



**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



**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 Introduction	1
1.1 Motivation	1
1.2 Objectives	2
1.3 Challenges	2
1.4 Organization	3
2 State-Of-The-Art Research	5
2.1 Emotion	5
2.1.1 Discrete Emotion	5
2.1.2 Dimensional Emotion	5
2.2 Emotion Recognition Datasets	6
2.3 Speech Emotion Recognition	8
2.3.1 Speech Features	8
2.3.2 Strategies	10
2.3.3 Traditional Feature-Based Speech Emotion Recognition (SER)	11
2.3.4 Deep Learning-Based SER	13
2.4 Multimodal Emotion Recognition	15
2.5 Emotion Recognition Services	17
2.6 Software Tools	17
2.7 Summary	18
3 Methodology	19
3.1 Proposal	19
3.1.1 Requirements Gathering	19
3.1.2 Dataset	19
3.1.3 Audio Feature Engineering	20

3.1.4	Models Implementation	21
3.1.5	Speech Models Evaluation	21
3.1.6	Data Stratification	22
3.1.7	Deployment	22
3.1.8	Dissertation Writing	22
3.2	Ethical Procedures and Concerns	22
4	SER Development	23
4.1	Datasets	23
4.1.1	eINTERFACE'05	23
4.1.2	Interactive Emotional Dyadic Motion Capture (IEMOCAP)	24
4.2	Audio Preprocessing	26
4.2.1	Noise Reduction	26
4.2.2	Audio Trim	26
4.3	Traditional Feature-Based SER	28
4.3.1	Feature Extraction	28
4.3.2	Feature Analysis	28
4.3.3	Feature Selection	31
4.3.4	Classifiers Evaluation and Selection	35
4.4	Deep Learning-Based SER	37
4.4.1	Spectrogram as a Feature	37
4.4.2	Mel Spectrogram as a Feature	37
4.4.3	Mel-frequency Cepstral Coefficients (MFCC) as a Feature	37
4.4.4	Classifiers	37
4.4.5	Numeric Data Classification	37
4.4.6	Image Classification	37
4.5	Classifiers Results and Discussion	38
5	Appendix	41
.1	Audio and Descriptive Features Study	41
.2	Features Mean Values Overview	43
.3	Feature Interception Study	45
.4	Feature Variation Plots	46
.5	Confusion Matrices	47
.6	Something else	52
Bibliography		57

List of Figures

2.1	Basic emotions spanned across different emotional models.	6
2.2	Frequency distribution of databases on reviewed articles from a State-Of-The-Art (SOTA) study [35].	8
2.3	Average pitch comparison of male and female speakers' [38].	9
2.4	Traditional machine learning flow vs. automatic feature extraction with Deep Learning (DL) flow [42].	11
2.5	Multimodal emotion recognition illustration.	16
3.1	Gantt chart with the proposed work plan for this research.	19
3.2	Voice Activity Detection employment on an audio signal [104].	21
4.1	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.	24
4.2	Visual representations of audio features before and after preprocessing.	27
4.3	Zero crossing rate wave plot annotated with spikes.	29
4.4	Bar plots mean for metrics used on the mel-scaled spectrogram feature	29
4.5	Zero crossing rate wave plot with a surrounding area of five male subjects for the same utterance with the anger emotion.	30
4.6	Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions.	30
4.7	Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions	31
4.8	Zero crossing rate mean values box plot for all emotions and different subjects	31
4.9	Audio Features' Pearson Correlation Matrices Before and After High Correlation Elimination.	32
4.10	Sequential Feature Selection with Backward Propagation using the Mean Accuracy as the Selection Method.	34
4.11	Graphical representations of the features used as input for the DL classifiers.	37
.1	Audio Signal Wave Plots of one Audio Segment for All Emotions	41
.2	Log Mel Magnitude Spectrograms of one Audio Segment for All Emotions	41
.3	Mel-Frequency Cepstral Coefficients Spectrogram of one Audio Segment for All Emotions	42
.4	Chromogram Spectrograms of one Audio Segment for All Emotions	42
.5	Spectral Wave Plots of one Audio Segment for All Emotions	42
.6	Root-Mean-Square Energy Wave Plots of one Audio Segment for All Emotions	43
.7	Zero Crossing Rate Wave Plots of one Audio Segment for All Emotions	43

.8	Bar plot with the mean values of the mean spectral centroid, bandwidth, rolloff and contrast features.	43
.9	Bar plot with the mean values of the mean chromogram, root-mean-square and zero-crossing rate features	44
.10	Zero Crossing Rate Wave Plots with a Surrounding Area of Five Male Subjects and the same Sentence for All Emotions	45
.11	Zero crossing rate standard deviation values variation plot along 50 audios of speech utterances for all emotions	46
.12	Zero crossing rate spikes values variation plot along 50 audios of speech utterances for all emotions	46
.13	Zero crossing rate sum values variation plot along 50 audios of speech utterances for all emotions	46
.14	Confusion Matrix of a Ridge Classifier Trained using all 327 Extracted Features	47
.15	Confusion Matrix of a Ridge Classifier Trained using 97 Features After High Correlation Feature Elimination	47
.16	Confusion Matrix of a Ridge Classifier Trained using 35 Features After High Correlation and Backward Selection	48
.17	Confusion Matrix of a AdaBoost Classifier.	48
.18	Confusion Matrix of a Random Forest Classifier.	49
.19	Confusion Matrix of a Histogram Gradient Boosting Classifier.	49
.20	Confusion Matrix of a Balanced Random Forest Classifier.	50
.21	Confusion Matrix of a Long-Short Term Memory (LSTM) Classifier.	50
.22	Confusion Matrix of a Convolutional Neural Network (CNN) Classifier.	51
.23	Confusion Matrix of a Linear Discriminant Analysis Classifier.	51
.24	Confusion Matrix of a Ridge Classifier.	52
.25	Scatter Plot of the Annotated Emotions in the Activation Dimension	52
.26	Scatter Plot of the Annotated Emotions in the Dominance Dimension	52
.27	Scatter and violin plots of emotional content of the IEMOCAP corpus in terms of valence, activation, and dominance relative to the neutral emotion	53
.28	Scatter and violin plots of emotional content of the IEMOCAP corpus in terms of valence, activation, and dominance relative to the sad emotion	54
.29	Scatter and violin plots of emotional content of the IEMOCAP corpus in terms of valence, activation, and dominance relative to the happiness emotion	55

List of Tables

2.1	Summary and description of emotion recognition datasets.	7
2.2	Overview on features commonly used for acoustic emotion recognition [37].	9
2.3	Overview of audio classification articles with strategies.	15
2.4	Emotion recognition services for facial, textural, and speech contents [71].	17
4.1	eINTERFACE'05 subjects nationalities	23
4.2	Number of Audio Files Used per Emotion from the IEMOCAP dataset.	24
4.3	Extracted Audio Features and Statistical Functions Applied to Them	28
4.4	Performance of various classifiers in 5-fold cross-validation using the 97 features obtained after high correlation elimination.	33
4.5	Selected features.	34
4.6	Evaluation Metrics of a Ridge Classifier Predictions Using Different Sets of Features	34
4.7	Tested Classification Models Performance on IEMOCAP.	35
4.8	SOTA Traditional Classification Models Performance on IEMOCAP.	36
4.9	DL Classification Models Performance on IEMOCAP.	39

Glossary

SER	Speech Emotion Recognition	UAR	Unweighted Average Recall
VAD	Voice Activity Detection	DL	Deep Learning
MFCC	Mel-frequency Cepstral Coefficients	SOTA	State-Of-The-Art
VAD	Valence-Arousal-Dominance	FT	Fourier Transform
VAD	Voice Activity Detection	SVM	Support Vector Machine
CNN	Convolutional Neural Network	CFS	Correlation-based feature selection
DNN	Deep Neural Network	HMM	Hidden Markov Model
IEMOCAP	Interactive Emotional Dyadic Motion Capture	GeMAPS	Geneva Minimalistic Acoustic Parameter Set
LSTM	Long-Short Term Memory	RNN	Recurrent Neural Network
		MCC	Matthews Correlation Coefficient

Introduction

“By humanizing technology, we have this golden opportunity to reimagine how we connect with machines, and therefore, how we, as human beings, connect with one another.”

- Rana el Kaliouby [1]

1.1 MOTIVATION

Affective computing also referred to as emotion artificial intelligence, is a branch of artificial intelligence that is focused on the measurement, understanding, simulation, and response to human emotions. It is a continuously growing multidisciplinary field that explores how technology can inform an understanding of human affect, how interactions between humans and technologies can be impacted by affect, how systems can be designed to utilize affect to enhance capabilities, and how sensing and affective strategies can transform human and computer interaction [2].

Recently, emotion recognition and sentiment analysis gained interest in the para-linguistic processing world and became an expanding research topic. Emotion recognition and sentiment analysis focus on developing techniques for automatically recognizing emotional states. Both can be treated as two affective computing subtasks on different levels [3]. Sentiment analysis allows the understanding of the general feelings and emotions experienced by a person, typically associated with sentiment polarity, which determines data as positive, negative, and neutral. In contrast, emotion recognition uses a more objective and precise system, it focuses on identifying a broad spectrum of distinct human emotions such as anger, joy, or fear.

A quote by MIT Sloan professor Erik Brynjolfsson highlights the importance of emotional awareness for genuine and satisfying human-computer communication: “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”. By comprehending emotional states expressed by human beings, a machine can give appropriate responses to each situation.

For example, in several contexts of a video conversation, detecting emotions could affect the dialog strategies:

- In teaching lectures, whether the teacher is a human or an intelligent spoken tutoring system, extracting and analyzing information about student’s emotional state is a very important part of teaching because it allows the teacher to adapt [4];

- In call centers, it helps to detect potential problems that arise from an unsatisfactory course of interaction [5]–[7]. Calls can be evaluated regarding customer satisfaction and the evolution of states throughout the call;
- In interviews, for example, it can serve as an aid to analyze suspects’ behavior during police questioning;
- For product reviews it can be used to evaluate the customer’s level of satisfaction while using or reviewing the product;
- In Human-Robotic interfaces, such as a robotic pet, it allows it to detect and manage tension in human interactions by reading the emotional state of its human commander, for example, Kanda, Iwase, Shiomi, *et al.* constructed a robot that focused more on detecting tensions [8];

For these reasons, multimodal affective computing in which a variety of data sources, including voice, facial expression, gestures, and linguistic content are employed [9], is a very challenging and promising research area.

1.2 OBJECTIVES

The main objective of this dissertation is to research and develop a SER system that can accurately identify emotions from audio streams of natural and unstructured conversations in a video conference setting.

To achieve this, we will conduct a thorough examination of speech features and evaluate various machine-learning models to determine the most suitable approach. The goal is to develop a unique set of audio features and achieve accuracy levels of emotion recognition that are comparable to or exceed current SOTA methods.

An additional objective is to create a system that can perform emotional analysis both offline and in real-time, which may involve designing separate audio stream processing pipelines for each mode.

Finally, we will also investigate the potential benefits of combining the speech model with a given facial emotion recognition model. The goal is to explore how multimodality fusion techniques can improve the overall performance of the system.

1.3 CHALLENGES

The ability to detect emotions is a difficult task for computers and even humans. One of the main challenges in the field of emotional analysis is the lack of datasets with authentic interactions that include information from various channels. Many existing data sets rely on subjects being asked to “act” or “simulate” emotions while being recorded, which can provide a controlled and simplified data collection process, but may not accurately represent real-life emotions. As a result, the performance of emotion recognition can suffer when models developed from these datasets are used in real-world situations where a mix of emotions is present. Another issue with existing data is that they often consist of isolated utterances or short dialogues, which do not consider the role of context in emotion perception and expression. Additionally, current emotional datasets are typically limited in size and in the number of subjects. The labeling of these datasets is prone to errors due to subjectivity in the evaluation, which implies the need for several evaluators, and the decision being made by majority vote.

The most natural form of human communication is considered to be speech [10], consequently, recognizing emotions from speech is extremely useful, as the way we speak can convey a lot of information about our emotional state. A lot of research has been conducted in order to understand and extract low-level audio features that are good descriptors of the actual affective state, also known

as feature engineering. It demands a deep understanding of the data, as it can cause loss of information from the original audio signal due to the fact it selects the most representative of the data.

Although speech is a valuable input signal for emotion recognition, it is generally considered to be more effective to use multiple modalities. This is because distinct emotions can be expressed differently depending on the modality. For example, while happiness and excitement may be expressed through smiling, laughing, and high-pitched speech, sadness may be expressed through crying, a low-pitched voice, and a slumped posture. It can also make the system more robust to noisy information from some modalities, e.g., background noises, light, blur, etc. Consequently, building multimodal systems introduces additional complexities, such as fusing different modalities and understanding how each source affects the output, including dealing with noisy or even faulty data. In addition, each input source, brings its own biases and covariates, for example, the speech modality has biases toward gender, age, and language. Previous studies have also shown that facial emotion recognition programs fail the racial bias test and have trouble reading faces with darker skin tones [11].

More challenges arise when building an emotion recognizer to be used on a video conference system. It requires sophisticated algorithms that can deal with spontaneous and dynamic interactions, as well as potential privacy concerns. Analyzing these systems at different times, in real-time or offline, can provide useful insights. Real-time analysis can be useful in specific contexts, but achieving high accuracy is more demanding. Another option is to process data in specific timed intervals.

1.4 ORGANIZATION

This document is composed of three chapters.

The introduction chapter 1 begins by briefly presenting the emotion recognition area and explaining the main objectives and challenges of this dissertation.

The SOTA chapter 2 describes previous work in the area of emotion recognition and speech analysis, providing background information and discusses related concepts, including emotion, emotional datasets, speech, and its audio features, SER strategies and applications, software tools, etc.

The proposal chapter 3 outlines the workflow, schedule, and methodology that will be followed during the development of the research.

2

CHAPTER

State-Of-The-Art Research

2.1 EMOTION

The first step to building a successful SER model is to have the concept of emotion carefully defined. There are various approaches to modeling emotions, and there is still no consensus on an accurate representation due to its natural subjectivity.

Currently, the main accepted and utilized for the emotion recognition task are the discrete and dimensional models.

2.1.1 Discrete Emotion

Discrete/categorical theories of emotions claim that some core emotions are characterized by stable patterns of triggers, behavioral expression, and associated distinct subjective experiences [12].

The emotions considered in these theories are typically the *basic* emotions, including happiness, sadness, fear, anger, disgust, and surprise.

Most of the existing systems focus on these basic emotional categories. Instinctively, people use this model in their daily life, consequently, models based on discrete labels are more intuitive.

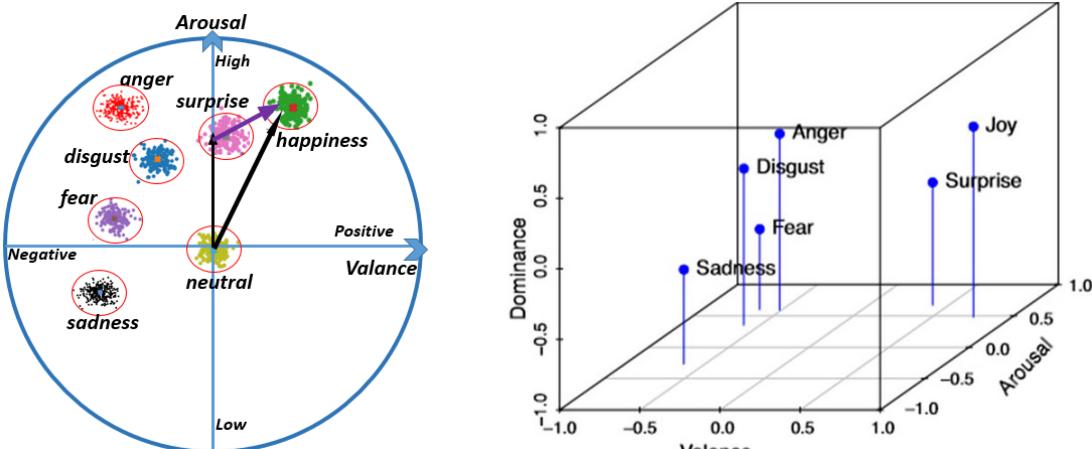
2.1.2 Dimensional Emotion

The dimensional emotional model uses dimensions such as valence, arousal, control, or power, to characterize emotions. It represents emotions as coordinates in a multidimensional space. These dimensions are definitive and generic aspects of emotion. One of the most preferred dimensional models is a two-dimensional model, as shown in figure 2.1a below, that uses arousal or activation, versus valence or evaluation on the other [13].

The arousal dimension refers to the level of energy or activation associated with an emotion, or in other words, the strength of the emotion. It measures whether humans are more or less likely to take some action under an emotional state.

The valence dimension measures how a human feels or the nature of the emotion, from positive to negative.

Also, a three-dimensional model may include a dimension of dominance or power, which refers to the seeming strength of the person, or the degree to which an individual feels in control or in charge of the situation that is causing the emotion, usually referred to as Valence-Arousal-Dominance (VAD) model [16], displayed on figure 2.1b.



(a) The 2D Valence-Arousal model [14].

(b) The 3D Valence-Arousal-Dominance model [15].

Figure 2.1: Basic emotions spanned across different emotional models.

Dimensional representations of emotions are useful because they provide a wider range of describing an emotional state. Also, they are more suitable for quantifying variations over time, since changing successively from one discrete emotion to another may not make much sense in a real scenario. However, this model isn't as intuitive, therefore, it makes the distinction between a few basic emotions harder, such as surprise, since depending on the context it can have opposite values of valence [13].

