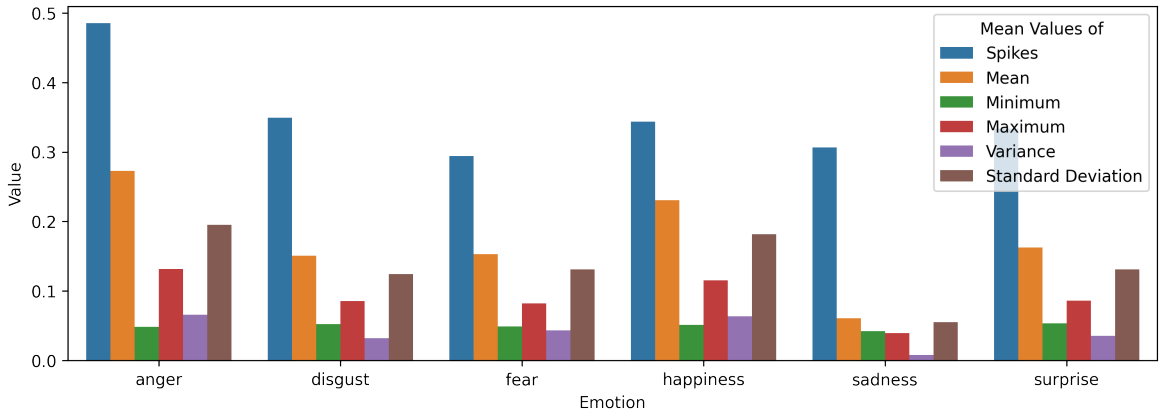


Figure 1.2: Bar plots mean for metrics used on the mel-scaled spectrogram feature

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.

The figure 1.3 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.

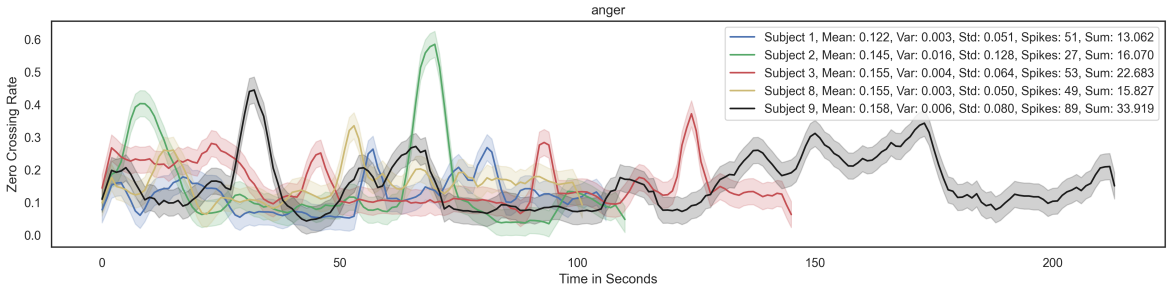


Figure 1.3: Zero crossing rate wave plot with a surrounding area of five male subjects for the same utterance with the anger emotion.

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 1.4 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.

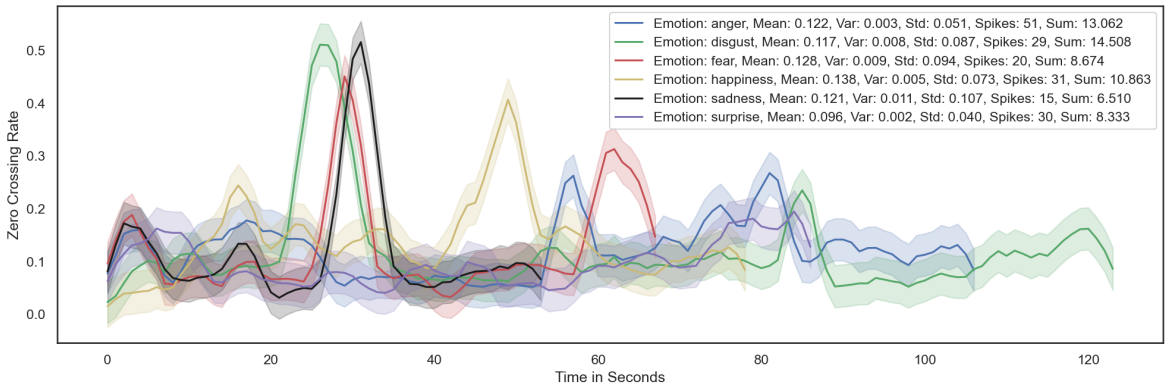


Figure 1.4: Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions.

