Mário Francisco
Costa Silva

Desenvolvimento de modelos para avaliação do áudio em computação afetiva

Models development for audio evaluation in affective computing

Mário Francisco
Costa Silva

Desenvolvimento de modelos para avaliação do áudio em computação afetiva

Models development for audio evaluation in affective computing

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

— MIT Sloan professor Erik Brynjolfsson

Universidade de Aveiro
2023

Mário Francisco
Costa Silva

Desenvolvimento de modelos para avaliação do áudio em computação afetiva

Models development for audio evaluation in affective computing

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, Investigadora no Instituto de Engenharia Eletrónica e Informática de Aveiro do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro, e do Professor Doutor Ilídio Castro Oliveira, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática Universidade de Aveiro.

o júri / the jury

presidente / president

Professor Doutor Armando José Formoso de Pinho
Professor Catedrático da Universidade de Aveiro

vogais / examiners committee

Professor Doutor Daniel Filipe Albuquerque
Professor Adjunto do Instituto Politécnico de Viseu - Escola Superior de Tecnologia e Gestão de Viseu

Doutora Susana Manuela Martinho dos Santos Baía Brás
Investigadora Doutorada (nível 1) da Universidade de Aveiro

Palavras Chave

Aprendizagem Automática, Computação Afetiva, Processamento de Voz, Reconhecimento de Emoção da Fala

Resumo

Esta dissertação apresenta um estudo do Reconhecimento de Emoção na Fala (REF) usando abordagens tradicionais de aprendizagem automática e aprendizagem profunda. A principal contribuição deste trabalho é o desenvolvimento e avaliação de dois modelos propostos em múltiplos datasets. Além de explorar o impacto de diferentes conjuntos de características do áudio e metodologias de validação, também investigamos a importância das técnicas de pré-processamento de áudio e seu efeito no desempenho do modelo. Por meio de estudos experimentais, desenvolvemos dois modelos para REF: um modelo eXtreme Gradient Boosting (XGBoost) para a abordagem tradicional utilizando um vetor unidimensional de 33 características do áudio, e um modelo ResNet50 ajustado usando imagens de espectrograma para aprendizagem profunda. Estes modelos alcançaram precisões de 60,69% e 58,24%, respectivamente, para validação cruzada estratificada de 5 vezes no dataset Interactive Emotional Dyadic Motion Capture (IEMOCAP). Além disso, o modelo de aprendizagem profunda superou o modelo tradicional na avaliação cruzada de datasets devido à sua maior capacidade de extração e generalização de recursos, enquanto o modelo tradicional é mais adequado para aplicações em tempo real devido à sua velocidade de processamento mais rápida. Além disso, uma análise detalhada dos dados usando várias estratégias de estratificação levou à identificação de um conjunto de condições para o IEMOCAP que melhorou o desempenho geral dos modelos na avaliação cruzada de datasets. Também foi desenvolvido um pipeline que automatiza o processo de REF para classificação em tempo real ou offline, criando segmentos de áudio com uma determinada duração e classificando as emoções presentes neles usando os modelos desenvolvidos. No geral, esta dissertação fornece uma base para o desenvolvimento de modelos REF mais robustos e precisos, oferecendo uma implementação abrangente e processo de pensamento, juntamente com conclusões e interpretações dos resultados obtidos. As nossas conclusões contribuem para o crescente corpo de pesquisa de REF e fornecem informações valiosas para investigadores e profissionais da área.

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

Abstract

This dissertation presents a comprehensive study of Speech Emotion Recognition (SER) using traditional machine learning and deep learning approaches. The main contribution of this work is the development and evaluation of two models on multiple datasets. In addition to exploring the impact of different feature sets and validation methodologies, we also investigate the importance of audio preprocessing techniques and their effect on model performance. Through experimental studies, we developed two models for SER: an eXtreme Gradient Boosting (XGBoost) model for the traditional approach utilizing a 1-dimensional vector of 33 audio features, and a fine-tuned ResNet50 model using spectrogram images for deep learning. These models achieved accuracies of 60.69% and 58.24%, respectively, for 5-fold stratified cross-validation on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Moreover, the deep learning model outperformed the traditional model in the cross-dataset evaluation due to its higher feature extraction and generalization capabilities, while the traditional model is more suitable for real-time applications due to its faster processing speed. Furthermore, a detailed analysis of the data using several stratification strategies led to the identification of a set of conditions for IEMOCAP that improved the general performance of the models in cross-dataset evaluation. A pipeline was also developed that automates the SER process for both real-time and offline classification by creating voice audio segments with a certain duration and classifying the emotions present in them using the developed models. Overall, this dissertation provides a proof of concept and line of action for future research in developing more robust and accurate SER models, by offering a comprehensive implementation and thought process, along with conclusions and interpretations of the obtained results. These findings contribute to the growing body of research on SER and provide valuable insights for researchers and practitioners in the field.

Contents

List of Figures

List of Tables

Glossary

SER	Speech Emotion Recognition
VAD	Voice Activity Detection
MFCC	Mel-frequency Cepstral Coefficients
VAD	Valence-Arousal-Dominance
VAD	Voice Activity Detection
CNN	Convolutional Neural Network
DNN	Deep Neural Network
IEMOCAP	Interactive Emotional Dyadic Motion Capture
LSTM	Long-Short Term Memory
UAR	Unweighted Average Recall
DL	Deep Learning
SOTA	State-Of-The-Art
FT	Fourier Transform
RMSE	Root-Mean-Squared-Error
MAE	Mean-Absolute-Error
SVM	Support Vector Machine
CFS	Correlation-based feature selection
HMM	Hidden Markov Model
GeMAPS	Geneva Minimalistic Acoustic Parameter Set
RNN	Recurrent Neural Network
MCC	Matthews Correlation Coefficient
AutoML	Automatic Machine Learning
RF	Random Forest
XGBoost	eXtreme Gradient Boosting
CV	Cross-Validation
CM	Confusion Matrix
PNG	Portable Network Graphic

CHAPTER 1

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 system based on the identification of 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 an important part of

teaching process, since the learning and cognition are directly related with the emotional state [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 feature engineering process, evaluate various machine-learning models, and determine the most suitable approach for SER. The goal is to develop a unique set of audio features, study the different approaches to SER, and propose models that 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 requires designing an audio stream processing pipeline for processing the audio.