2.2 EMOTION RECOGNITION DATASETS

The quality of the data affects the success of the recognition process. Incomplete, low-quality, or faulty data may lead to incorrect predictions; hence, data should be carefully designed and collected.

When creating a database for emotion recognition research, researchers must consider factors such as the age, gender, and language of the speakers. Most databases include adult speakers, but there are also databases featuring children and elderly individuals. To improve the representativeness and usefulness of the database, researchers may also consider using different actors to express different emotions.

There are various databases containing emotional expressions in different languages. According to Scherer, vocal emotion expressions may be driven by universal psychobiological mechanisms, as individuals from diverse cultures and speaking different languages can accurately recognize emotions expressed through speech better than would be expected by chance [17]. This is supported by research demonstrating that even infants who cannot yet speak can recognize emotional cues in adult speech [18]. In another study, Rajoo and Aun investigated the influence of language on the ability of a SER system, using spoken expressions of four selected emotions (anger, sad, happiness, and neutral) in three languages (Malay, English, and Mandarin). Their study revealed that there are language-specific differences in emotion recognition, with English showing a higher recognition rate compared to Malay and Mandarin. Additionally, the results support the conclusion that SER is language-independent, but also suggest that emotions expressed by native speakers are more accurately recognized [19].

Databases for emotion recognition that can be used have three types:

- **Acted:** recorded by professional or semi-professional actors in sound-proof studios;
- **Elicited:** created by, for example, placing speakers in a simulated emotional situation that can stimulate various emotions;

- **Natural:** mostly obtained from talk shows, call-center recordings, radio talks, and similar sources. Sometimes, these real-world speeches are referred to as spontaneous speech.

In video conferencing applications, the conversations take place in natural contexts, which is a fundamental factor to consider when choosing a dataset. Research has demonstrated that there are differences in the characteristics and accuracy of acted and genuine emotions [20]. Some studies have also suggested that simulated emotional speech may not be fully genuine and may be perceived more intensely than genuine emotional speech [21]. Therefore, the studies on emotion recognition have shifted from focusing on induced expressions to spontaneous ones, but data on audiovisual expressions in natural contexts is still limited. Emotion expressions are infrequent, fleeting, and associated with complex contextual structures, making it difficult to collect a large amount of spontaneous data, as shown in a survey on audiovisual emotion recognition [22]. Ostensibly, different types of databases are suitable for different purposes. For example, a database designed for mainly theoretical research may be more suitable in certain cases rather than one that is intended for use in a real-life industry application.

A summary of prominent datasets and their description is provided in the table 2.1.

Table 2.1: Summary and description of emotion recognition datasets.

Dataset	Format	Language	Content	Emotions	Type
EMO-DB [23]	Audio	German	800 recording spoken by 10 actors (5 males and 5 females).	7 emotions: anger, neutral, fear, boredom, happiness, sadness, disgust.	Acted
eINTERFACE'05 [24]	Audio Video	English	Videos by 42 subjects, coming from 14 different nationalities.	6 emotions: anger, fear, surprise, happiness, sadness and disgust.	Elicited
IEMOCAP [25]	Audio Video Text	English	Conversations among pairs of 10 speakers (gender balanced) spanning 12 hours.	10 categorical emotions & VAD	Acted & Elicited
MOUD [26]	Audio Video	Spanish	80 product reviews YouTube videos. 7442 clips of 12	Positive, negative, or neutral sentiment. 6 emotions: angry, disgusted, fearful, happy, neutral, and sad.	Natural
CREMA-D [27]	Audio Video Text	English	sentences spoken by 91 actors (gender balanced).	6 emotions: angry, disgusted, fearful, happy, neutral, and sad.	Acted
CMU-MOSI [28]	Audio Video	English	2199 movie reviews with annotated sentiment.	Very negative to very positive in seven Likert steps.	Natural
MSP-Improv [29]	Audio Video Text	English	8438 recordings by 12 actors.	4 emotions: angry, sad, happy, neutral.	Elicited
CMU-MOSEI [30]	Audio Video Text	English	65 hours of YouTube videos from more than 1000 speakers (gender balanced) and 250 topics.	6 emotions: happiness, sadness, anger, fear, disgust, surprise. Polarity in Likert scale.	Natural
MELD [31]	Audio Video Text	English	1400 dialogues and 14000 utterances from Friends TV series by multiple speakers.	7 emotions: Anger, disgust, sadness, joy, neutral, surprise and fear. Also positive, negative and neutral.	Acted
RAVDESS [32]	Audio Video	English	7356 recordings by 24 actors with 2 statements only.	7 emotions: calm, happy, sad, angry, fearful, surprise, and disgust.	Acted
MSP-Podcast [33]	Audio	English	100 hours by over 100 speakers	Activation, dominance and valence and 9 categorical labels	Natural
TESS [34]	Audio	English	2800 recording by 2 actresses.	7 emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral.	Acted

In a 2021 SER SOTA review article, Jahangir, Teh, Hanif, *et al.* exhibited the most commonly used databases in their selected studies, figure 2.2 [35]. They realized the most popular ones had a

gender-balanced number of speakers with high-quality recordings.

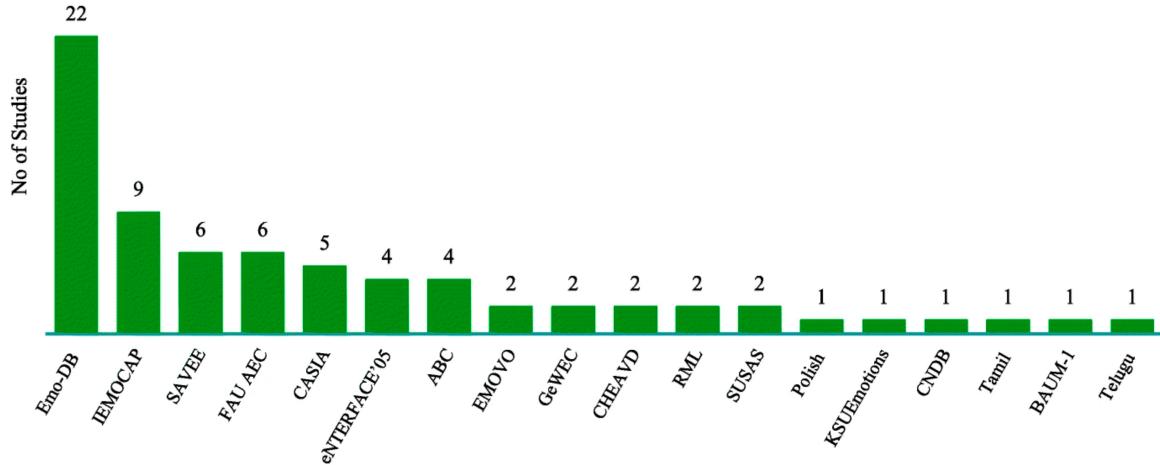


Figure 2.2: Frequency distribution of databases on reviewed articles from a SOTA study [35].

2.3 SPEECH EMOTION RECOGNITION

SER is a subfield of artificial intelligence that focuses on the development of systems that can identify and classify the emotional state of a speaker based on their speech.

Speech is a vital tool for human communication and social interaction. Fundamentally, it is a continuous signal that can convey information, express emotions, and share meaning [36].

Words alone do not always transmit sufficient emotion, for instance, text messages can be easily misconceived. Hence, the necessity of SER to aid and complete this task of figuring out the emotions/intent of human interaction.

2.3.1 Speech Features

Speech features are characteristics of an audio signal that can be extracted and analyzed to understand its properties. These features are commonly categorized in the following three classes:

- **Prosodic Features;**
- **Spectral Features;**
- **Voice Quality Features.**

Prosodic and spectral features are commonly employed in various fields, as they have been shown to effectively improve outcomes. By combining these features, it is possible to further optimize their usefulness and attain even better results.

Currently, there are a wide variety of acoustic properties that can be used to extract a large feature vector. Examples of these features include pitch, intensity, duration, and voice quality, which then several common statistics can be applied to them, such as minimum, mean, maximum, etc., as shown in table 2.2 adapted from a book [37].

Table 2.2: Overview on features commonly used for acoustic emotion recognition [37].

Intonation (F0 or pitch modelling)	Deriving (raw LLD, deltas, regression coefficients, LDA, PCA, ...)	Filtering (smoothing, normalizing, ...)	Chunking (absolute, relative, ...)	Extremes (min, max, range, ...)	Deriving (raw LLD, deltas, regression coefficients, LDA, PCA, ...)	Filtering (smoothing, normalizing, ...)
Intensity (energy, Teager, ...)				Mean (arithmetic, absolute, ...)		
Linear Prediction (LPCC, PLP, ...)				Percentiles (quartiles, ranges, ...)		
Cepstral Coefficients (MFCC, ...)				Higher Moments (std. dev., kurtosis, ...)		
Formants (amplitude, position, ...)				Peaks (number, distances, ...)		
Spectrum (MFB, NMF, roll-off, ...)				Segments (number, duration, ...)		
TF-Transformation (Wavelets, Gabor, ...)				Regression (coefficients, error, ...)		
Harmonicity (HNR, spectral tilt, ...)				Spectral (DCT coefficients, ...)		
Perturbation (jitter, shimmer, ...)				Temporal (durations, positions, ...)		

Low-Level-Descriptors

Functionals

Prosodic Features

Prosodic features are those that can be perceived by humans, such as intonation, rhythm, and loudness. The most widely used ones are based on **pitch**, **energy**, and **duration**.

Pitch, also known as the fundamental frequency or F0, refers to the lowest frequency of a periodic waveform. It is often used to convey emotion because it is influenced by the tension in the vocal folds and the sub-glottal air pressure. Studies have shown that the average pitch for males is typically about an octave lower than that of females, with an average pitch of 120 Hertz for males, 229 Hertz for females, and 155–185 Hertz for those with ambiguous gender. According to the study conducted by Rathina in 2012, there are clear differences in the pitch of the female and male genders when expressing different emotions. These differences may present a challenge for a gender-independent emotion classifier that relies solely on pitch [38].

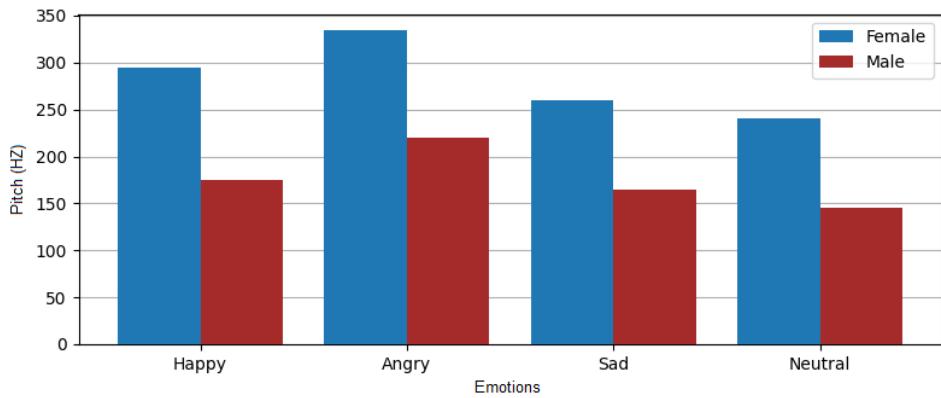


Figure 2.3: Average pitch comparison of male and female speakers' [38].

Prosodic features have been indicated as the most distinctive properties of emotional content for speech emotion recognition [39].

Spectral Features

In the field of speech and audio processing, spectral features are used to analyze the characteristics of the human voice, which can convey information such as emotion, identity, and speaker traits. The

spectral content of speech signals varies significantly depending on the sounds being produced and the language being spoken, and it is often highly dynamic and time-varying.

One commonly used spectral feature is the MFCC. MFCC are calculated by performing a Fourier Transform (FT) on a window of audio data to obtain the spectral power, which is then mapped to the Mel-scale (a frequency scale that reflects how the human auditory system perceives sound). The Mel-scaled spectral power is then transformed into the cepstral domain using a discrete cosine transform, resulting in the MFCC that represent the spectral power of the signal in the cepstral domain and can be used as input for tasks such as emotion recognition.

Another spectral feature is the gammatone frequency cepstral coefficient, which is similar to MFCC in that they are derived using a FT and a filter bank (in this case, a gammatone filter bank that models the auditory system's response to different frequencies). Linear prediction cepstral coefficients are another type of feature that is derived from the linear prediction of a signal. Log-frequency power coefficients are a type of feature that mimic the logarithmic filtering characteristics of the human auditory system and are obtained by measuring spectral band energies using a fast Fourier transform.

Voice Quality Features

Voice quality is the set of characteristics that distinguish one person's voice from another. Some common methods for analyzing voice quality involve examining the physical properties of the vocal tract, such as jitter, shimmer, and the harmonics-to-noise ratio.

These properties may change involuntarily and can be used to differentiate emotions in the speech signal. Jitter refers to the variability in the period, or the time between successive peaks, of the fundamental frequency, whereas shimmer refers to the variability in the amplitude. Harmonics-to-noise ratio is a measure of the relative strength of the harmonics in a speech signal compared to the noise.

These three measures are often used together to assess the quality of a person's voice and can be useful in several tasks, e.g., treatment of voice disorders, age detection, speaker or emotion recognition, etc.

2.3.2 Strategies

The use case for this specific system of SER is a video conferencing application where conversations are unpredictable, therefore, it is necessary to have a robust speech recognition system, however, some approaches can create privacy issues, for users who may not want to have their private conversations public, and even, ethical issues. Therefore, combining acoustic and linguistic content, by transcribing the speech to text, was not intended in this work, nonetheless, it is a remarkable approach that would potentially improve the system.

A SER system requires a large and diverse dataset of speech samples, as well as feature extraction and preprocessing techniques to obtain relevant features from the audio signal, such as the ones mentioned above. Appropriate machine learning algorithms are then used to classify the emotional content of the speech based on the extracted features. In the past several years, classical machine learning algorithms, such as Hidden Markov Model (HMM) [40], Support Vector Machine (SVM), and decision tree-based methods, have been utilized for audio sentiment analysis. More recently, researchers have proposed various neural network-based architectures to improve audio sentiment analysis. With the development of DL, more complex neural-based architectures are proposed. For example, CNN-based models have been used to train with the audio spectrograms or features derived from the signal.

Currently, there are two main types of implementations [41]:

- **The traditional feature-based SER:** hand-crafted features are extracted by applying a series of statistical aggregation functions to acoustic features of the audio signal, which are then passed on to a classifier;
- **Automatic feature extraction-based SER:** avoids manual feature engineering and leverages the abilities of DL techniques. This method typically involves using raw speech signals, Mel spectrograms, or MFCC as input for a Deep Neural Network (DNN) model.

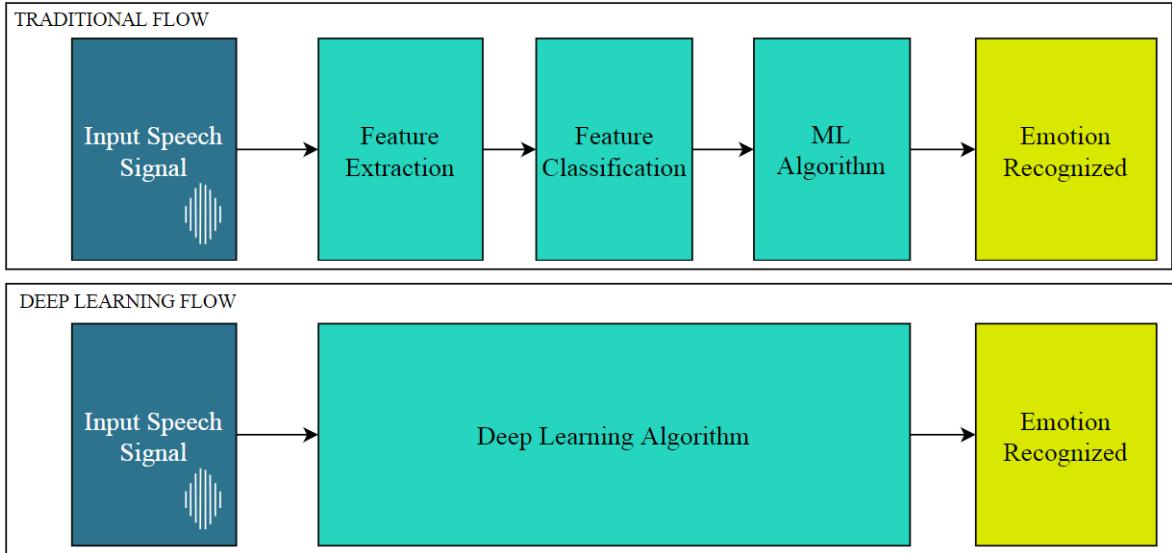


Figure 2.4: Traditional machine learning flow vs. automatic feature extraction with DL flow [42].

Figure 2.4, displays the main difference between the methods, the traditional implementation has more steps, because it mainly depends on the feature extraction and selection processes. The DL approaches avoid feature engineering and tend to improve recognition efficiency, however, it significantly increases complexity and the computational cost, which has a negative effect on SER systems used in real-time scenarios.

2.3.3 Traditional Feature-Based SER

The traditional feature-based SER is a challenging task for researchers. It consists of applying a series of statistical aggregation functions (e.g., mean, max, variance, etc.) on acoustic features extracted from the speech to produce a statistical feature vector. The feature vector obtained can describe the temporal variations of speech signals, which are considered to be associated with the subjacent emotion. This vector is then analyzed, and feature selection is performed on it to reduce its size and reduce overfitting to the data, and finally, various classifiers are given that vector as input and are evaluated to find the most suitable for the task.