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 1.5 shows an example of this type of plot for the mean zero crossing rate value across 50 speech utterances for all emotions. From this figure, it was observed that the values were not consistent across multiple audio segments for each emotion, which was a common observation for most extracted feature plots.

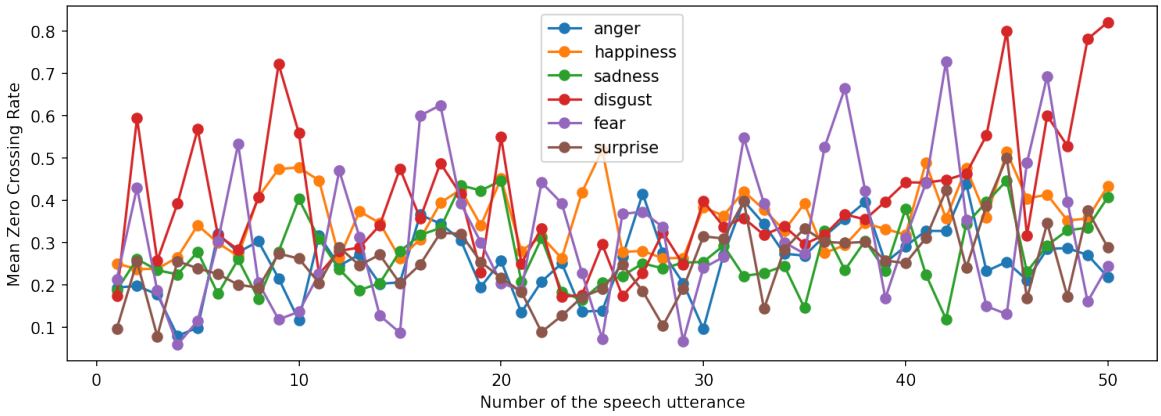


Figure 1.5: Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions

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, for the study of the features, box plots were utilized to visualize the distribution of the features on different subjects, as well as to compare the values for each emotion. An example of this is shown in Figure 1.6, which displays the mean zero crossing rate feature.

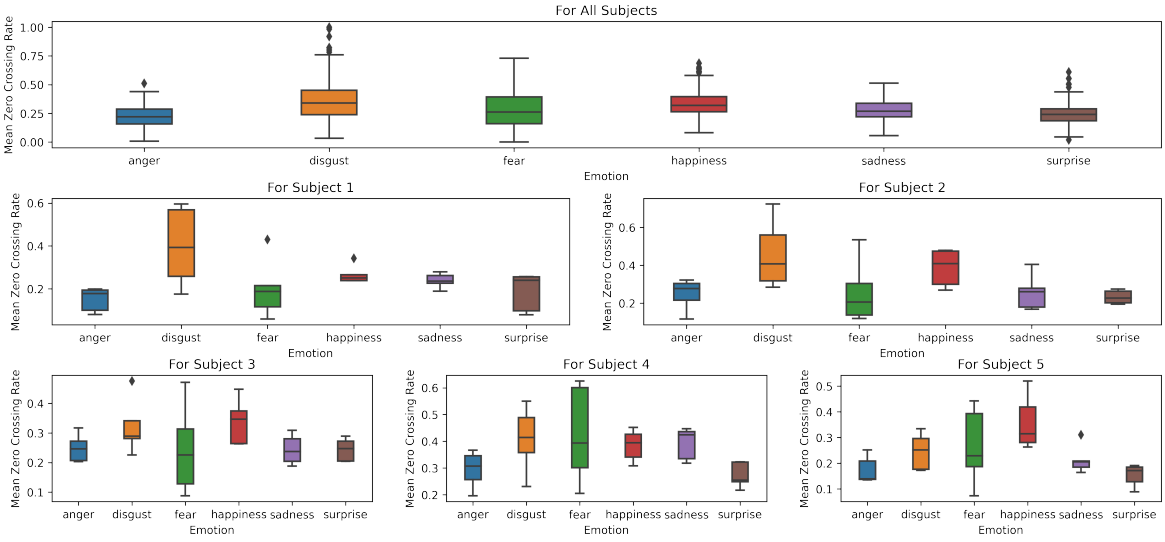


Figure 1.6: Zero crossing rate mean values box plot for all emotions and different subjects