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 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 handle 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 seven 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.

Chapter 4 presents the emotion recognition strategies and describes their process. This chapter also evaluates different classification models and compares their performance.

Based on the previously settled models, in chapter 5 we will stratify the training data, to analyze potential data biases, such as gender, and to explore the quality of the used data.

The sixth chapter 6 presents a SER pipeline that can be applied in a video conference system, detailing its implementation and results.

The final chapter 7, discusses the accomplishment of the objectives, with remarks regarding the development process, obtained results, conclusions, and future work.

CHAPTER 2

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.

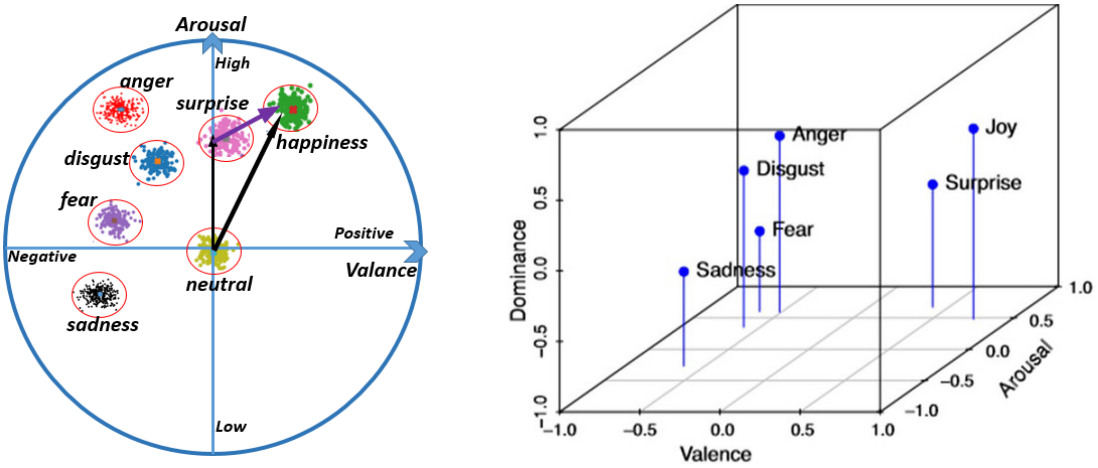
2.1.2 Dimensional Emotion

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 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
eNTERFACE'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.	Positive, negative, or neutral sentiment.	Natural
CREMA-D [27]	Audio Video Text	English	7442 clips of 12 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.

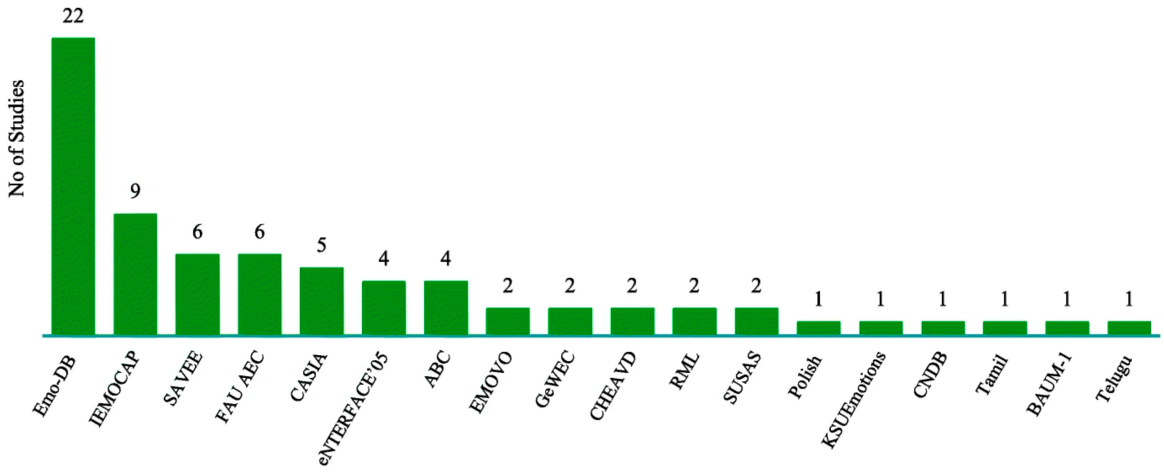


Figure 2.2: Frequency distribution of databases on reviewed articles from a SOTA study [35].

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 into 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].

Low-Level-Descriptors				**Functionals**		
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, ...)		

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].