In SOTA articles, the final set of selected features is often not the same across different studies. This selection of features for training classifiers has a significant impact on the overall performance of the SER system, which is why it is a major challenge.

The statistical feature vector may contain only global features, extracted from the full-length utterance, or several local features, extracted from speech segments obtained from framing the utterance. The performance of the features based on their granularity is also analyzed in several studies.

Schuller, Rigoll, and Lang found that using global features (computed from gross statistics) resulted in a recognition rate of 86.6% while using local features (computed from syllable duration, pitch, and

energy values) resulted in a recognition rate of 77.6%, while human judges had a recognition rate of 79.8% [43]. In a separate study, Rao, Koolagudi, and Vempada observed that combining local and global prosodic features slightly improved performance compared to using only local features [44].

Forward methods and backward methods are two common techniques for feature selection. Forward methods begin with an empty (or very limited) feature set and add features incrementally. The most well-known forward method is the Sequential Forward Selection algorithm, which at each step tries adding each feature to the set and keeps the one that results in the highest improvement in accuracy.

Luengo, Navas, Hernández, *et al.* (2005) applied a Forward 3-Backward 1 wrapper method to select features. This method involves choosing the feature that maximizes accuracy in each step using leave-one-out testing. After three consecutive selections, the least useful feature is eliminated. Without feature selection, the authors obtained a recognition rate of 93.50% with a SVM classifier and 84.79% with a Gaussian Mixture Model using 86 prosodic features. However, by using the selected six best features, the recognition rates changed to 92.38% and 86.71%. In conclusion, it was noted that a slight reduction in recognition rate is compensated by a much lower computational cost of extracting and training the features.

In 2015, Gosztolya demonstrated that classification and regression algorithms can benefit from even simple feature selection techniques. They proposed a simple greedy forward-backward feature selection algorithm, which is less computationally costly than any other methods, for speech conflict intensity estimation, which significantly outperformed previous scores [46].

Correlation-based feature selection (CFS) is another popular method used in machine learning and data mining. CFS involves calculating the correlation between each feature and the target variable, ranking the features based on their correlations, and selecting a subset of the highest-ranked features for use in a model. The goal of CFS is to select a subset of features that are highly correlated with the target variable and not highly correlated with each other, to improve the performance of the model and reduce overfitting. In a study by Schuller, this method was applied to reduce the number of acoustic features from 760, for each of valence, activation, and dominance dimensions, to 238, 109, and 88, respectively, and improved their results by doing so [47].

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) is a set of features used in speech emotion recognition developed by researchers at the University of Geneva [48]. It is designed to be a compact and efficient feature set for automatic emotion recognition from speech. It includes a range of prosodic and acoustic features that are relevant to emotion recognition, such as pitch, intensity, spectral properties, and temporal features. GeMAPS feature set has been utilized in several studies on speech emotion recognition, e.g., [49], as well as in applications such as affective computing, virtual assistants, and human-computer interaction.

Many of the early researchers used traditional machine learning classification techniques such as Linear Discriminant Analysis, Regular Discriminant Analysis, SVM, K-Nearest Neighbors, Random Forests, and Gaussian Mixture Models [50]. Moreover, ensemble learning has also been employed by researchers to combine the predictions of multiple individual classifiers to improve performance.

Albornoz, Milone, and Rufiner built a hierarchical classifier with two stages, on each stage they predicted different emotions using a different set of features and classifiers [51]. They showed that 12 mean MFCC features, deltas, and acceleration coefficients are the best features for the first stage with a HMM with 30 Gaussians in the mixture classifier. For the second stage with a HMM, they obtained as the best features, 12 means MFCCs, 30 means log-spectrum, and mean and standard deviation pitch. Using the EMO-DB corpus, they achieved an accuracy of 71.75%.

In a study, Lee, Mower, Busso, *et al.* (2011) used a traditional strategy to extract 16 low-level

descriptors from audio data, including zero crossing rate, root-mean-square energy, pitch, harmonics-to-noise ratio, and 12 MFCC and deltas. They then applied 12 different statistical functions, resulting in a set of 384 acoustic features per utterance. The researchers used a multi-level binary decision tree structure with different classifiers at each level to classify the data by dividing the problem into sub-problems. For example, at the first level, they compared the "neutral" label to all other labels. This approach resulted in an Unweighted Average Recall (UAR) of 41.57% on the AIBO dataset and 58.46% on the IEMOCAP dataset [52], with 5 and 4 classes respectively.

In 2019, Sahu developed a set of hand-crafted features for an audio signal using the IEMOCAP dataset. The authors obtained an eight-dimensional vector by applying statistical functions to acoustic features such as pitch, harmonics, and pause [53]. They also found that ensembling models improved performance, and they achieved a maximum accuracy of 56.6% with a model that combined random forest, extreme gradient boosting, and a multilayer perceptron.

2.3.4 Deep Learning-Based SER

As DL technology has advanced, automatic feature learning algorithms have become increasingly popular for SER, and even in other areas, for example, in cardiac spectrogram analysis, it is one of the state-of-art approaches, [54] and [55], in which the spectrogram of the audio from the heartbeat is used to evaluate the cardiac activity. These algorithms are effective at learning task-specific features without the need for extensive manual feature engineering.

DNN techniques for SER are primarily trained using raw signals, spectrograms, low-level descriptors, and MFCC, which have been shown to produce satisfactory results. Various DNN architectures including Recurrent Neural Network (RNN), LSTM, and attention-based convolution RNN have been designed for SER, each with its strengths and weaknesses. CNNs and their variations are the most common, as they are known for their ability to extract a large number of hidden features from signals and images.

A new architecture was introduced in 2020 by Issa, Demirci, and Yazici, for identifying emotions in sound files using a one-dimensional CNN. They obtain a one-dimensional array by taking mean values along the time axis of various features, including MFCC, chromagram, Mel spectrogram, Tonnetz representation, and spectral contrast, from the sound files and use them as inputs for the CNN. The proposed framework was evaluated using samples from the RAVDESS, EMO-DB, and IEMOCAP datasets. The framework achieved accuracies of 71.61% for RAVDESS (with 8 emotions), 86.1% for EMO-DB (with 7 emotions), 95.71% for EMO-DB (with 7 emotions), and 64.3% for IEMOCAP (with 4 emotions) in speaker-independent audio classification tasks [56].

It is important to note that the way the information contained in an audio spectrogram is used can be different, as some authors use three channels of the spectrogram, such as the RGB of the plotted image, or even using the static, delta, and delta-deltas of it, essentially having a 3-D tensor, while other authors use a 2-D tensor with the spectrogram values.

In a paper published in 2021, García-Ordás, Alaiz-Moretón, Benítez-Andrade, *et al.* proposed a fully CNN that can process variable input lengths for near real-time sentiment analysis [57]. The use of MFCCs made it easier to identify emotions in audio signals. The model achieved superior performance compared to other machine learning models on the EmoDB, RAVDESS, and TESS datasets, with accuracies of 75.28%, 92.71%, and 99.03%, respectively.

The use of CNNs and RNNs increased in recent years, driven by their successes in various applications, and researchers have begun exploring the potential benefits of combining these models into a single architecture. For example, in a previous study, Ma, Wu, Jia, *et al.* (2018) proposed a neural network architecture designed specifically to handle variable-length speech by combining CNN-based

deep spectrogram representations with a RNN to process variable-length speech segments. Their model achieved a weighted and unweighted accuracy of 71.4% and 64.22% on the IEMOCAP dataset [58]. In another study that used the same corpus, Zhao, Bao, Zhao, *et al.* (2019) developed an attention-based bidirectional LSTM RNN with attention-based fully convolutional networks and achieved weighted and unweighted accuracies of 68.1% and 67.0% [59].

Luo, Xu, and Chen in 2018, proposed a model for audio sentiment analysis that combines multiple traditional acoustic features and spectrum graphs. The authors analyze the MFCC, spectral centroid, chroma-short-time FT, and spectral contrast. The best model used is a combination of a LSTM network and a CNN. The model was tested on the CMU-MOSI and MOUD datasets, achieving accuracies of 68.74% and 69.64% for 4 and 2 emotion classes, respectively.

In 2018, Chen, He, Yang, *et al.* proposed a 3-D attention-based convolutional RNN, which learns discriminative features from the input data. The input to the model is a Mel spectrogram with static, deltas, and delta-deltas, which is processed by a 3-D CNN and then combined with a LSTM for high-level feature extraction. To address silent and emotion-irrelevant frames, an additional attention model is used to focus on emotion-relevant frames and produce discriminative utterance-level features. Their proposed method achieved SOTA performance on the IEMOCAP and Emo-DB corpora, with UARs of $64.74 \pm 5.44\%$ and $82.82 \pm 4.99\%$ [61], using a 10-fold cross-validation technique.

In a paper published by Muppidi and Radfar, the authors proposed one of the first quaternion-based CNN for SER. Their model is significantly smaller than other DL models and achieved SOTA results in terms of accuracy on several datasets, including RAVDESS (77.67% accuracy with 8 emotions), IEMOCAP (70.46% accuracy with 4 emotions), and EMO-DB (88.78% accuracy with 7 emotions) [62].

Transfer Learning

Transfer learning is a machine learning technique that involves using the knowledge gained from training a model on one task to improve the performance of a model on a related task.

As stated before, finding a large amount of labeled data to train a SER system can be challenging. That is why leveraging the knowledge gained from training a model on a large dataset of speech samples can be particularly useful to reduce the amount of data and computational cost required.

Although there is a significant difference between audio signals and spectrograms and standard ImageNet images, the principles of transfer learning still apply. In a 2020 paper, Palanisamy, Singhania, and Yao demonstrated this by using pre-trained weights from image classification models led to improved performance compared to using randomly initialized weights. Their ensemble of models pre-trained on ImageNet achieved SOTA results on the ESC-50 and UrbanSound8K datasets [63].

It is also common to extract features from the spectrogram images using pre-trained models and use them as input to more traditional machine learning models, as shown in a study by Zhang, Zhang, Huang, *et al.*, that used the pre-trained CNN AlexNet model to learn deep features and feed them to a SVM [64].

In table 2.3 we provide a list of our reviewed articles for both strategies commonly used in audio classification tasks, and, subsequently, SER tasks.

Table 2.3: Overview of audio classification articles with strategies.

Paper Title (Publication Date)	Audio Features	Classifier	Datasets (Nº of Labels) & Accuracies (%)
Traditional Feature-Based Approaches			
Spoken emotion recognition using hierarchical classifiers (Jul. 2011) [51]	MFCC Log-spectrum Pitch Zero crossing rate Root-mean-square energy Pitch MFCC Harmonics-to-noise ratio Pitch Harmonics Pause	HMM Multi-level binary decision trees Random forest, extreme gradient boosting and multilayer perceptron	EMO-DB (7): 71.75 (UARs) AIBO (5): 41.57 IEMOCAP (4): 58.46 IEMOCAP (4): 56.6
Emotion recognition using a hierarchical binary decision tree approach (Nov. 2011) [52]			
Multimodal SER and Ambiguity Resolution (2019) [53]			
Deep Learning-Based Approaches			
SER with deep CNNs (May 2020) [56]	MFCC Chromagram Mel spectrogram Spectral contrast	One-dimensional CNN	RAVDESS (8): 71.61 IEMOCAP (4): 64.30 EMO-DB (7): 86.10
Sentiment analysis in non-fixed length audios using a Fully CNN (2021) [57]	Mel-Spectrogram MFCC	Fully CNN	RAVDESS (8): 75.28 EMO-DB (7): 92.71 TESS: 99.03
Emotion Recognition from Variable-Length Speech Segments Using DL on Spectrograms (2018) [58]	Log-Spectrograms	CNN and RNN	IEMOCAP (4): 64.22
Exploring Deep Spectrum Representations via Attention-Based RNN and CNN for SER (2019) [59]	Mel-Spectrogram	Attention-based bidirectional LSTM and fully CNN	IEMOCAP (4): 67.0
Audio Sentiment Analysis by Heterogeneous Signal Features Learned from Utterance-Based Parallel Neural Network (2018) [60]	MFCC Spectral centroid Spectral contrast Chromagram	LSTM and CNN	MOSI (4): 68.74 MOUD (2): 69.64
3-D Convolutional RNN With Attention Model for SER (2018) [61]	Mel spectrogram	3-D attention-based convolutional RNN	(UARs) IEMOCAP (4): 64.74 EMO-DB (7): 82.82 RAVDESS (8): 77.87
SER Using Quaternion CNN (Jun. 2021) [62]	Mel spectrogram encoded in an RGB quaternion domain	Quaternion CNN	IEMOCAP (4): 70.46 EMO-DB (7): 88.78
Rethinking CNN Models for Audio Classification (2020) [63]	Mel spectrogram	Ensemble of ImageNet pre-trained DenseNets Deep CNN (Fine-tuned the AlexNet pre-trained on ImageNet)	ESC-50 (50): 92.89 UrbanSound8K (10): 87.42
SER Using Deep CNN and Discriminative Temporal Pyramid Matching (2018) [64]	Mel spectrogram		EMO-DB (7): 87.31 RML (6): 69.70 eINTERFACE05 (6): 76.56 BAUM-1s (6): 44.61

2.4 MULTIMODAL EMOTION RECOGNITION

There are several strategies to capture features for recognizing emotions. Multimodality, which refers to the use of multiple channels of information, is exploited to improve the accuracy of SER systems. By combining multiple modalities, such as speech, facial expressions, and speech transcriptions, as the figure 2.5 demonstrates, or even electrocardiogram readings [65], many researchers have been able to achieve improved results. However, multimodality increases complexity, since it becomes necessary to fuse the information across modalities in some way that achieves accurate emotional predictions.

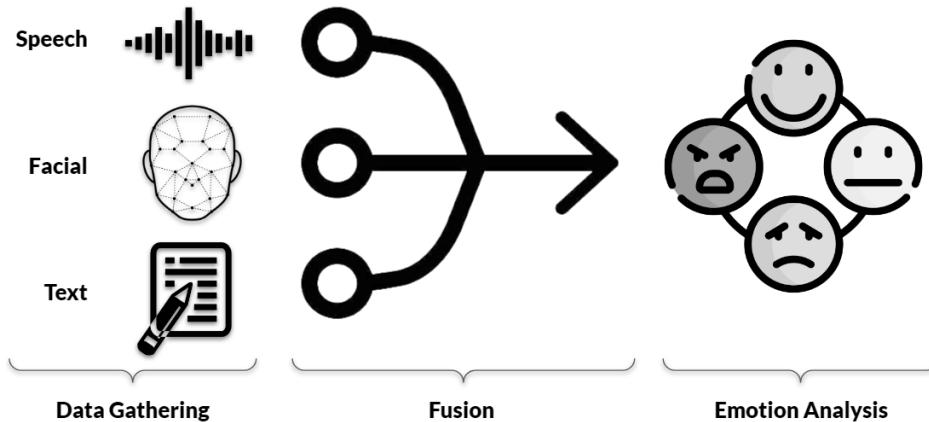


Figure 2.5: Multimodal emotion recognition illustration.

Four main methods of fusion can be used in multimodal systems:

- **Feature level:** This involves combining the features of different modalities and creating a new, high-dimensional feature set.
- **Decision level:** Each feature set from different modalities is given as input to different classifiers, and then the prediction results are combined.
- **Model level:** The goal of this method is to build a model that captures the complex relationships among the different modalities.
- **Hybrid:** This approach combines and takes advantage of the different fusion methods.

One of the strategies is to use speech transcriptions to understand the context of the utterances correctly and predict their intent. Human oral communication consists of both linguistic and acoustic content, any subtle change in these cues can alter the significance of an expression. However, this approach alone tends to fail, word meanings are very ambiguous, and it fails to capture some low-level features of the speech itself, therefore not achieving the best results. Thus, emotion and sentiment recognition could benefit from considering both linguistic and acoustic features. Bhaskar, Sruthi, and Nedungadi evidenced this improvement by using a hybrid approach based on text and speech mining [66]. Similarly, Tripathi, Kumar, Ramesh, *et al.* also corroborated this idea by using a model that fuses MFCC indicators with text transcriptions to predict emotions [67].

In a paper published in 2020 by Lu, Cao, Zhang, *et al.*, it was used a pre-trained end-to-end automatic speech recognition model to extract text features from audio data and combined them with other acoustic features. These features were then fed into a RNN with multi-head self-attention, which improved the accuracy of sentiment analysis using only audio data from 67.4% to 71.7% on the IEMOCAP dataset.

In a study by Handa, Agarwal, and Kohli (2021), a model was proposed for fusing facial and speech features using a SVM. The model employed an AlexNet and a deep LSTM network to extract the facial and speech features, respectively, which were then combined and fed into the SVM for classification. The model achieved SOTA results on the eINTERFACE'05 dataset for classifying seven emotions, with an overall accuracy of 76.61% [69].

Yan, Xue, Jiang, *et al.* in 2021, proposed a multi-tensor fusion network with cross-modal modeling with video, audio, and text cues for emotion recognition [70]. To perform their classification experiences, and obtained results that outperformed their current SOTA, achieving an F1 score of 81.0% and 81.3% on the CMU-MOSI and the CMU-MOSEI datasets, respectively.

2.5 EMOTION RECOGNITION SERVICES

The widespread use of video conferencing applications has increased significantly during the COVID-19 pandemic, as many people have been forced to communicate remotely. These systems typically use audio and video inputs, as well as text input, to facilitate communication. This growth of video conference usage increased the demand for tools that can improve the virtual communication experience.