The main 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

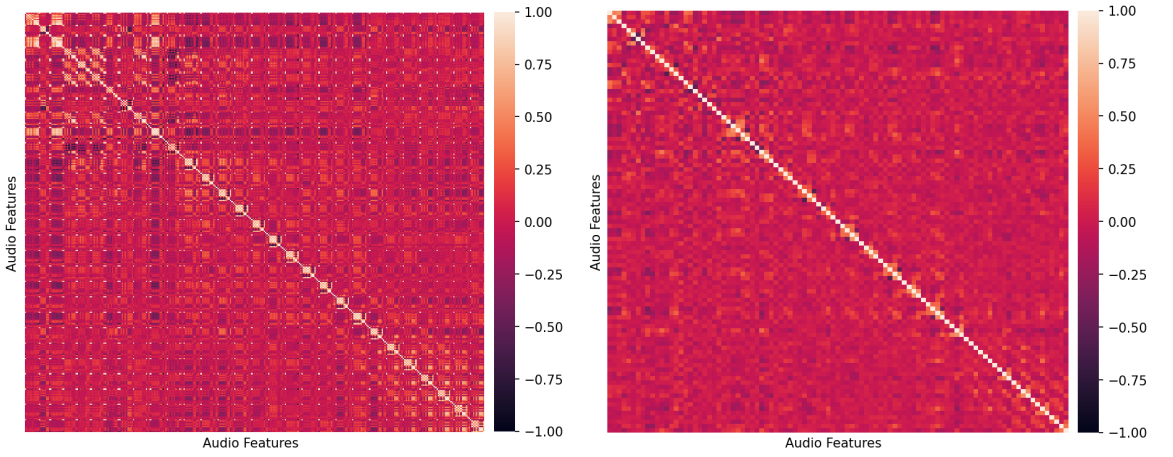
the given task. The process of feature selection is aimed to improve the accuracy of the model and reducing the problem's complexity by removing redundant or irrelevant features.

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 the figure **??**.

A high correlation elimination was performed by selecting every pair 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 230 features, leaving 97 features for subsequent analysis. The correlation matrix after the feature selection process is presented in the figure 1.7b.



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

Figure 1.7: 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 computanioally 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 **scikit-learn**, using the features obtained after the previous process, as shown in Table 1.3.

Table 1.3: Performance of various classifiers in 5-fold cross-validation using the 97 features obtained after high correlation elimination.

Classifiers	Mean Accuracies	Standard Deviation Accuracies	Total Time (s)
Ridge	0.652	0.024	0.038
Extra Trees	0.641	0.048	1.789
Random Forest	0.629	0.028	4.529
XGBoost	0.619	0.037	2.456
AdaBoost	0.585	0.025	9.018
Decision Tree	0.408	0.042	0.227
Extra Tree	0.364	0.045	0.026
C-Support Vector	0.329	0.046	0.202
Multi-layer Perceptron	0.301	0.035	0.716

Based on the evaluation results, the Ridge classifier was chosen for further analysis. This model exhibited the best prediction metrics and was also the fastest among all the evaluated classifiers. Therefore, it was deemed suitable for performing computationally expensive feature selection methods.

Feature Elimination with Backwards Propagation

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 Ridge classifier, using all features except one, and then removes 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 highest accuracy. This method involves finding a maximum and multiplying it by a threshold value of 0.99 to balance accuracy with the number of removed features. Figure 1.8 displays the mean accuracies obtained at each step and the chosen furthest highest accuracy.