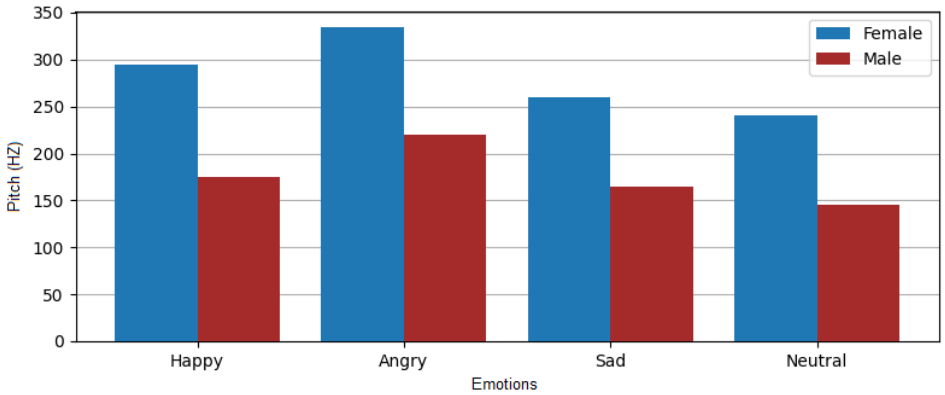


Figure 2.3: Average pitch comparison of male and female speakers' [38].

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 Mel-frequency Cepstral Coefficients (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. The 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, Convolutional Neural Network (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.

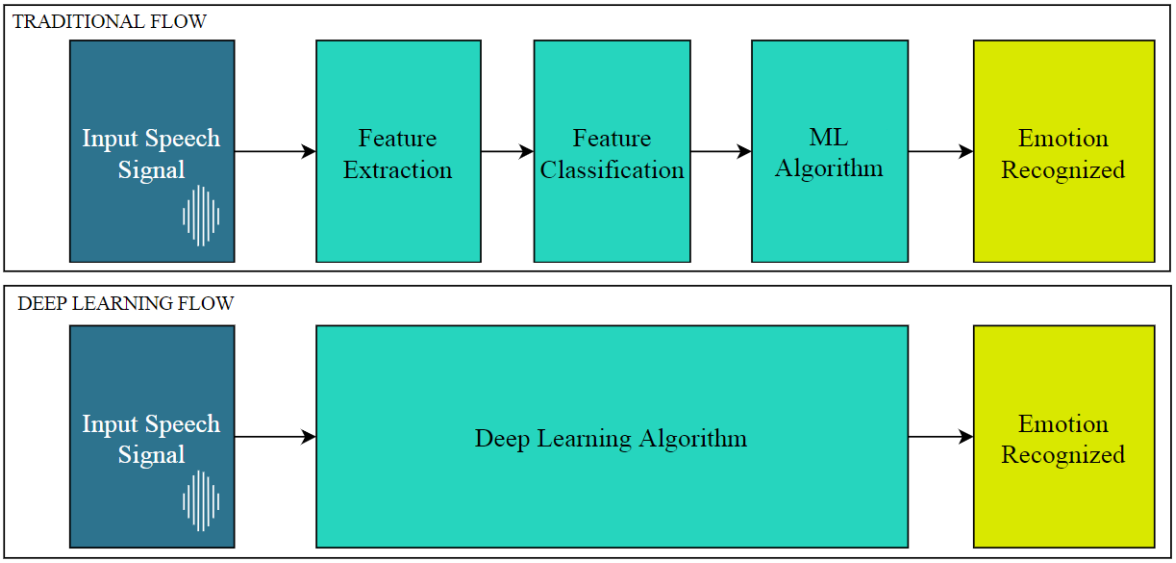


Figure 2.4: Traditional machine learning flow vs. automatic feature extraction with DL flow [42].

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, Hernáez, *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].

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.

Some past works used traditional machine learning classification techniques such as Linear Discriminant Analysis, Regular Discriminant Analysis, SVM, K-Nearest Neighbor, RF, and Gaussian Mixture Model [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 RF, eXtreme Gradient Boosting (XGBoost), and a multilayer perceptron.

A new architecture was introduced in 2020 by Issa, Demirci, and Yazici, for identifying emotions in sound files using a one-dimensional CNN, which is not typically considered a traditional model, however, the extraction of features was still performed manually. They obtain a one-dimensional array by taking mean values along the time axis of 193 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 [54].

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, [55] and [56], 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), Long-Short Term Memory (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. It is also 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 numeric spectrogram values.

In a paper published in 2021, García-Ordás, Alaiz-Moretón, Benítez-Andrades, *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±5.44% and 82.82±4.99% [61].

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): Accuracy (%)
Traditional Feature-Based SER Approaches			
Spoken emotion recognition using hierarchical classifiers (2011) [51]	MFCC Log-spectrum Pitch	HMM	EMO-DB (7): 71.75
Emotion recognition using a hierarchical binary decision tree approach (2011) [52]	Zero crossing rate Root-mean-square energy Pitch MFCC Harmonics-to-noise ratio	Multi-level binary decision trees	(UARs) AIBO (5): 41.57 IEMOCAP (4): 58.46
Multimodal SER and Ambiguity Resolution (2019) [53]	Pitch Harmonics Pause	Random forest, extreme gradient boosting and multilayer perceptron	IEMOCAP (4): 56.6
SER with deep CNNs (2020) [54]	MFCC Chromagram Mel spectrogram Spectral contrast	One-dimensional CNN	RAVDESS (8): 71.61 IEMOCAP (4): 64.30 EMO-DB (7): 86.10
Deep Learning-Based SER Approaches			
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
SER Using Quaternion CNN (2021) [62]	Mel spectrogram encoded in an RGB quaternion domain	Quaternion CNN	RAVDESS (8): 77.87 IEMOCAP (4): 70.46 EMO-DB (7): 88.78
Rethinking CNN Models for Audio Classification (2020) [63]	Mel spectrogram	Ensemble of ImageNet pre-trained DenseNets	ESC-50 (50): 92.89 UrbanSound8K (10): 87.42
SER Using Deep CNN and Discriminant Temporal Pyramid Matching (2018) [64]	Mel spectrogram	Deep CNN (Fine-tuned the AlexNet pre-trained on ImageNet)	EMO-DB (7): 87.31 RML (6): 69.70 eNTERFACE05 (6): 76.56 BAUM-1s (6): 44.61

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.