In the customer service industry, video conferencing has become very popular, but it can be a challenging task, such as dealing with agitated customers. By providing real-time feedback on the emotions of participants, emotion recognition technologies can prevent problematic situations and evaluate the effectiveness and engagement of the customer service experience.

Buitelaar, Wood, Negi, *et al.* in 2018 summarized some other known emotion recognition services, that are helping to advance the field, along with their characteristics, displayed below in the table 2.4. It is visible that most of the services at the time of this study are paid and not open-source [71].

Table 2.4: Emotion recognition services for facial, textural, and speech contents [71].

Service	Modality	Open-Source	Free
IBM Watson Alchemy Language [72]	Text	No	No
Bitext [73]			
Synesketch [74]	Text	Yes	Yes
Microsoft Cognitive Services [75]			
IMOTIONS [76]	Facial	No	No
Affectiva Emotion API [77]	Facial	No	Free/Enterprise Editions
EmoVu [78]	Facial	No	No
Nviso [79]			
SkyBiometry [80]	Facial	No	Limited/Non-Free Editions
audEERING SensAI [81]	Speech	Yes	Free Research Edition
Good Vibrations [82]	Speech	No	No
Vokaturi [83]	Speech	No	Limited/Enterprise Editions

More recently, companies have been working on developing technology for recognizing human emotions, and they are, gradually, adding multimodality to perform emotion analysis. Affectiva [77], has added the audio modality to their emotion recognition system that initially was only with facial cues. The company primarily uses its technology for market research, but it has also been applied in the automotive industry to monitor drivers' emotions and cognitive states. Microsoft [75] has also integrated into their emotion API other modalities over time, such as text and speech. EmoVoice [84], is another framework that uses acoustic features of speech to recognize human behavior in real-time, without the need for speech transcription.

2.6 SOFTWARE TOOLS

For working and manipulating speech data and also machine learning techniques, different software tools are necessary. Generally, **Python** [85], is the most widely used programming language in all machine learning-related tasks, along with **Pip** [86] to install external resources, labeled as Python libraries. Important Python libraries in this area are **Numpy** [87] and **Pandas** [88], which provide high-level mathematical functions for dealing with big vectors and matrices, **Matplotlib** [89] and **Seaborn** [90], for statistical data visualization, **Keras** [91] and **scikit-learn** [92], for implementing and manipulating machine learning algorithms.

For extracting and processing speech, there are several tools available in Python, such as:

- **Librosa** [93]: provides tools for music and audio analysis. It helps to visualize audio signals and extract features using signal processing techniques;

- **OpenSmile** [94]: open-source toolkit for audio analysis, processing, and classification that focuses on speech and music applications;
- **COVAREP** [95]: an open-source repository of advanced speech processing algorithms that are useful for tasks such as speech analysis, synthesis, and conversion;
- **Spafe** [96]: simplifies the process of extracting features from mono audio files and includes various filter bank modules and other spectral statistics;
- **Pydiogment** [97]: makes it easy to augment audio files by generating multiple versions with different speeds, tones, and other modifications;
- **Noise Reduce** [98], [99]: helps to remove noise from time-based signals such as speech.

Frequently, Voice Activity Detection (VAD) algorithms are employed in speech processing systems. These algorithms are designed to be highly accurate and efficient in detecting when a user is speaking and filtering out background noise and other non-speech signals. Examples of VAD toolkits are:

- **Silero** [100]: intended to be used in real-time speech processing applications, it is optimized for performance on a wide range of devices, including mobile phones and other embedded systems;
- **Py-webrtcvad** [101]: python interface to the Google's WebRTC VAD project, which is an open-source project that provides web browsers and mobile applications with real-time communication capabilities.

2.7 SUMMARY

Through the SOTA research, it is possible to state the significant progress over time of not only SER, but emotion recognition systems in general. Here are a few key takeaways that emerged from it:

- Due to the subjectivity of emotions, the lack of consensus on how to label them, and the need for a large and diverse emotional corpus (e.g. different contexts, genders, languages, etc.), makes the whole process of creating and evaluating an emotion recognition system a very challenging task.
- Audio feature engineering is a critical aspect of SER due to the large variety of features available and the need to carefully select them, including prosodic, spectral, and voice quality features.
- Traditional feature-based approaches use more classical models and focus mainly on feature selection methods.
- DL strategies with automatic feature extraction capabilities have, in general, outperformed the traditional acoustic feature extraction approach. The most popular models for this are CNN, RNN (LSTM), and hybrid CNN-RNN architectures, with Mel spectrogram as input for them.
- The field of affective computing, which originally focused mainly on using facial cues to recognize emotions, has seen success in using multiple channels of information (multimodality). This shift has allowed for more accurate emotion recognition, and many current solutions in the market are taking advantage of this approach.
- There are still areas in which emotion recognition systems need to be improved, such as being able to handle a wider range of input signals and increase reliability. For example, most systems overlook the case of input channels with noise, where the system should weigh each differently, and also, research for online classification contexts is still underdeveloped.

Overall, there is a significant amount of ongoing research and development that is taking place on emotion recognition systems and has successfully advanced further their capability and reliability, however, there remain a lot of limitations to overcome and areas to improve.

3

CHAPTER

Methodology

3.1 PROPOSAL

In this chapter, we propose our plan and schedule to achieve the goals of building a SER system, integrating another modality, and evaluating its effectiveness in a video conference context. Figure 3.1 illustrates our research plan and schedule through a Gantt chart.

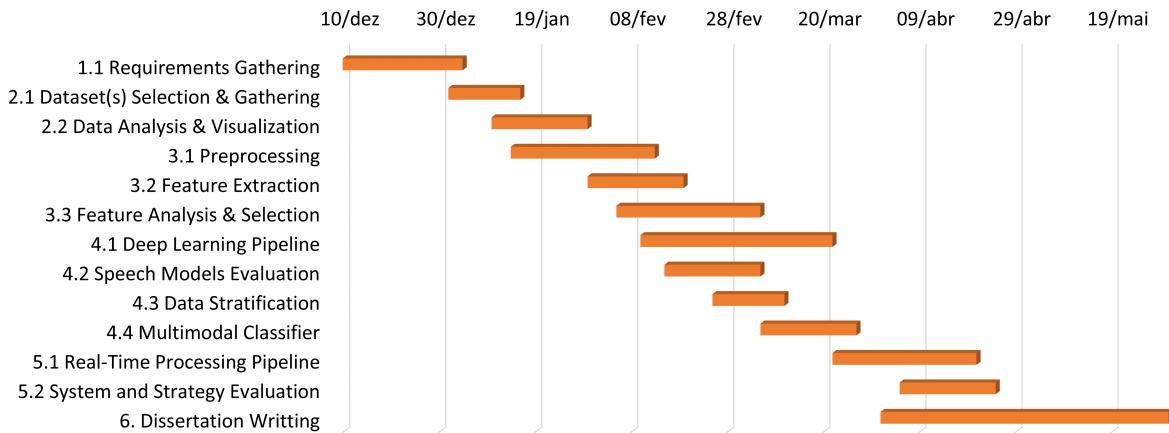


Figure 3.1: Gantt chart with the proposed work plan for this research.

3.1.1 Requirements Gathering

Initially, we conducted a requirements-gathering process to understand the needs, expectations, and constraints of our emotion recognition system. We identified specific use cases in which the system would be applied, and from there we were able to determine various requirements, including ethical considerations and technical requirements. These requirements served as the foundation for our defined objectives and strategies.

3.1.2 Dataset

Dataset(s) Selection & Gathering

The next step is the selection of one or more audiovisual datasets that, in line with the SOTA research, are preferably large and diverse (e.g., by featuring individuals of both genders, different scripts, contexts, etc.).

Data Analysis & Visualization

Data analysis and visualization will allow us to explore, understand and display the data. Using various techniques such as statistical analysis, machine learning, and computational methods, we aim to extract information and then create graphical representations of it, in a way that allows us to better understand the dataset.

3.1.3 Audio Feature Engineering

Once the dataset(s) have been selected, following the traditional feature-based approach, audio feature engineering will be required. In the context of machine learning, audio feature engineering involves preprocessing, extracting, and selecting relevant features from audio signals to build effective systems.

Preprocessing

Framing. Speech framing, also called signal segmentation, is the process of dividing continuous speech signals into fixed-length segments. Speech typically stays consistent for a relatively short period, such as 20 to 30 ms. By dividing the speech signal into frames, local features can be extracted. Additionally, the relationship and information between the frames can be preserved by deliberately overlapping 30% to 50% of these segments. After creating the speech frames, researchers often apply a windowing function to smooth the edges and improve the time-frequency resolution of the signal [102].

Noise reduction. Noise can have a significant negative impact on the quality of an audio signal by masking or distorting important features of the speech. It can also introduce variability into the audio data, making it harder for the machine learning model to generalize to new situations. The most successful methods for noise reduction are the Minimum Mean Square Error (MMSE) and Log-spectral Amplitude MMSE estimators [103].

Normalization. Scale differences in features' values may affect the models' performance, especially those based on distance metrics to parameter's estimation. So, it is advised to normalize features to prevent this effect. This can lead to more accurate and consistent results when applying the model to new data. The most widely used normalization techniques are z-normalization (also known as standardization) and min-max normalization [102].

Voice Activity Detection. VAD is a technique used in speech processing to identify periods of speech in an audio signal. By focusing on analyzing the voiced speech and ignoring noisy or silent periods, VAD can improve the accuracy of a system by reducing the influence of non-speech segments on the analysis. Furthermore, it can decrease the computational cost by reducing the amount of data to be processed. Commonly used methods for VAD include zero crossing rate, short-time energy, and auto-correlation [102].

Figure 3.2 demonstrates an instance where a VAD system is applied to a speech signal, in this case, it is processing chunks with a duration of 1 second. It reduces the original signal of 10 seconds by half by identifying 5 voiced segments that are then fed to a SER model [104].

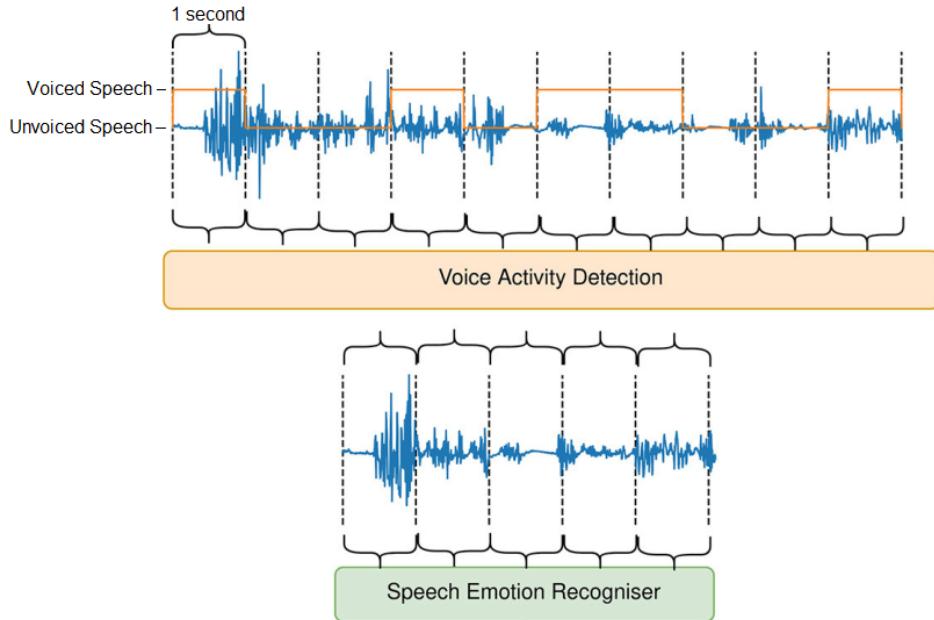


Figure 3.2: Voice Activity Detection employment on an audio signal [104].

Feature Extraction

Various techniques can be used to extract features from audio signals, such as time-domain analysis, frequency-domain analysis, and signal processing algorithms. These techniques can be used to extract features such as spectral features, temporal features, and statistical features. When studying the features extracted from speech audio signals, it is common to consider both global and local features [3]. Some emotions are more prominent at the beginning or end of a speech, so it is frequent to consider both global and local features to fully capture the temporal and emotional content of the signal.

Feature Selection

In order to achieve optimal performance, it is important to carefully select the features used in any model. Including redundant or uninformative features can degrade performance. By selecting and reducing the number of features, it is possible to maintain or even improve performance while using fewer resources and avoiding overfitting, also it becomes easier to understand the model's decision-making process and which features are most important.

3.1.4 Models Implementation

Deep Learning Pipeline

In addition, we will experiment and compare the effectiveness of automatic audio feature extraction-based approaches to SER, using either the raw audio signals, MFCC, or Mel spectrograms as input to deep learning models, some transfer learning techniques will also be explored.

3.1.5 Speech Models Evaluation

After we have chosen the relevant features through the feature selection process and completed the deep learning pipeline, we will test various models, and measure their performance by using appropriate metrics, to then identify the best classifiers for the traditional feature-based and DL-based approaches.

3.1.6 Data Stratification

Data stratification refers to a process of dividing a dataset into subgroups, based on certain properties. We will test different ways of grouping the labeled data and interpret their effects on the models' predictions, for example, by relabeling the angry and sadness samples as *negative* samples. By doing this, it allows discovering potential biases in both the classifiers and the dataset(s), such as the individual's age, gender, and even the evaluation of the samples given by juries.

Multimodal Classifier

After selecting the most suitable models, a fusion model will be developed by combining it with a facial emotion recognition model. This fusion can improve the accuracy of emotion predictions by leveraging information from both modalities. It can be trained to weight the contributions of each modality differently or to use a weighted combination of the outputs from the individual models based on external factors such as brightness for video and noise for audio.

3.1.7 Deployment

Real-Time Processing Pipeline Creation

The most suitable model for offline and online scenarios will be analyzed and tested in a real-time scenario using an audio segmentation pipeline with VAD.

System and Strategy Evaluation

We will evaluate the performance of the fusion model and compare it to the performance of the individual models to determine the potential improvement in emotion prediction.

3.1.8 Dissertation Writing

To conclude the work plan, we will focus on a comprehensive discussion of our research, including any potential challenges and limitations that we encountered and the steps we took to address them.

3.2 ETHICAL PROCEDURES AND CONCERNs

There are several ethical concerns when building emotion recognition systems, and in our case, we will address those related to SER systems. One of them is that they may not be accurate in recognizing emotions for people with certain characteristics (age, gender, accents, speech disabilities, etc.), which is why we will address and quantify some potential biases in our models.

Another concern is that these systems may be used to monitor or control persons without their consent. Our models will only be in use on individuals that fully consented to submit their data for emotional evaluation purposes.

Additionally, there may be privacy concerns about personal data collected by these systems being vulnerable to hacks or used for unintended purposes. Our processing pipelines will only utilize the participant's streaming data for the SER models input, and, will not store or use the sensitive data for any other purposes than the emotional analysis. We also decided to not use speech transcriptions, as it raises even greater privacy issues, due to capturing the actual words spoken, and can reveal confidential data such as personal, financial, or business information.

In summary, we intend to be transparent about the system's capabilities and limitations and obtain informed consent before collecting and using any personal data.

4

CHAPTER

SER Development

4.1 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.

4.1.1 eINTERFACE'05

The eINTERFACE'05 emotion database [24] was designed and collected during the eINTERFACE'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 4.1: eINTERFACE'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.

4.1.2 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 VAD emotion model. Similar to the development dataset, this data was collected using emotion elicitation techniques such as improvisations and scripts. Figure 4.1 from the research article ??, demonstrates that there is a similar amount of annotated labels on scripted and spontaneous sessions on this dataset.

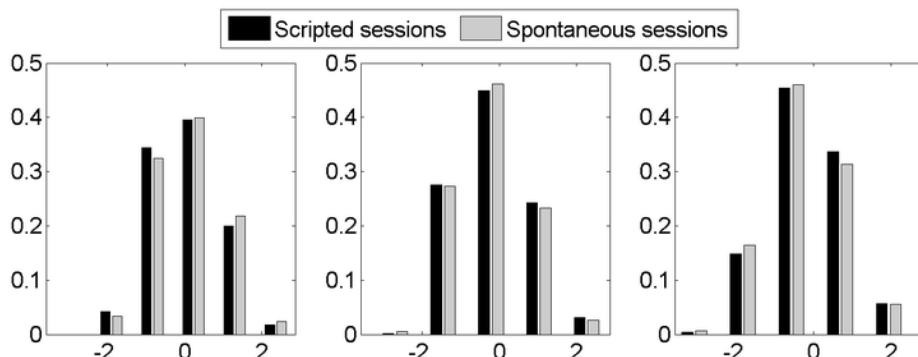


Figure 4.1: Distribution of the emotional content of the IEMOCAP corpus in terms of (a) valence, (b) activation, and (c) dominance. The results are separately displayed for scripted (black) and spontaneous (gray) sessions.

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 4.2, ending up with a total of 5531 audio files, recorded with a sample rate of 16000 Hertz.

Table 4.2: 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.

4.2 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.

4.2.1 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

4.2.2 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.