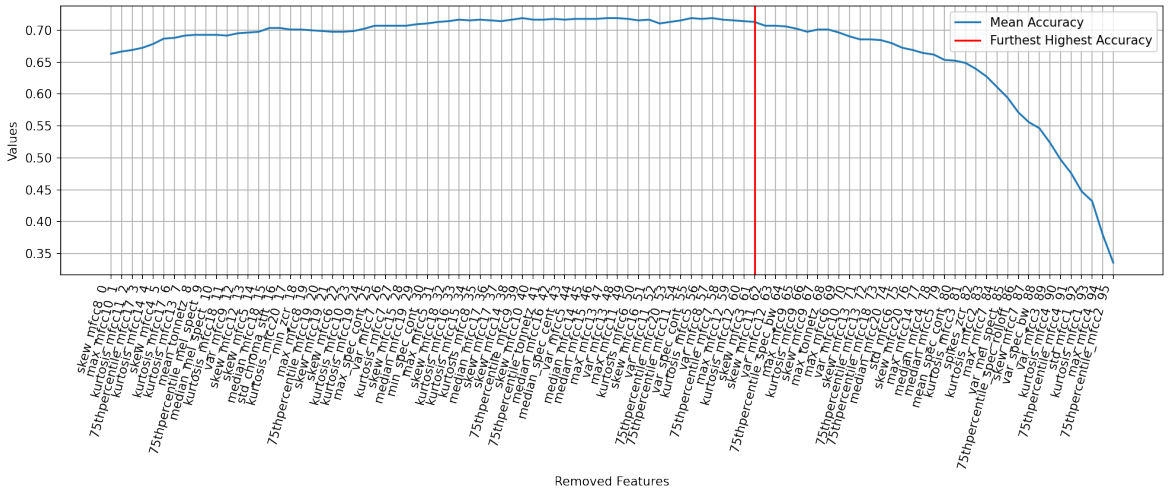


Figure 1.8: Sequential Feature Selection with Backward Propagation using the Mean Accuracy as the Selection Method.

This process led to the elimination of 62 features from the initial set of 97 obtained after the high correlation elimination, leaving a total of 35 features, as shown in Table 1.4.

Table 1.4: Selected features.

Metric	Audio Features
75th Percentile	MFCC-2, MFCC-4, MFCC-13, MFCC-18, Spectral Bandwidth, Spectral Roll-Off
Variance	MFCC-3, MFCC-4, MFCC-10, MFCC-12, Spectral Bandwidth, Mel-Spectrogram
Maximum	MFCC-4, MFCC-6, MFCC-7, MFCC-9, MFCC-14, Tonnetz
Skew	MFCC-7, MFCC-9, MFCC-11, MFCC-13, MFCC-20
Kurtosis	MFCC-1, MFCC-2, MFCC-3, MFCC-4, MFCC-9
Median	MFCC-4, MFCC-5, MFCC-20
Standard Deviation	MFCC-1, MFCC-6
Mean	Spectral Contrast
Spikes	Zero-Crossing-Rate

Testing Classifiers

Finally, we intended to corroborate the quality of the feature selection. To do so, several evaluation metrics were calculated from the estimations: 1. Accuracy 2. Macro F1 3. Precision 4. Recall 5. Hamming 6. Matthews Correlation Coefficient (MCC)

A series of predictions were performed and plotted the confusion matrix of the predictions **??**, furthermore, evaluation metrics were also calculated, as the table 1.5 shows.

In the first situation, a Random Forest model was trained with 194 of all the initial features (case 1), secondly, trained with the 50 features after the high correlation elimination (case 2), then, with the remaining 24 features after both the high correlation elimination, and, the feature elimination with backward propagation (case 3).

Moreover, one vs rest, also known as one vs all classifier, was implemented. This strategy consisted in fitting one Random Forest classifier per label. For each classifier, the label is fitted against all the other labels. For this strategy, it used all 194 initial features (case 4), and, the same of case 3, as input for each classifier (case 5).

Table 1.5: Predictions' Evaluations Metrics with Different Strategies and Input Data

Case	N.º of Features	Accuracy	Macro F1	Precision	Recall	MCC.
(1)	194	0.461	0.453	0.456	0.461	0.354
(2)	50	0.417	0.408	0.411	0.417	0.302
(3)	24	0.417	0.411	0.413	0.417	0.302
(4)	194	0.450	0.440	0.443	0.450	0.342
(5)	24	0.433	0.424	0.426	0.433	0.321

It is possible to observe from the table 1.5, that, there is some loss of accuracy from the first situation to the rest, and, the second and third situations have almost the same results. From this, if the short loss in prediction performance could be an acceptable price to pay. the third situation is optimal since there are 170 fewer features used as input data, therefore, speeding up the prediction calculation time.

In addition, the 4th case and 5th cases show similar results to the 1st case and 3rth cases respectively, so, it was concluded that using a one vs rest wasn't a good strategy, since the prediction performance wasn't improved a lot, and it increased the computational cost by using more classifiers.

1.3.4 Conclusion

During the feature analysis, most of the displayed graphs represent univariate studies of the features, therefore, if, visually, a feature is perceived as a bad descriptor, it doesn't invalidate its usage for the emotion classification problem, and it is essential to employ some other multivariate studies.

The data were also analyzed in terms of gender bias, by comparing the data of male and female subjects' audios, and, reasoned that the dataset does not contain sufficient data to take any well-founded conclusion.

1.4 Deep Learning Classifiers

notebook 3 models study categorical and the google collabs one with transfer learning

1.4.1 Raw Audio as a Feature

1.4.2 MFCC as a Feature

1.4.3 Mel-Spectrogram as a Feature

1.5 Classifiers Discussion (pros & cons)

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.