While we will not be implementing multimodality in our work, it is an important concept to consider for future developments in emotion recognition systems. The use of multiple modalities could potentially enhance the accuracy and reliability of emotion recognition models and contribute to the development of emotionally intelligent systems.

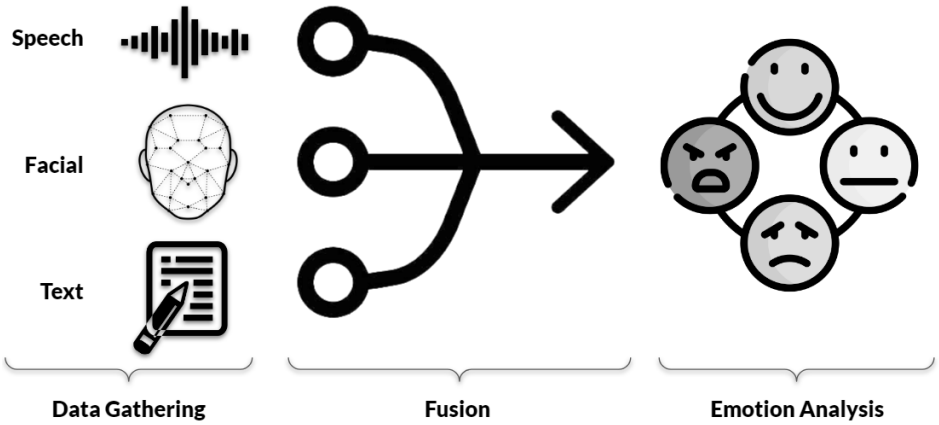


Figure 2.5: Multimodal emotion recognition illustration.

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 eNTERFACE'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] Bitext [73]	Text	No	No
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.

CHAPTER 3

Methodology

3.1 Introduction

In this chapter, we present an overview of our research methods, highlighting key steps and techniques utilized, with the main objective of developing an accurate and robust SER system.

3.2 Dataset Collection

The data collection process involved the selection and acquisition of audio-emotional datasets. The chosen datasets were selected based on their alignment with SOTA research. Preference was given to large and diverse datasets that contained gender-balanced samples and were widely used in the field.

3.3 Data Preprocessing

To enhance the quality of the audio data, a series of preprocessing steps were employed. These included noise reduction techniques to minimize background noise and silence trimming to remove unnecessary portions of the audio signal.

3.4 Feature Extraction and Analysis

The feature extraction process involved extracting relevant audio features from the preprocessed data, 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. However, our study focused on global features to simplify the feature representation and streamline the analysis process. Global features provide a holistic representation of the entire audio signal, summarizing its overall characteristics without focusing on specific local segments.

The features were then selected based on their relevance to the SER task. Various analysis techniques were applied to the extracted features to gain insights into the underlying patterns and characteristics of the audio data, such as time-domain analysis, frequency-domain analysis, and signal processing algorithms.

3.5 Models Implementation

Two main approaches were employed in model development: a traditional feature-based approach and a deep learning-based approach. The traditional approach involved audio feature engineering, to find a set of suitable audio features, and the selection of appropriate classifiers. The deep learning approach focused on studying the effectiveness of different feature representations and selecting suitable classifiers, which also involved transfer-learning techniques. The models were also then evaluated through cross-dataset validation and compared the results obtained to other SOTA approaches.

3.6 Data Stratification

Data stratification refers to a process of dividing a dataset into subgroups, based on certain properties. We explored different ways of grouping the data and interpreted their effects on the models, this involved stratifying the audio recordings based on various factors, such as duration, speaker genders, and discrete and dimensional annotations. The purpose of data stratification was to address potential biases and improve the performance of the models across different conditions.

3.7 SER Pipeline

A versatile and scalable SER pipeline was developed, which involved consuming audio data, detecting voice using a VAD tool, and subsequently performing emotion classification on the detected speech segments. The pipeline enabled efficient processing of audio data, in both online and offline time, for performing SER tasks.

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. Figure 3.1 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 [102].

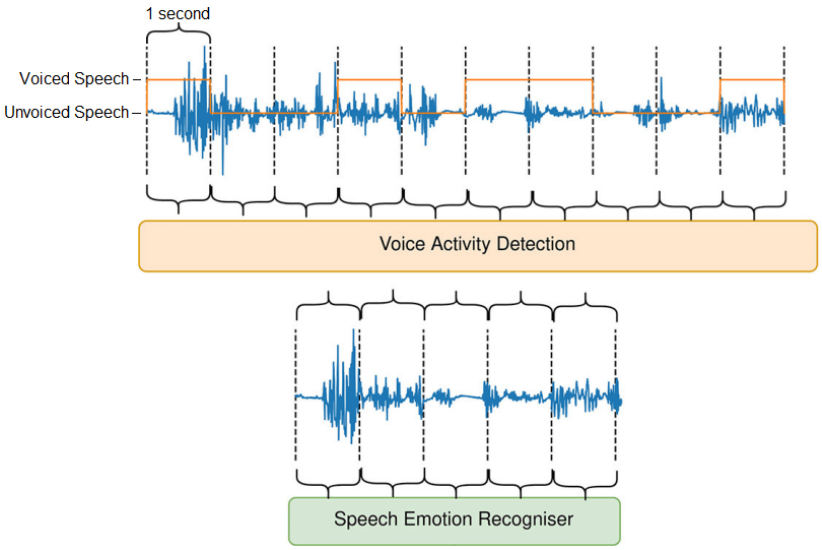


Figure 3.1: Voice Activity Detection employment on an audio signal [102].