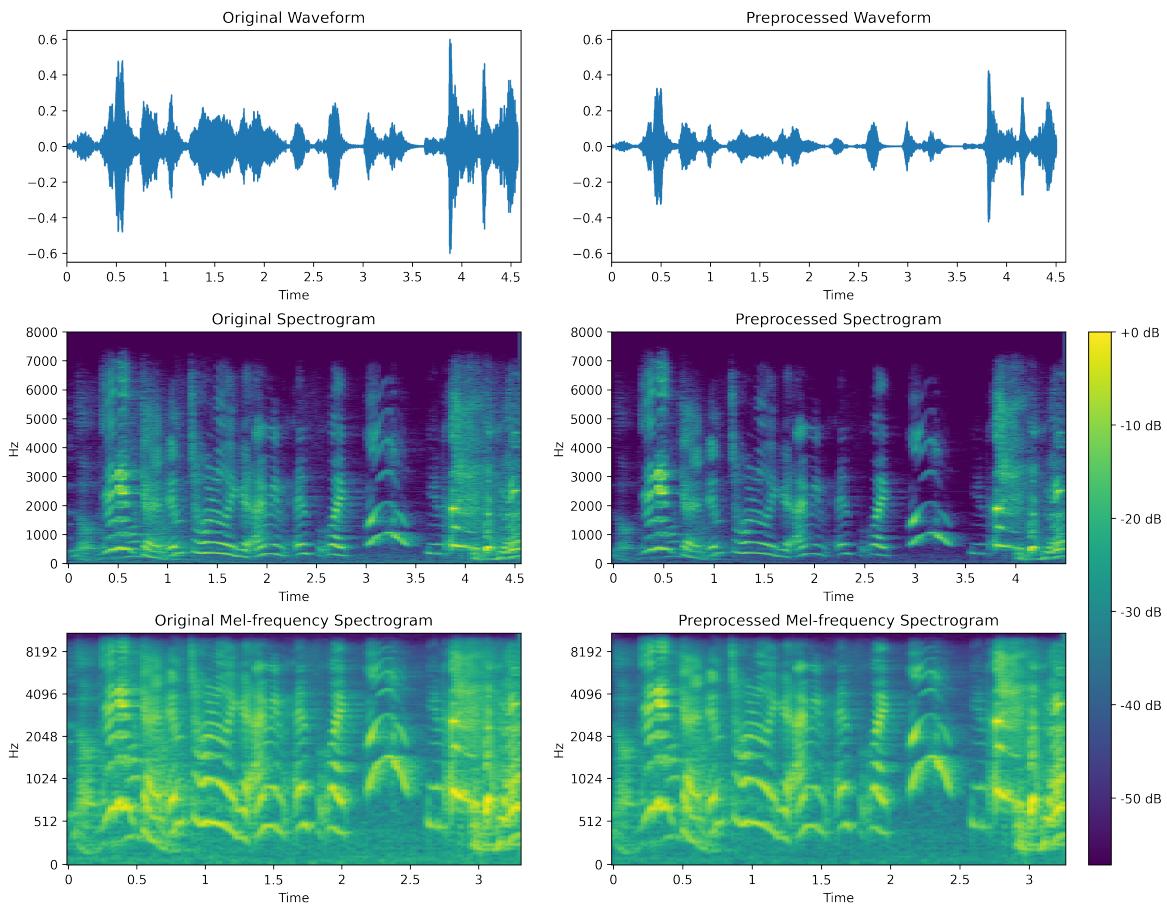


Figure 4.2: Visual representations of audio features before and after preprocessing.

4.3 TRADITIONAL FEATURE-BASED SER

4.3.1 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 eINTERFACE'05 dataset, which were subsequently processed using statistical metrics. The extracted features and associated metrics are summarized in Table 4.3, having in total, extracted 327 features.

Table 4.3: Extracted Audio Features and Statistical Functions Applied to Them

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

¹Custom function detailed on the Feature Selection section.

²Only for the MFCCs.

4.3.2 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 .1 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 4.3, 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. Consequently, this metric was also tested and applied to every other audio feature.

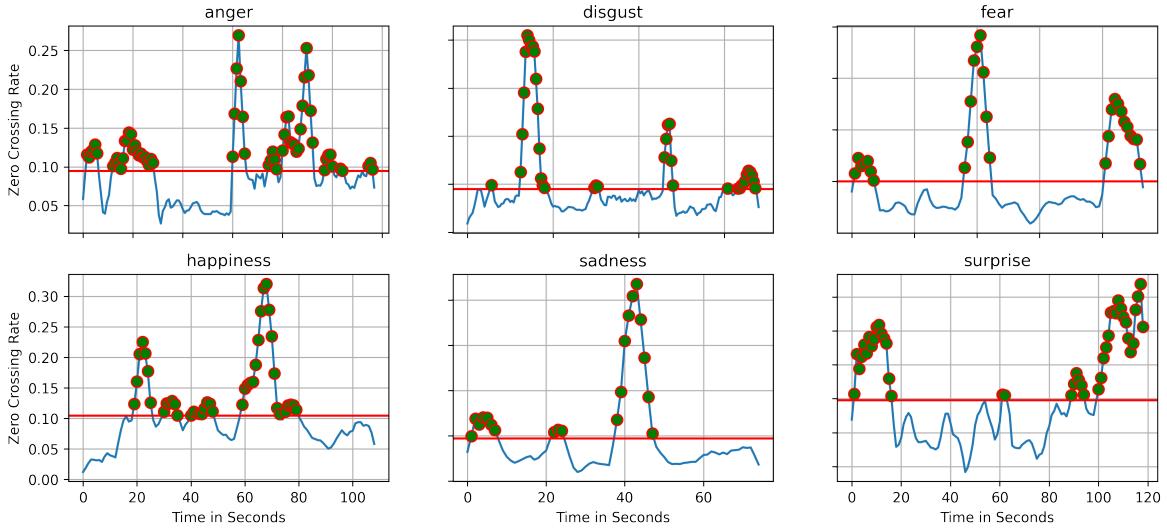


Figure 4.3: Zero crossing rate wave plot annotated with spikes.

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

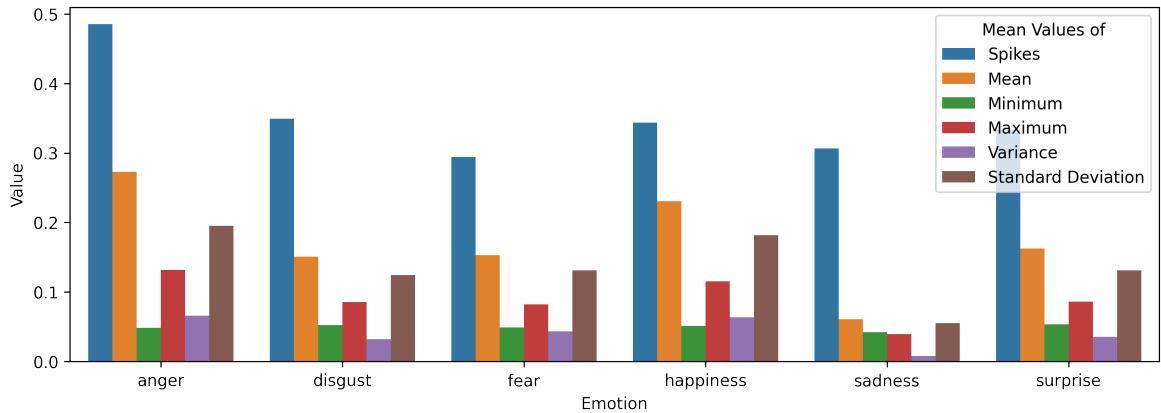


Figure 4.4: Bar plots mean for metrics used on the mel-scaled spectrogram feature

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 4.5 is an excerpt of the figure .10 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.

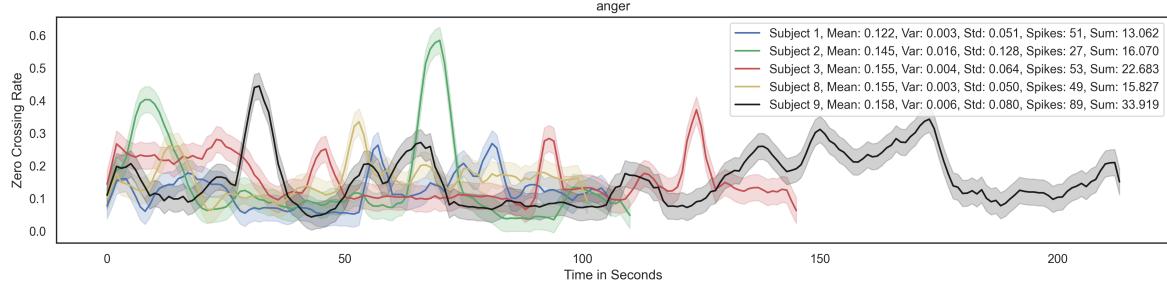


Figure 4.5: Zero crossing rate wave plot with a surrounding area of five male subjects for the same utterance with the anger emotion.

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 4.6 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.

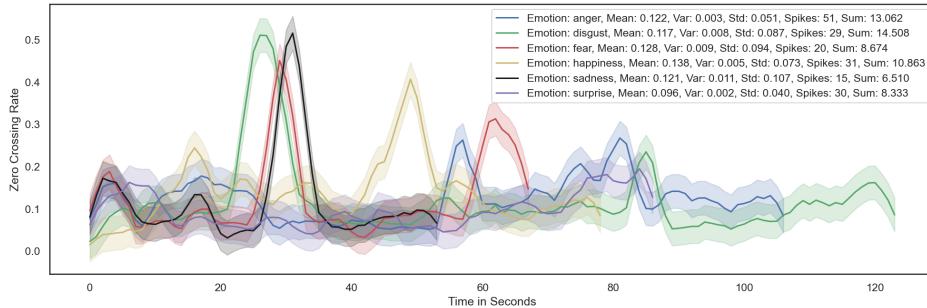


Figure 4.6: Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions.

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 4.7 shows an example of this type of plot for the mean zero crossing rate value across 50 speech utterances for all emotions.

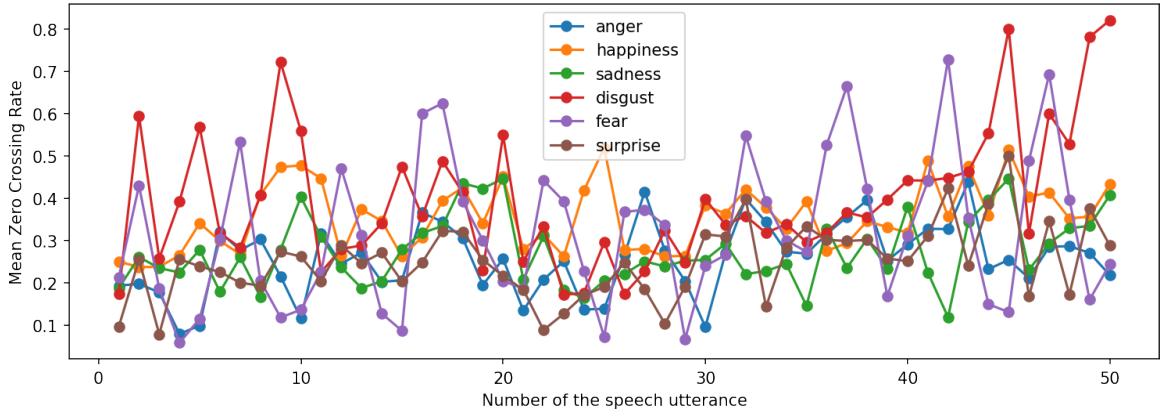


Figure 4.7: Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions

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 4.8, which displays the mean zero crossing rate feature for all emotions and different subjects.

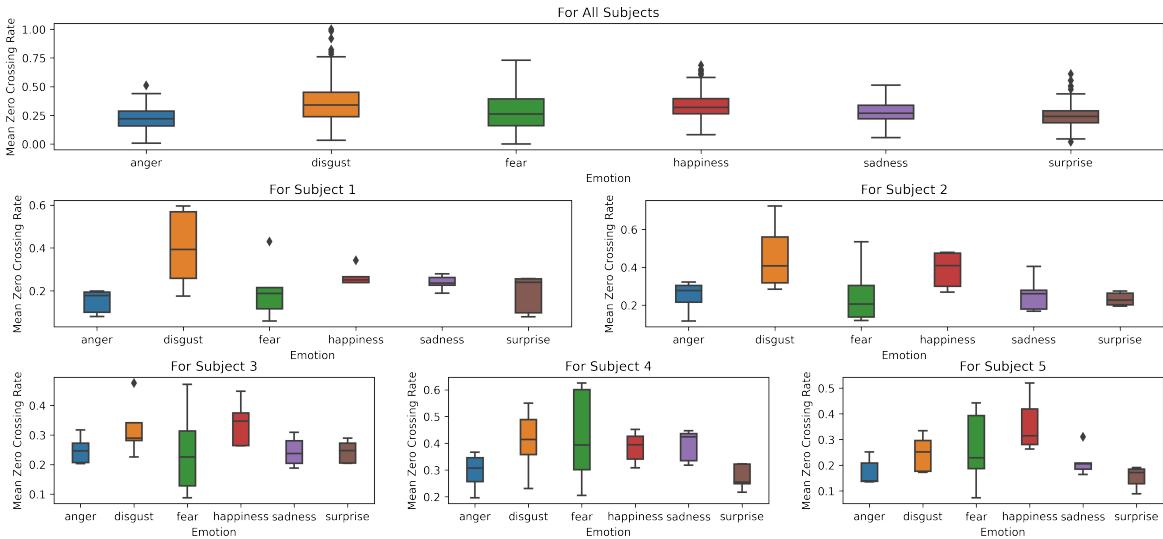


Figure 4.8: Zero crossing rate mean values box plot for all emotions and different subjects

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.

4.3.3 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

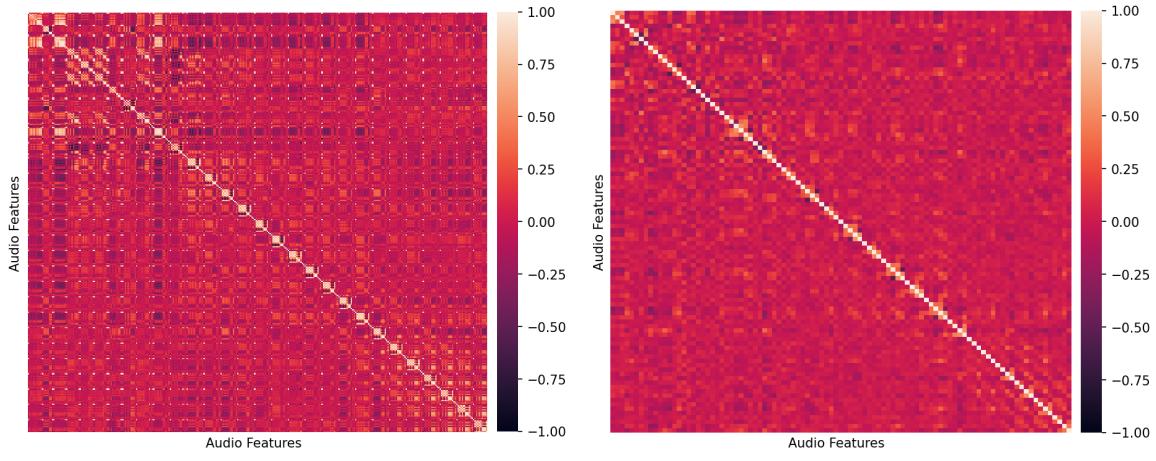
the given task. The process of feature selection is aimed to improve the accuracy of the model and reducing the problem's complexity by removing redundant or irrelevant features.

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 230 features, leaving 97 features for subsequent analysis. The correlation matrix after the feature selection process is presented in figure 4.9b.



(a) Correlation Matrix of all the Features.

(b) Correlation Matrix after High Correlation Elimination.

Figure 4.9: 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 4.4.

Table 4.4: 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.

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 Ridge 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 involves finding a maximum and multiplying it by a threshold value of 0.99 to balance accuracy with the number of removed features. Figure 4.10 displays the mean accuracies obtained at each step and the chosen furthest highest accuracy.

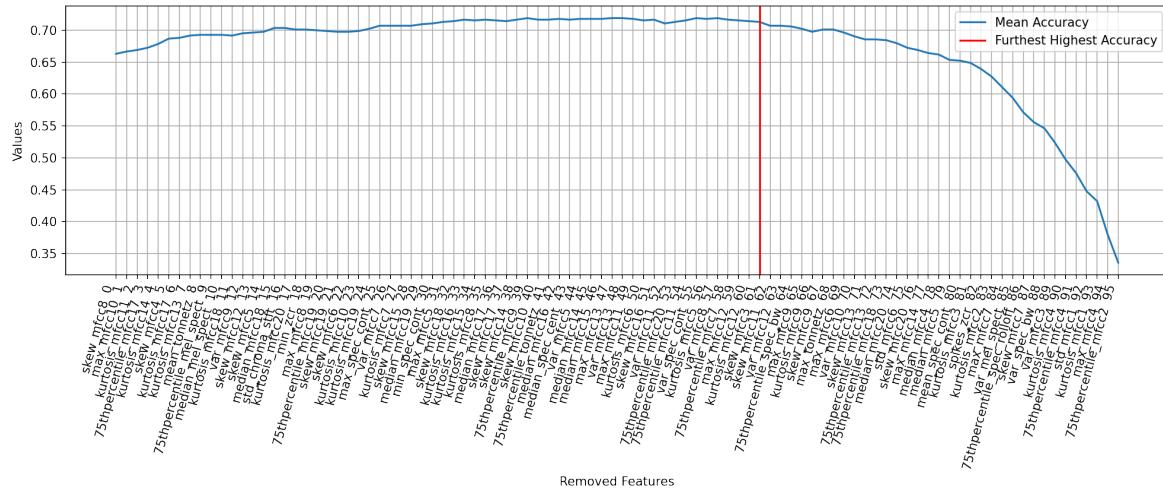


Figure 4.10: Sequential Feature Selection with Backward Propagation using the Mean Accuracy as the Selection Method.