3.8 Ethical Procedures and Concerns

Ethical considerations were given due attention throughout the research process. The study strictly adhered to ethical guidelines, prioritizing transparency, informed consent, and the protection of personal data.

One ethical concern addressed was the potential bias in the proposed SER system's accuracy for individuals with certain characteristics, such as age, gender, accents, or speech disabilities. We conducted an analysis to identify and quantify any biases present in our models, promoting fairness and transparency.

Another ethical concern was the potential misuse of such systems for monitoring or controlling individuals without their consent. Our models are only applied to individuals who have willingly provided their data for emotional evaluation purposes.

To address privacy concerns, our data processing pipelines strictly adhere to principles of data protection. We utilize only the streaming data of participants as input for the SER models, and we do not store or use sensitive data for any purposes other than emotional analysis. Additionally, we have chosen not to use speech transcriptions to further safeguard privacy, as they may reveal confidential information.

3.9 Conclusion

The methodology section provided an overview of the research design, data collection, and preprocessing steps, model development approaches, data stratification techniques, the development of the SER pipeline, and ethical procedures and concerns. This methodology explains the basis of the subsequent chapters, where the specific details and findings are discussed in more depth.

CHAPTER 4

SER Development

4.1 Datasets

In this section, we present a detailed account of the datasets employed in the development and evaluation of our SER system.

The first dataset was utilized to investigate and select the most optimal features and evaluate the performance of classification models and effective classification strategies.

4.1.1 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 [25], demonstrates a similar amount of annotated labels on scripted and spontaneous sessions on this dataset.

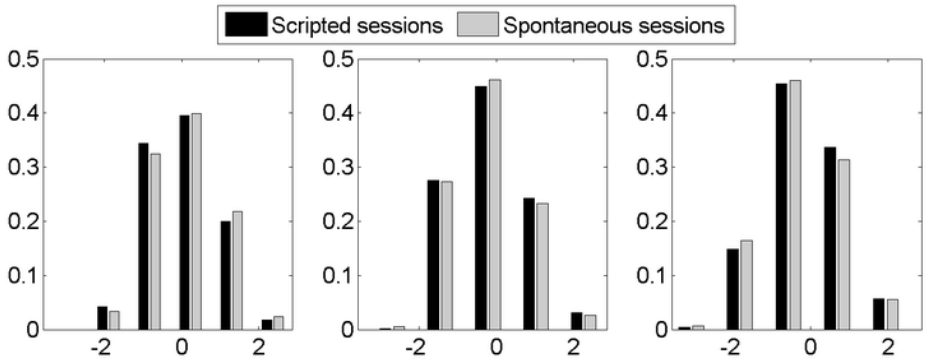


Figure 4.1: Distribution of the emotional content of the IEMOCAP corpus in terms of (a) valence, (b) activation, and (c) dominance. The results are separately displayed for scripted (black) and spontaneous (gray) sessions.

Most researchers when using this dataset perform 4 class emotion recognition, and also, consider the emotion excitement as happiness, due to their similarities and to even out the distribution of files per emotion. We decided to use the same data, as shown in Table 4.1, ending up with a total of 5531 audio files recorded with a sample rate of 16000 Hertz.

Overall, IEMOCAP 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 operations. Several 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 compare our developed models effectively, which is why we utilized it as a training and testing dataset for our SER models and to explore classification strategies and their biases.

4.1.2 Testing Datasets

To evaluate the generalization ability of our final models trained on the IEMOCAP dataset, we tested them on three additional emotion datasets: eNTERFACE'05, CREMA-D, and EMO-DB. The inclusion of these datasets allows us to assess the efficacy and applicability of our proposed SER system across diverse contexts and conditions. For all three testing datasets, we only used the set of 4 emotions present in the IEMOCAP.

The eNTERFACE'05 emotion database [24] was designed and collected during the eNTERFACE'05 workshop. This dataset contains discrete annotated emotions, where each subject was asked to listen to six successive short stories, each eliciting a particular emotion. 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. The diversity of accents present in the dataset and its authenticity (due to the data being elicited), make it a suitable choice for testing the models' performance with different factors than the IEMOCAP. The number of used files per emotion is presented in Table 4.1.

The EMO-DB database is an acted German emotional database created by the Institute of Communication Science, Technical University, Berlin, Germany. Ten professional speakers (gender-balanced) participated in the data recording. This dataset is composed of a total of 339 audio files annotated with the four emotions used in the IEMOCAP, shown in table 4.1, making it a small-sized dataset. However, it allows us to test more directly the models' limitations, mostly the language bias since it provides files in a different spoken language than the training dataset.

The CREMA-D dataset includes clips from 91 actors spoken in English. These clips are from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African American, Asian, Caucasian, Hispanic, and Unspecified). It provides a variety of data which is ideal to test a model's generalization ability across new datasets. The sentences were presented using one of six different emotions, but as mentioned before, we selected only the emotions present in the IEMOCAP, resulting in 4898 audio files 4.1.

Table 4.1: Number of audio files per emotion from the described datasets.

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

4.2 Audio Preprocessing

Audio preprocessing is an essential step that is performed on audio data before it is fed into machine learning models. The objective of audio preprocessing is to clean the collected audio data, extract relevant features, and transform the data into a format that can be easily interpreted by the models. This section discusses important aspects of audio preprocessing, including noise reduction and audio trimming, and how we employed them.

4.2.1 Noise Reduction

One of the key techniques used in audio preprocessing is noise reduction. In audio data, noise can arise from a variety of sources, such as background noise, microphone interference, or electrical interference. The presence of noise in the audio data can significantly impact the performance of any audio classification system.