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 4.5.

Table 4.5: 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

Feature Selection Evaluation

To assess the feature selection quality on the development dataset, we trained and evaluated the predictions of a Ridge Classifier model with different sets of features, using accuracy and macro f1-score as the evaluation metrics. The results are presented in Table 4.6. Additionally, we plotted the confusion matrix of the predictions, which can be found in Appendix .5.

Table 4.6: Evaluation Metrics of a Ridge Classifier Predictions Using Different Sets of Features

Feature Selection Method	N. ^o of Features	Accuracy	Macro F1
None	327	0.640	0.639
High Correlation Elimination	97	0.652	0.649
High Correlation Elimination & Backward selection	35	0.713	0.710

The results indicate that there is not a significant change in the metrics between the initial set of extracted features and the 97 features obtained after the high feature correlation elimination, despite the elimination of 70% of the initial set. Moreover, applying a backward selection technique to these

97 features yielded optimal results, slightly improving both metrics and reducing the original set by approximately 90%.

4.3.4 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.

Classifier Selection

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.

The results obtained from the tested models were compiled and exhibited in Table 4.7. 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 4.7: Tested Classification Models Performance on IEMOCAP.

Model	Accuracy	Macro F1	Precision	Recall	MCC
AdaBoost	56.92±1.56	57.4	58.66	56.98	0.416
Random Forest	56.54±1.69	57.01	58.25	56.60	0.41
Histogram Gradient Boosting	55.58±0.83	56.38	57.08	56.0	0.397
Balanced Random Forest	54.64±0.74	54.97	54.65	57.59	0.402
LSTM	49.92±0.48	50.32	51.16	50.58	0.324
CNN	48.87±1.99	49.28	51.71	48.36	0.302
Linear Discriminant Analysis	47.55±0.9	48.24	48.79	47.94	0.288
Ridge	47.01±0.82	47.43	47.93	47.28	0.281

Table 4.8: SOTA Traditional Classification Models Performance on IEMOCAP.

Model	Input	Accuracy
Dilated Residual Network Li_2019	Audio Features	67.4
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, this process can potentially worsen the predictions by removing necessary information from the audio signal that might be required for accurate classification. In some cases, noise may be part of the signal of interest, and removing it may cause the algorithm to misinterpret the remaining signal.

The state-of-the-art results presented in Table 4.8 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 compared to the chosen strategies.

Conclusion

Upon analyzing the results of the tested models, AdaBoost with Random Forest as the base estimator is the best candidate thus far in the research, reaching an average accuracy of 56.92% while utilizing only 35 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 and is presented on the following Code Snippet 2.

```
AdaBoostClassifier(  
    estimator=RandomForestClassifier(  
        bootstrap=False, max_features=6, min_samples_leaf=2, min_samples_split=3, n_estimators=512  
    )  
)
```

Code Snippet 2: Python code for the selected AdaBoost classifier using the traditional-based SER approach.

4.4 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.

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.

These three used features are displayed in figure 4.11.

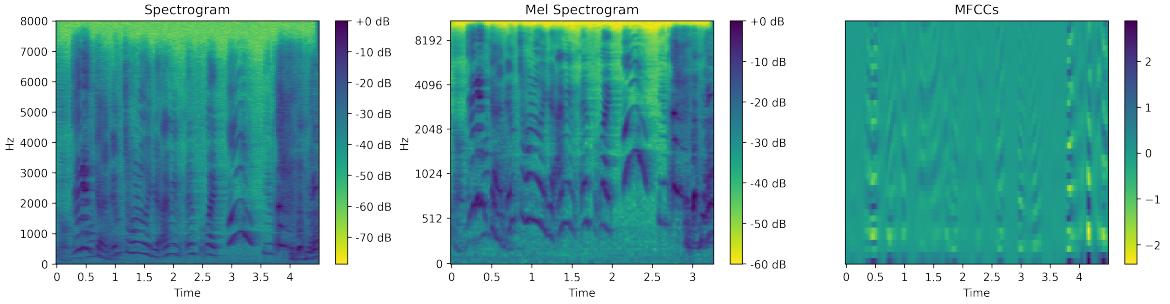


Figure 4.11: Graphical representations of the features used as input for the DL classifiers.

4.4.1 Spectrogram as a Feature

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 1025×188 . The amplitude spectrogram was converted to a dB-scaled spectrogram, which was then used for the PNG file.

4.4.2 Mel Spectrogram as a Feature

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 256×188 . The dB-scaled Mel Spectrogram was also used for the PNG file.

4.4.3 MFCC as a Feature

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 40×188 .

4.4.4 Classifiers

4.4.5 Numeric Data Classification

4.4.6 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 $224 \times 224 \times 3$ 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.

4.5 CLASSIFIERS RESULTS AND DISCUSSION

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. Table 4.9 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 4.9: 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

Appendix

.1 AUDIO AND DESCRIPTIVE FEATURES STUDY

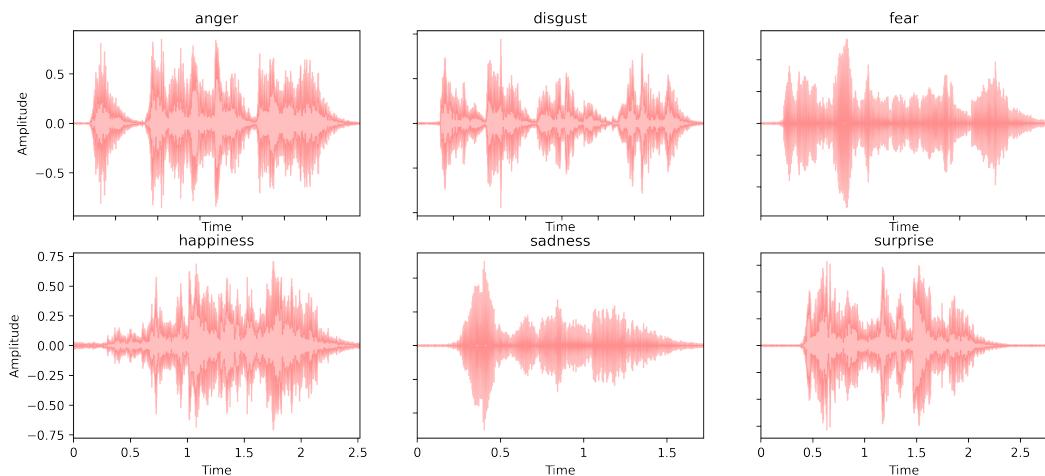


Figure .1: Audio Signal Wave Plots of one Audio Segment for All Emotions

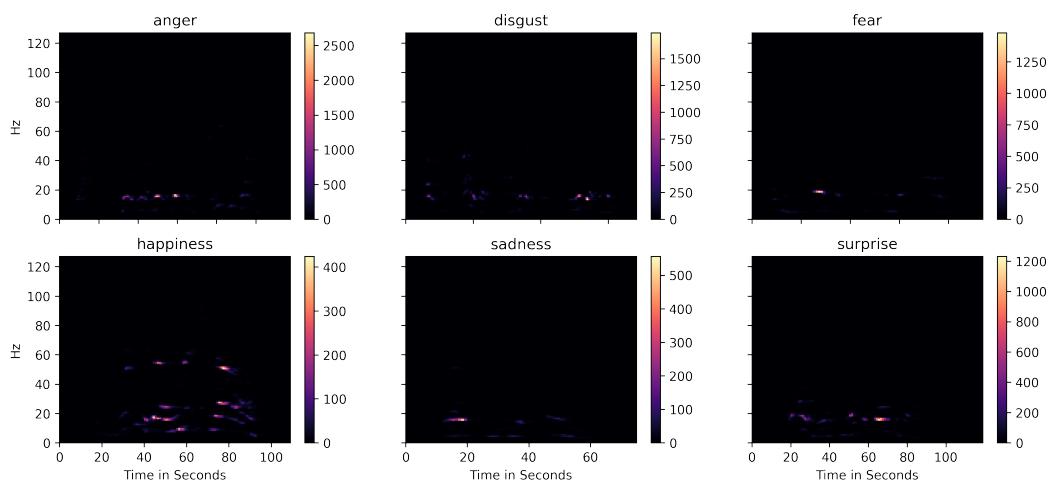


Figure .2: Log Mel Magnitude Spectrograms of one Audio Segment for All Emotions

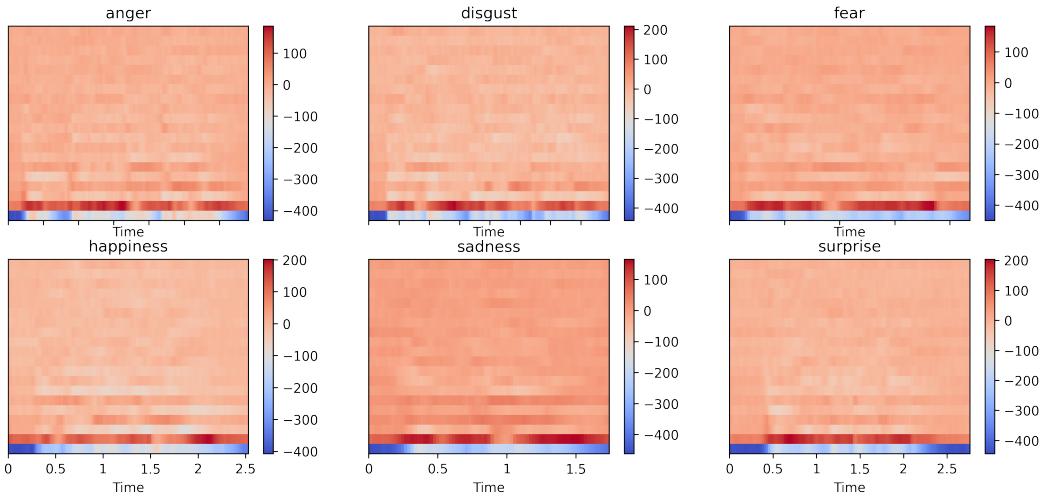


Figure .3: Mel-Frequency Cepstral Coefficients Spectrogram of one Audio Segment for All Emotions

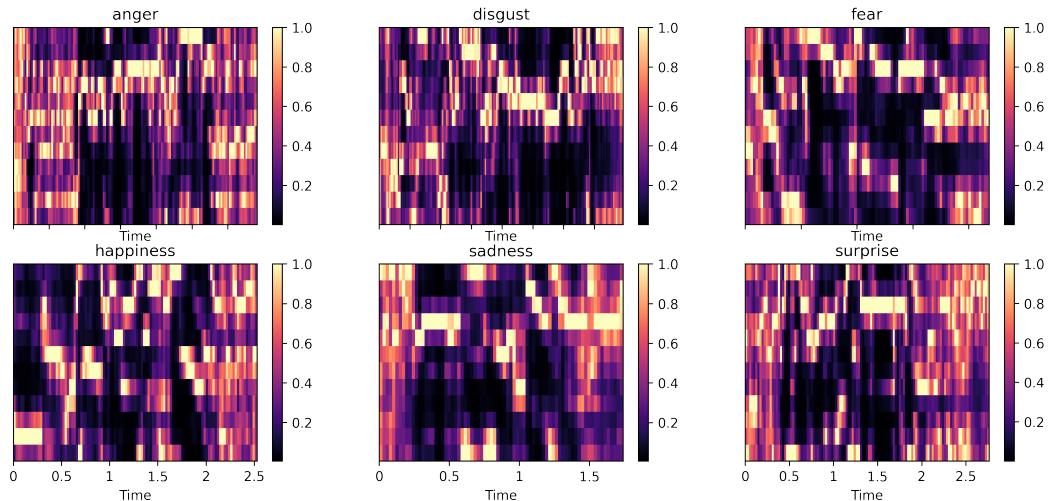


Figure .4: Chromogram Spectrograms of one Audio Segment for All Emotions

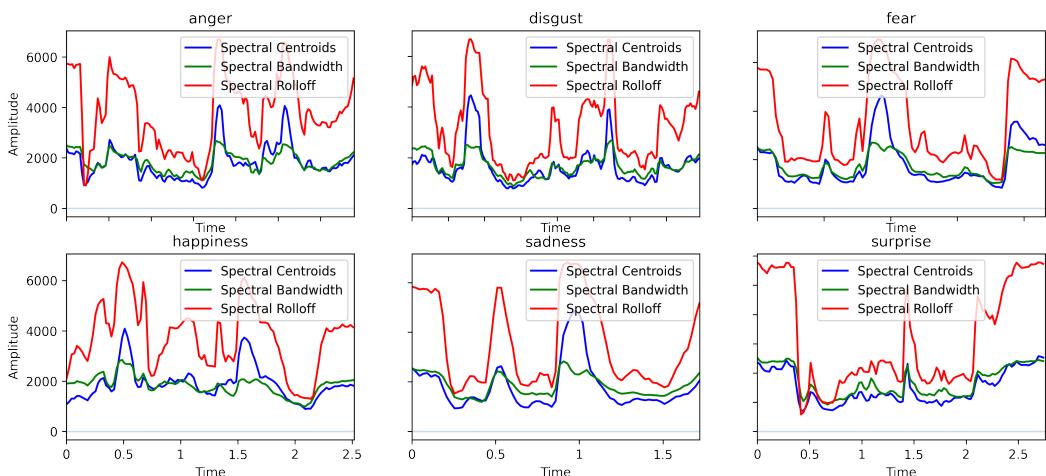


Figure .5: Spectral Wave Plots of one Audio Segment for All Emotions

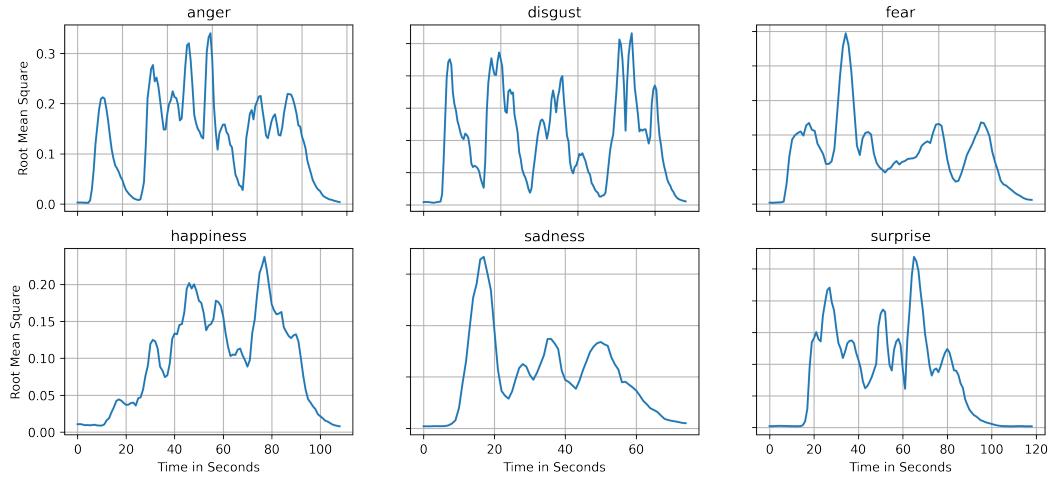


Figure .6: Root-Mean-Square Energy Wave Plots of one Audio Segment for All Emotions

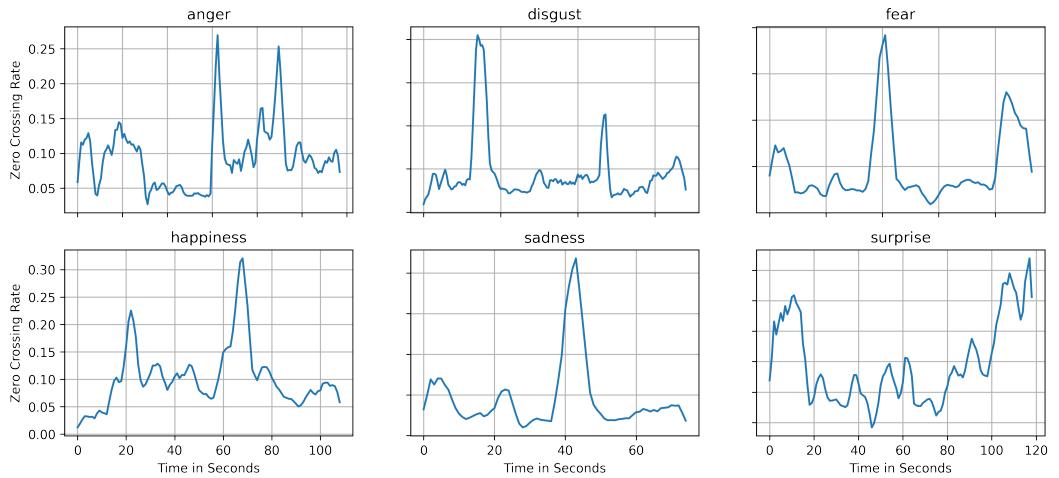


Figure .7: Zero Crossing Rate Wave Plots of one Audio Segment for All Emotions

.2 FEATURES MEAN VALUES OVERVIEW

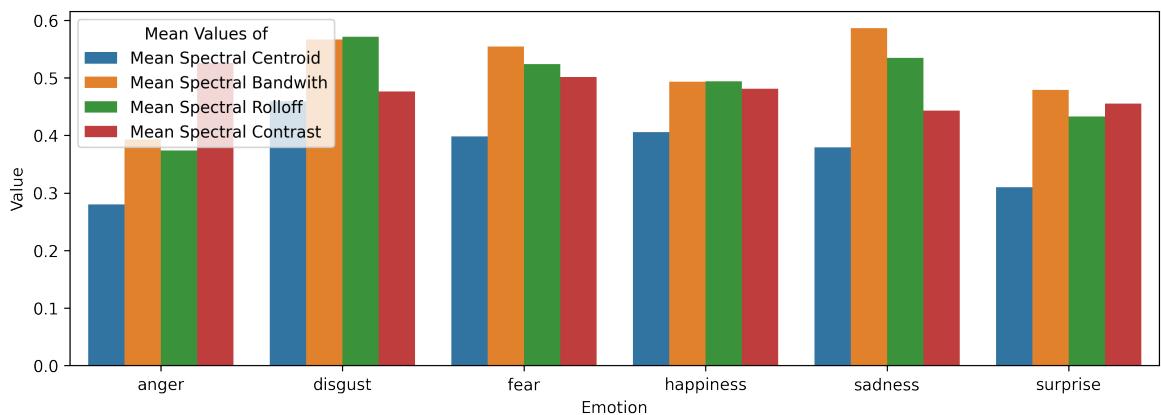


Figure .8: Bar plot with the mean values of the mean spectral centroid, bandwidth, roll-off, and contrast features.