One common strategy for denoising audio is spectral gating, which involves gating the signal only on high-level sounds. We recurred to the *noisereduce* library to implement this noise reduction technique. This algorithm estimates a noise threshold for each frequency band of the signal. 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. Parameter "y" is the audio signal, and "sr" specifies its sampling rate in Hz. "n_fft" sets the number of points for each Fourier transform, while "hop_length" sets the number of samples between each frame. "prop_decrease" adjusts noise reduction strength, and "time_constant_s" sets the length of the filter. The function returns a new audio signal with noise removed, for use in downstream processing steps like feature extraction and classification.

```
noisereduce.reduce_noise(y=y, sr=sr, 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

Another important aspect of audio preprocessing is audio trimming. This technique involves removing unwanted portions of the audio signal that are not relevant. Trimming the silence from the beginning and end of an audio clip is a common technique in speech-based classifications. This reduces the computational overhead and also makes the data more manageable.

We defined 30 decibels as the threshold for considering silence and removed any segments, at the beginning and end of the audio, that fell below this threshold. This was done because 30 decibels is generally considered a very low level of noise, and is often used as a standard threshold for measuring background noise levels in quiet environments.

Table 4.2 shows the original and after trim durations for all of the IEMOCAP audio files, and it is clear that this process of removing unwanted noise at the beginning and end of the audio files resulted in overall shorter audio durations.

Table 4.2: Audio files' original and after trim durations.

Metric	Original Duration	After Trim Duration
Minimum	0.585	0.256
25th Percentile	2.324	1.952
Mean	4.549	4.065
75th Percentile	5.773	5.248
Maximum	34.139	33.184

4.2.3 Conclusion

Figure 4.2 compares an original and preprocessed audio file from the IEMOCAP dataset. We observed the amplitudes of the preprocessed waveform are reduced and that there are more silent parts compared to the original. The preprocessed spectrogram shows a cleaner representation and some reduction of decibels values. The preprocessed power spectral density also contains lower values of dB/Hz, and, for example, in frequencies around 1100 and 1900 Hz, the elimination of some spike values can be perceived. In addition to these observations, we also listen to the resulting audio and manually detected noise reduction and a clearer sound.

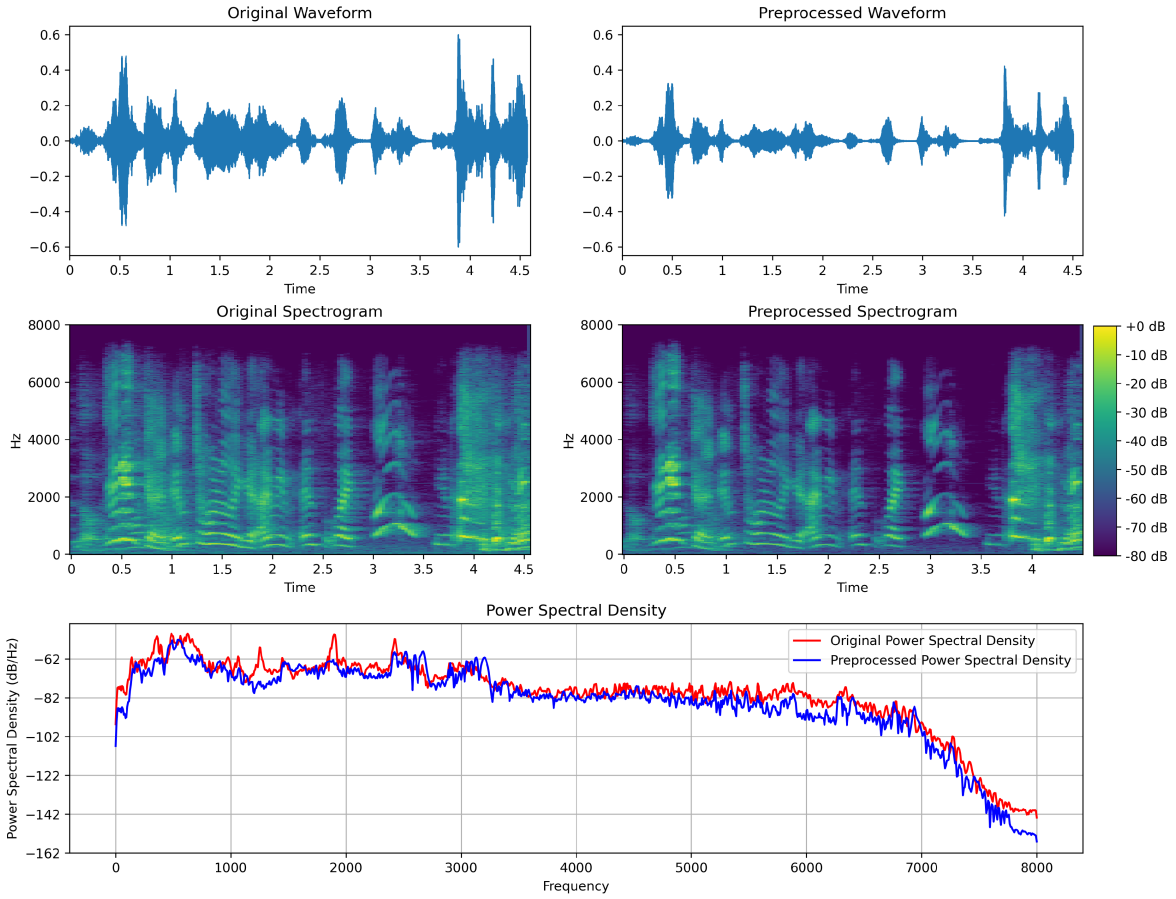


Figure 4.2: Visual representations of audio features before and after preprocessing.

These results demonstrate that the employed techniques cleaned the audio data, resulting in shorter durations and less noise. Overall, this audio preprocessing is an essential step to ensure the quality of any type of audio classification, but it requires an in-depth comprehension of audio signals to create a successful preprocessing pipeline since it can cause a great drawback in some scenarios where noise may be part of the signal of interest, and removing it may lead to misinterpretation.

4.3.1 Feature Extraction

The first step in a traditional approach is feature extraction to transform raw audio data into a set of informative features that can capture key characteristics of the signal.

In this regard, the widely-used Librosa toolkit was employed to extract various audio features, from the audio files of the IEMOCAP dataset, and subsequently, they were processed using statistical metrics. The extracted features and associated metrics are summarized in Table 4.3, with a total of 327 extracted features.

Table 4.3: Extracted audio features and the statistical functions applied.

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

[1] Custom function detailed on the feature selection subsection 4.3.2.
[2] Only for the MFCCs.

4.3.2 Feature Analysis

An important task following feature extraction is analyzing and interpreting the extracted data to gain a deeper understanding of the audio signals and the features that describe them.

Audio Features Visualization

In this process, we visually analyzed and interpreted the features' data by graphically representing each feature from an audio segment. The figures in Section .1.1 of the appendix demonstrate some of the graphics we used to visualize the features.

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. To account for different-length audio signals, it was also divided the number of spikes to the total length of the data, as the Code Snippet 2 demonstrates. Consequently, this metric was also tested and applied to every other audio feature.