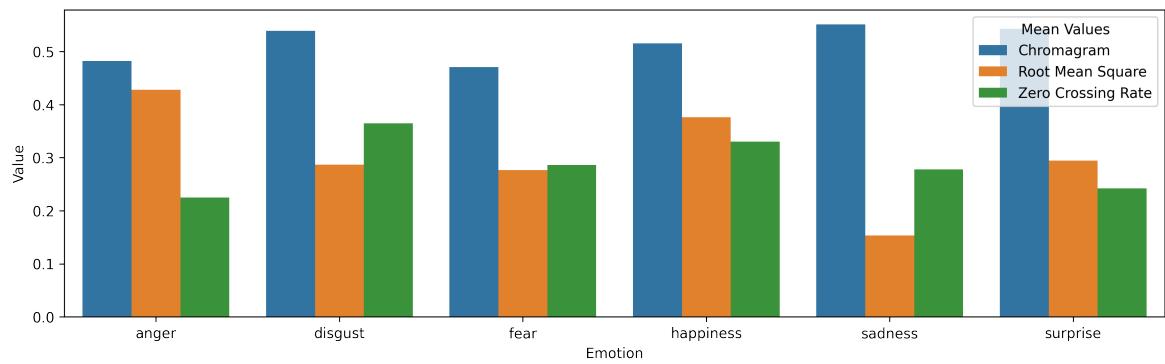


Figure .9: Bar plot with the mean values of the mean chromogram, root-mean-square and zero-crossing rate features

.3 FEATURE INTERCEPTION STUDY

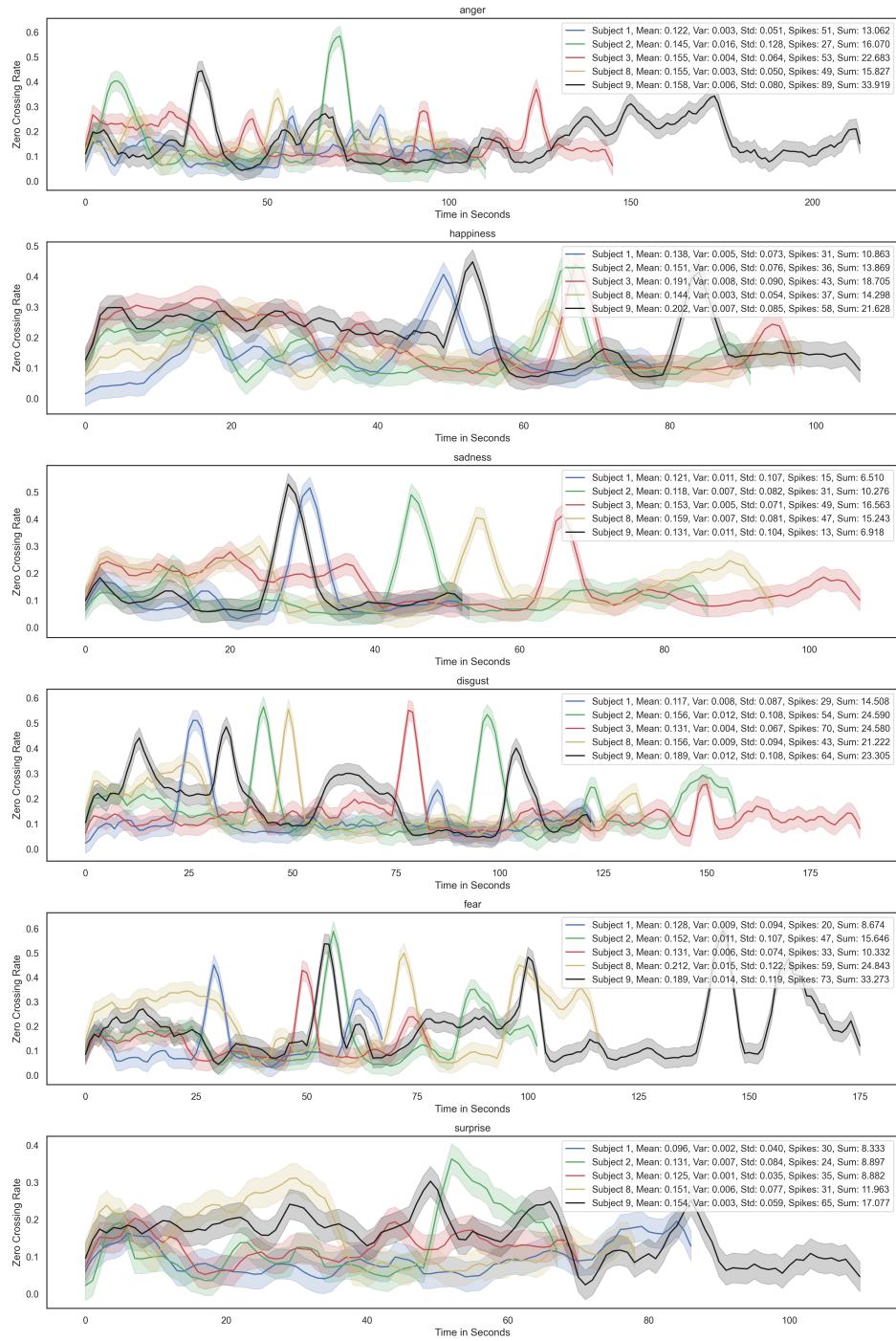


Figure .10: Zero Crossing Rate Wave Plots with a Surrounding Area of Five Male Subjects and the same Sentence for All Emotions

.4 FEATURE VARIATION PLOTS

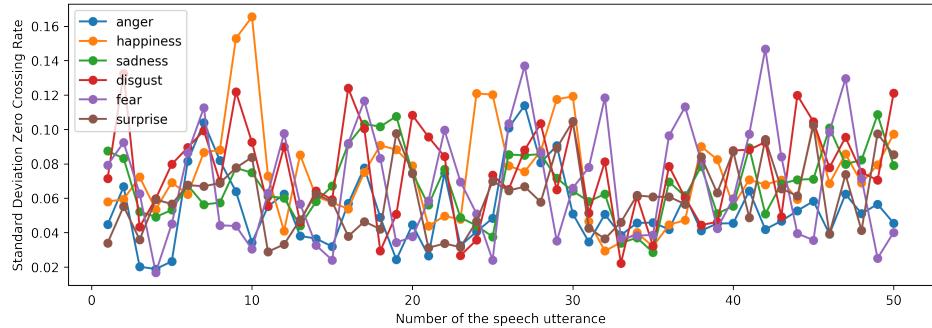


Figure .11: Zero crossing rate standard deviation values variation plot along 50 audios of speech utterances for all emotions

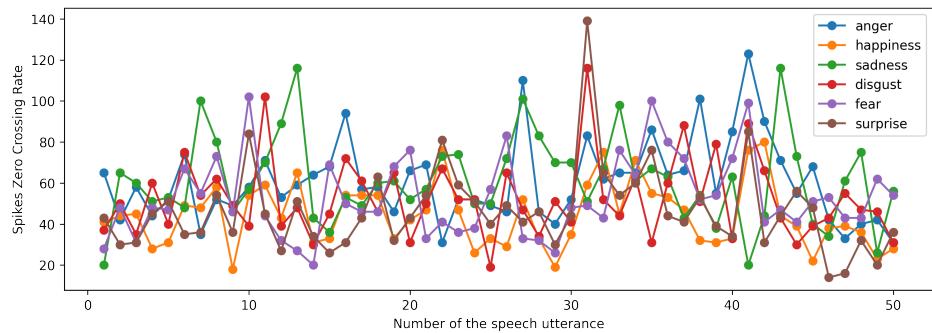


Figure .12: Zero crossing rate spikes values variation plot along 50 audios of speech utterances for all emotions

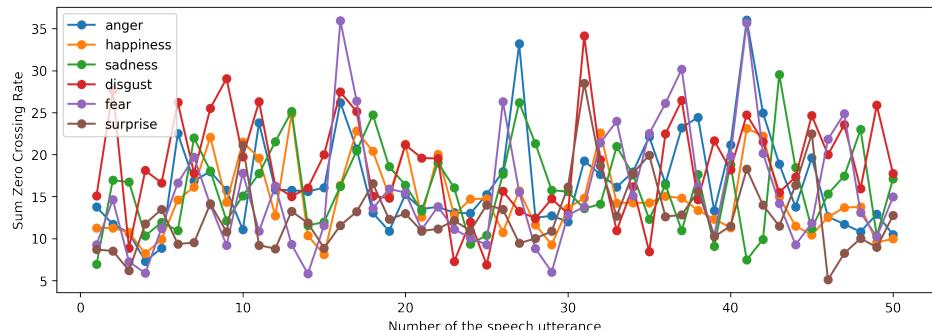


Figure .13: Zero crossing rate sum values variation plot along 50 audios of speech utterances for all emotions

.5 CONFUSION MATRICES

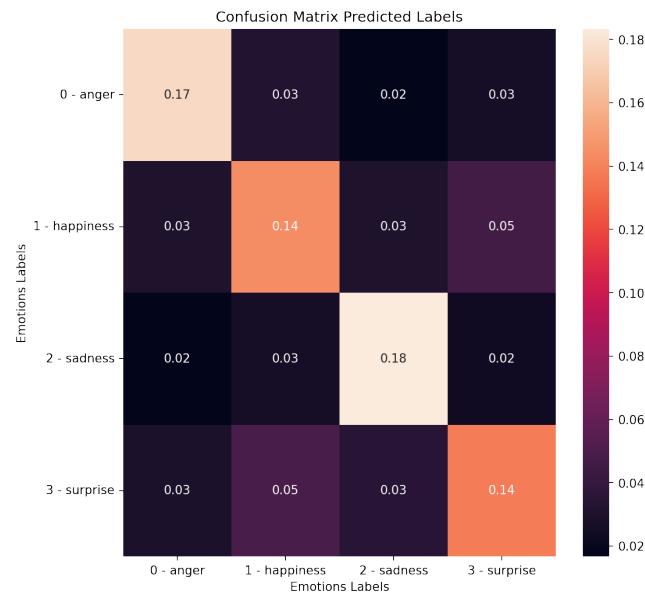


Figure .14: Confusion Matrix of a Ridge Classifier Trained using all 327 Extracted Features

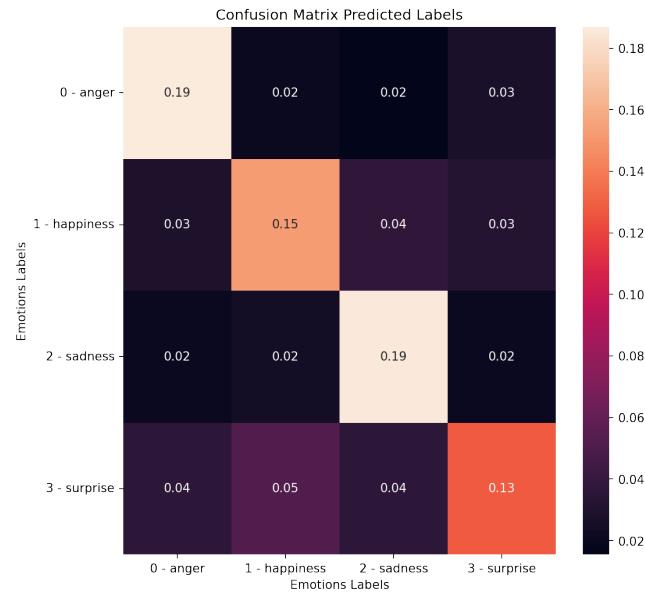


Figure .15: Confusion Matrix of a Ridge Classifier Trained using 97 Features After High Correlation Feature Elimination

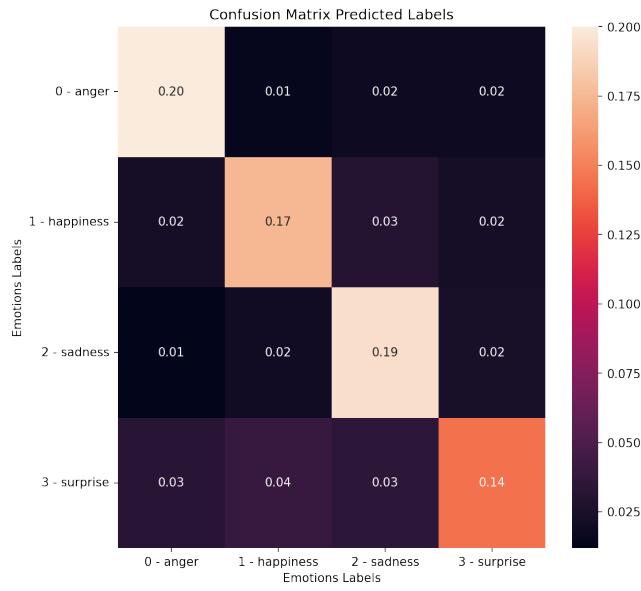


Figure .16: Confusion Matrix of a Ridge Classifier Trained using 35 Features After High Correlation and Backward Selection

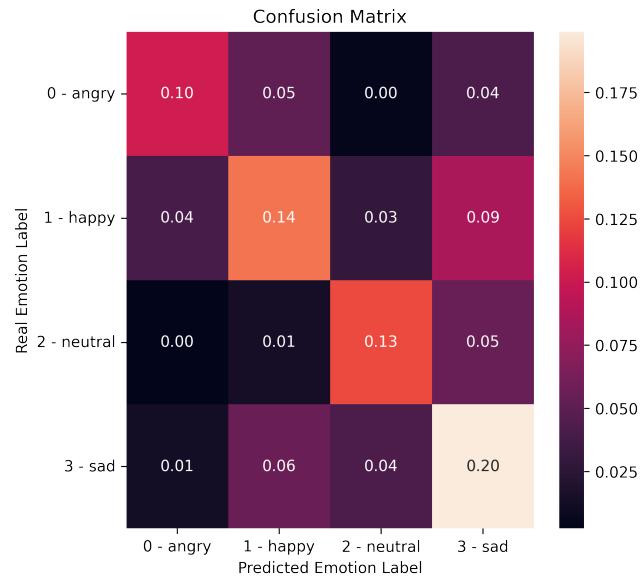


Figure .17: Confusion Matrix of an AdaBoost Classifier.

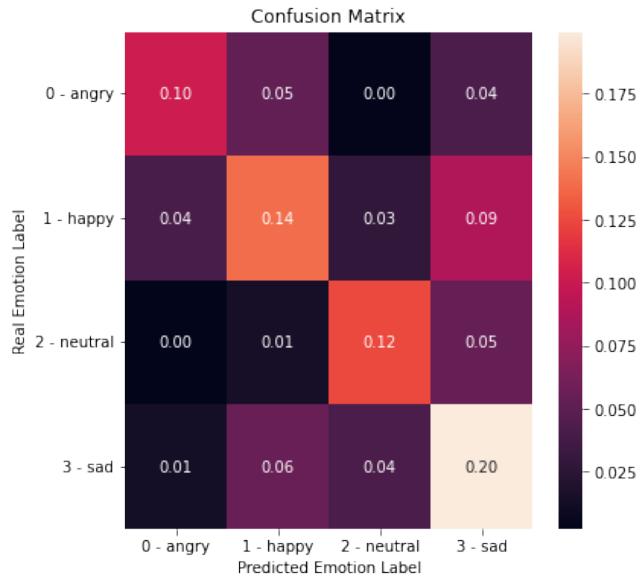


Figure .18: Confusion Matrix of a Random Forest Classifier.

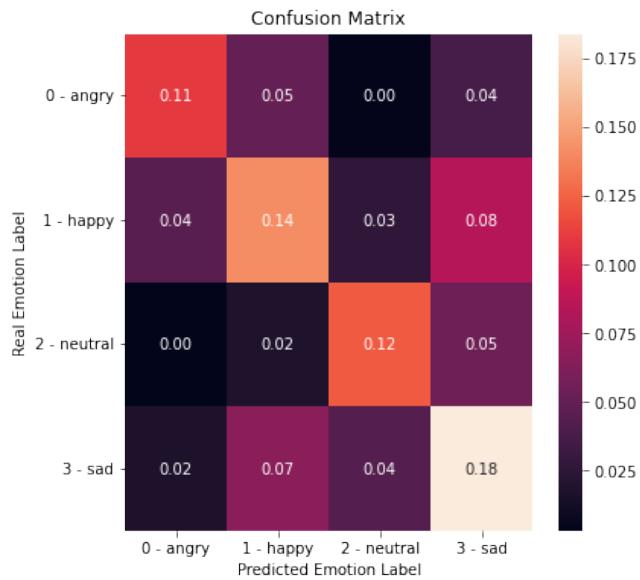


Figure .19: Confusion Matrix of a Histogram Gradient Boosting Classifier.

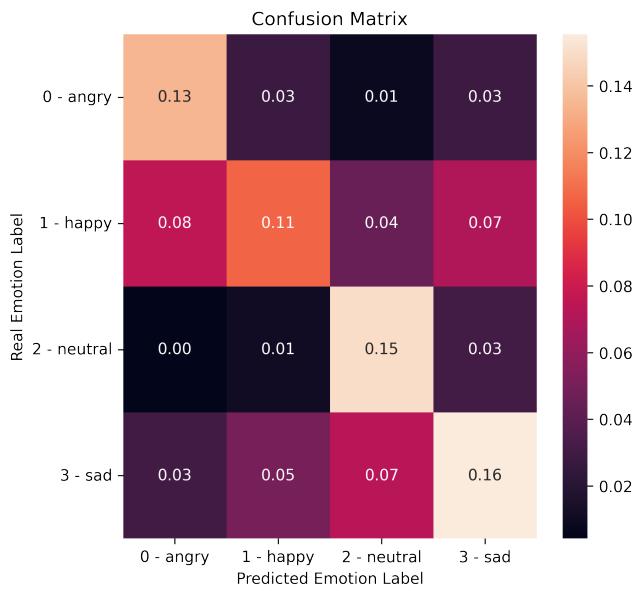


Figure .20: Confusion Matrix of a Balanced Random Forest Classifier.

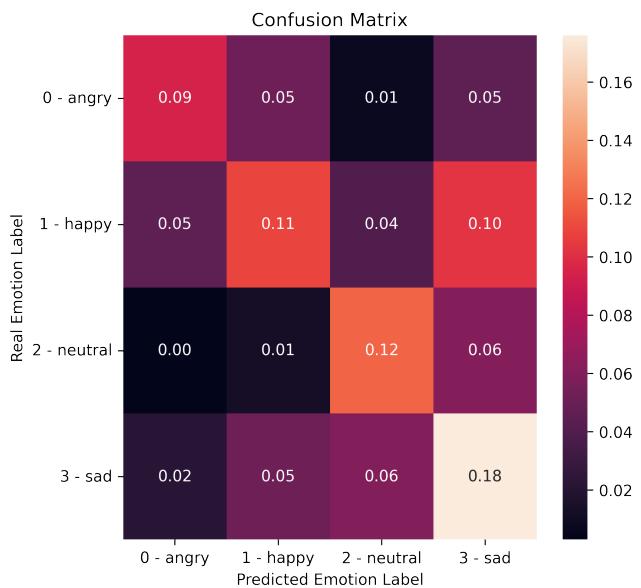


Figure .21: Confusion Matrix of a LSTM Classifier.

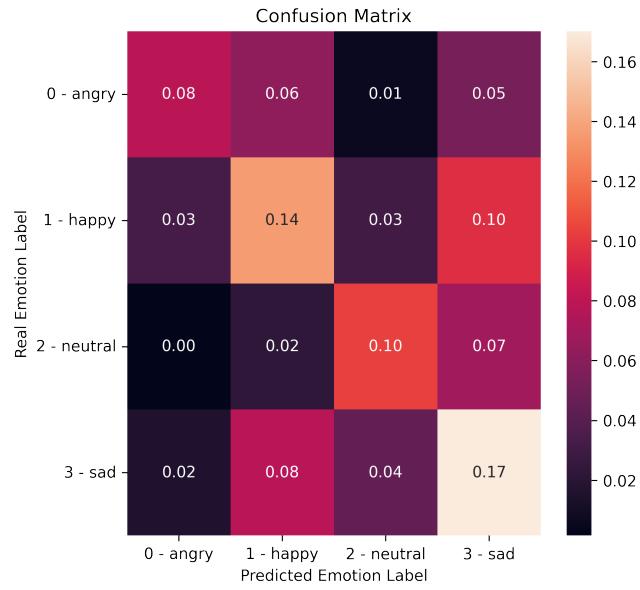


Figure .22: Confusion Matrix of a CNN Classifier.

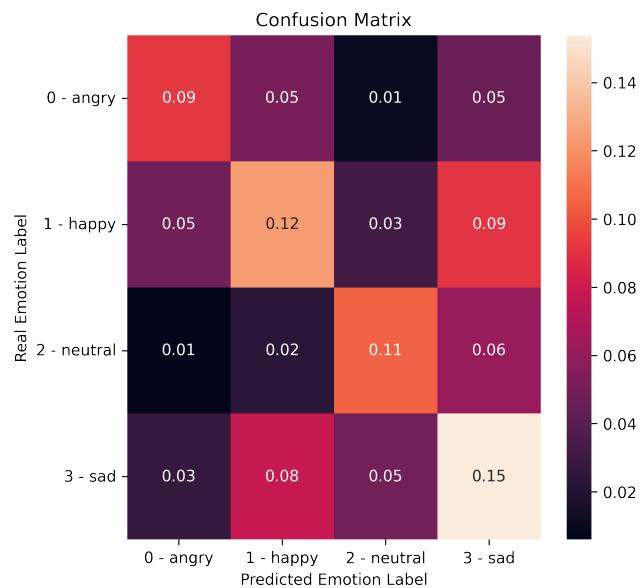


Figure .23: Confusion Matrix of a Linear Discriminant Analysis Classifier.

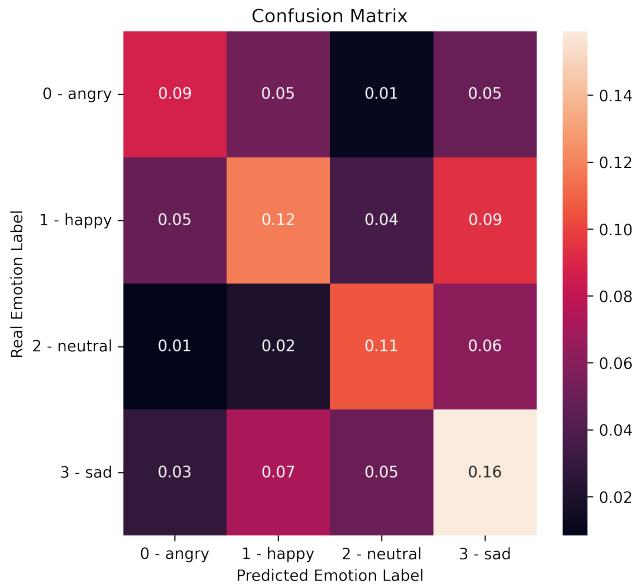


Figure .24: Confusion Matrix of a Ridge Classifier.

.6 SOMETHING ELSE

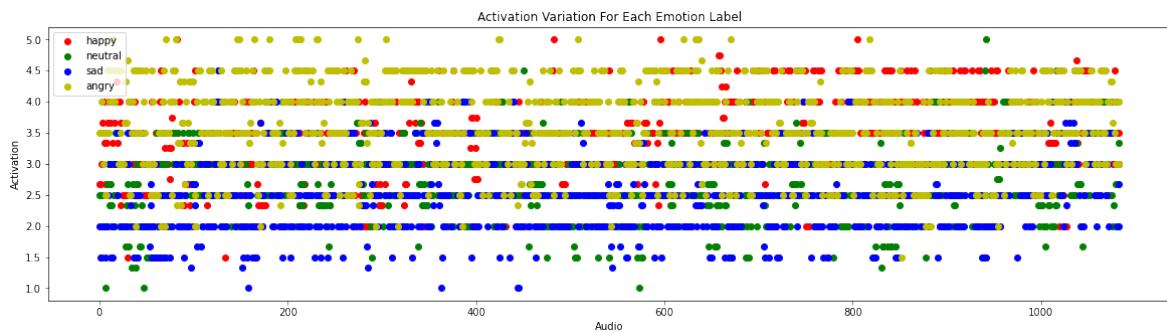


Figure .25: Scatter Plot of the Annotated Emotions in the Activation Dimension

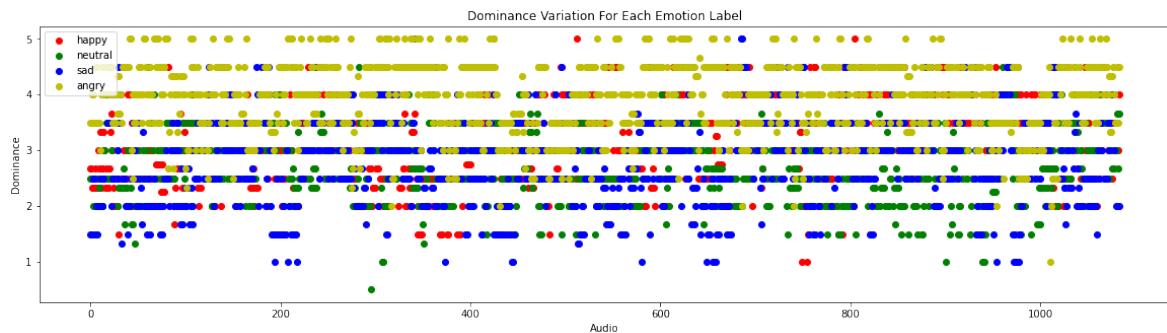


Figure .26: Scatter Plot of the Annotated Emotions in the Dominance Dimension

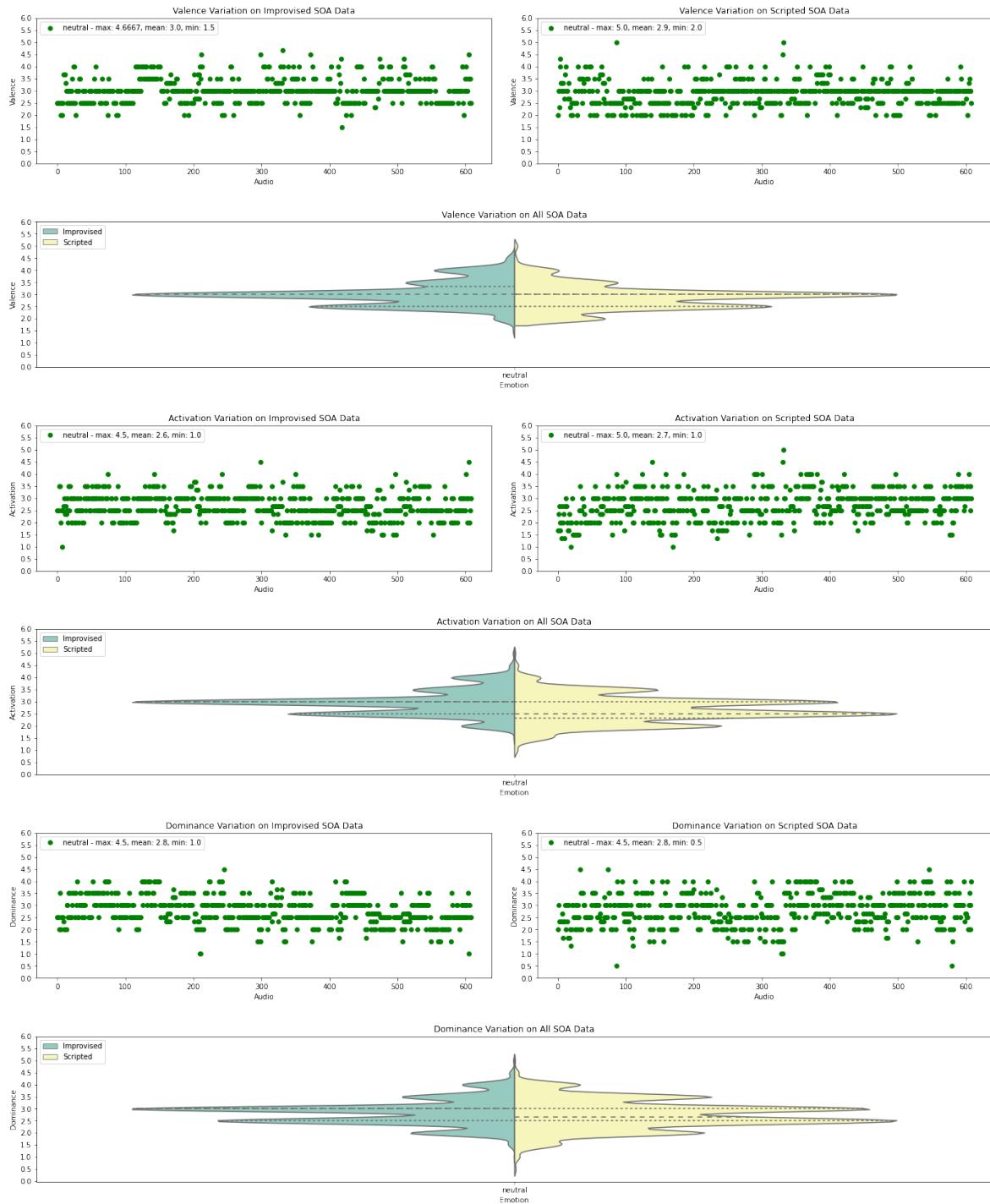


Figure .27: Scatter and violin plots of the emotional content of the IEMOCAP corpus in terms of valence, activation, and dominance relative to the neutral emotion



Figure .28: Scatter and violin plots of the emotional content of the IEMOCAP corpus in terms of valence, activation, and dominance relative to the sad emotion

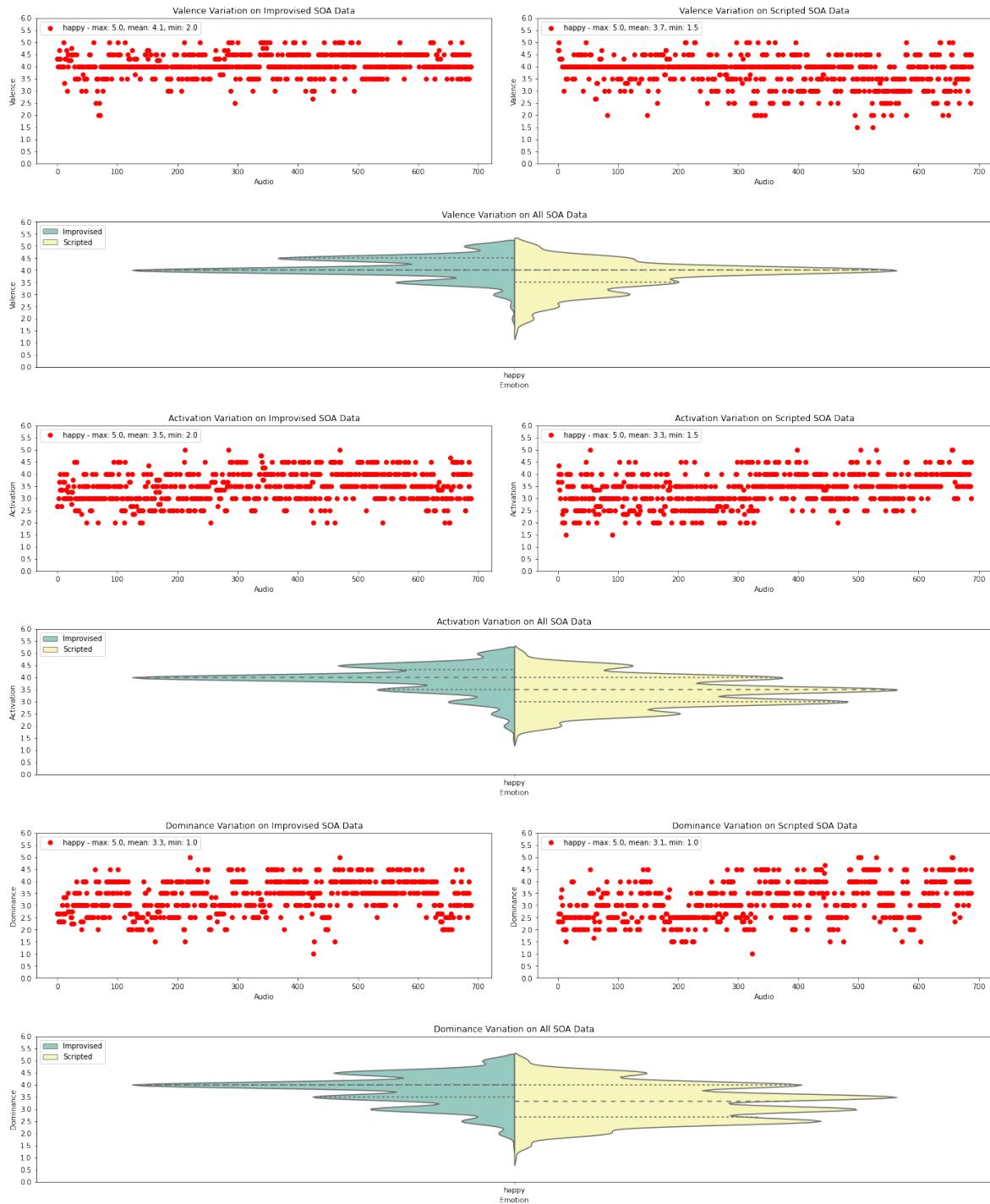


Figure .29: Scatter and violin plots of the emotional content of the IEMOCAP corpus in terms of valence, activation, and dominance relative to the happiness emotion

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 Ballegooij, 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](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 & 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. Rolfs, 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 eINTERFACE&amp;#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. Hernández, 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 & 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-Andrade, 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. Muppudi 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. Aleyani, 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 — 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>.