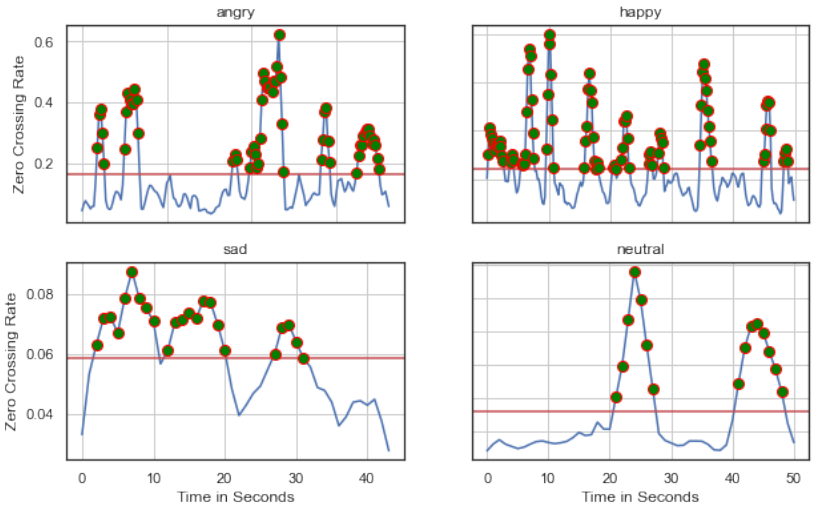


Figure 4.3: Zero crossing rate wave plot annotated with spikes.

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

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

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. Other bar plots are presented in the appendix .1.2.

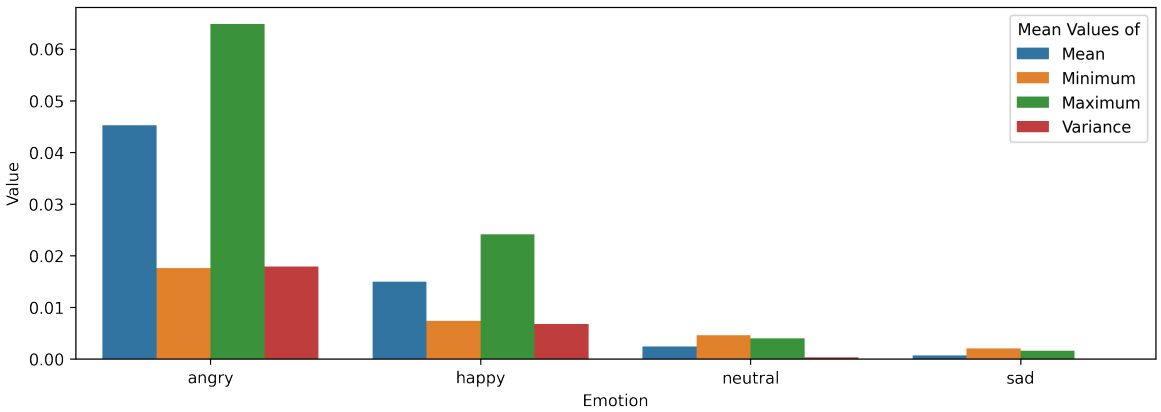


Figure 4.4: Bar plots mean for metrics used on the mel spectrogram feature.

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, for a given emotion, could indicate that the feature has consistent values and is relevant for identifying that emotion.

Figure 4.5 is an excerpt of the figure .9 in the appendix .1.3, and it demonstrates an example of this analysis for the zero crossing rate with 5 different subjects on the same sentence for the anger emotion. From this graphic, it was observed that there is an overlap between the surrounding areas for each emotion, which allows us to conclude that the feature has utility for describing each emotion. However, due to the different lengths of each audio segment, it is ambitious to guarantee this conclusion.

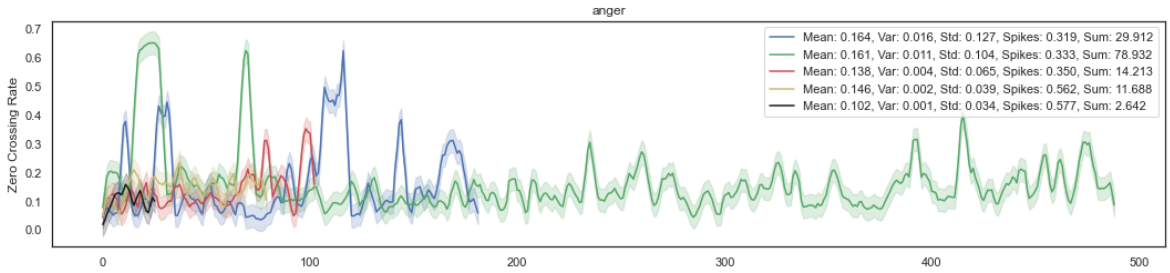


Figure 4.5: Zero crossing rate wave plot with a surrounding area of five male subjects for the same utterance with the anger emotion.

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 case. If the areas around the feature, on different emotions, do not heavily overlap, it may be an indicator that the feature can be able to discriminate 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 challenging to draw a direct and well-founded conclusion.

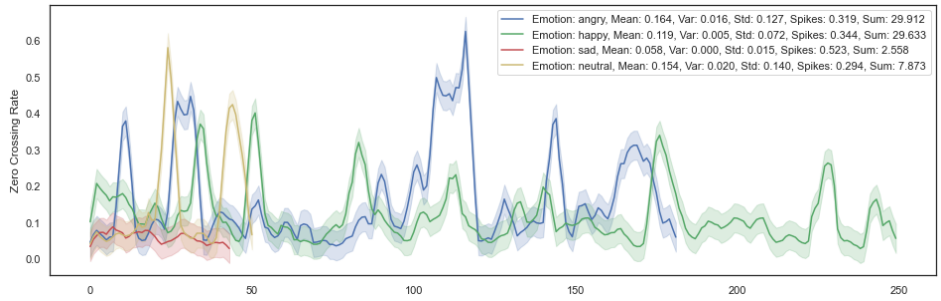


Figure 4.6: Zero crossing rate wave plots with a surrounding area of a single male subject and sentence for all different emotions.

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 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. Other examples of this type of graphic are also in the appendix .1.4.

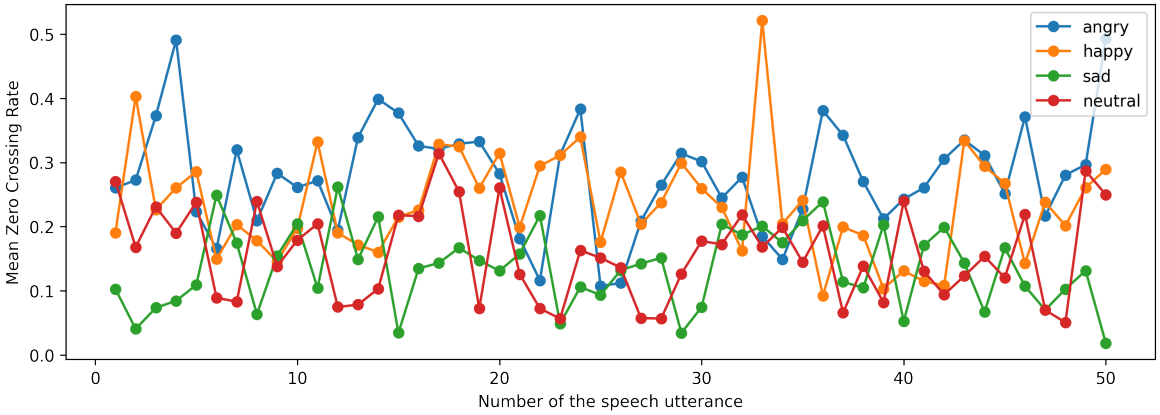


Figure 4.7: Zero crossing rate mean values variation plot along 50 audios of speech utterances for all emotions.

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.

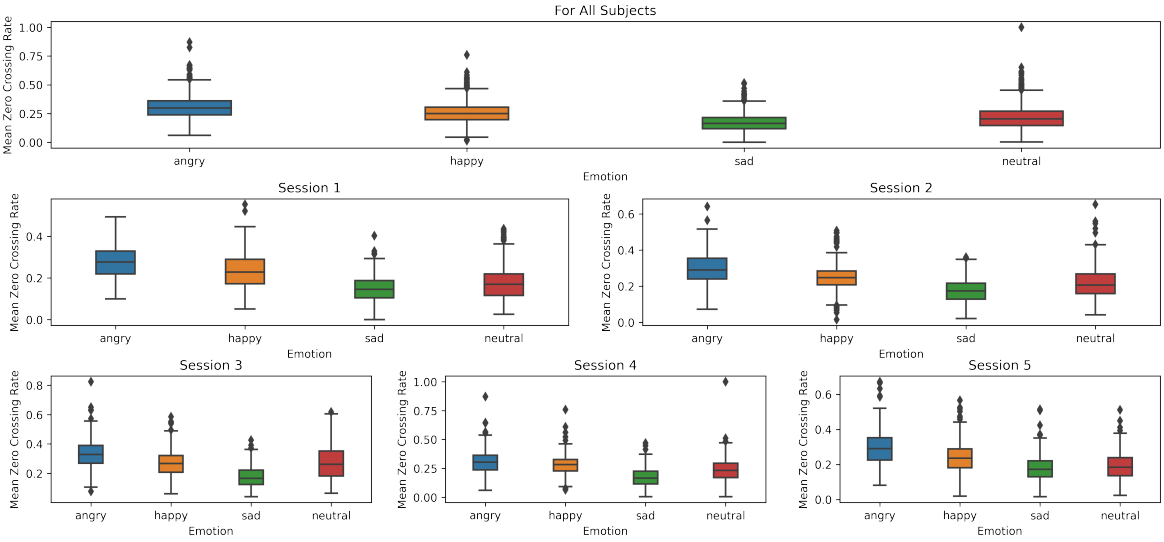


Figure 4.8: Zero crossing rate mean values box plot for all emotions and different subjects.

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

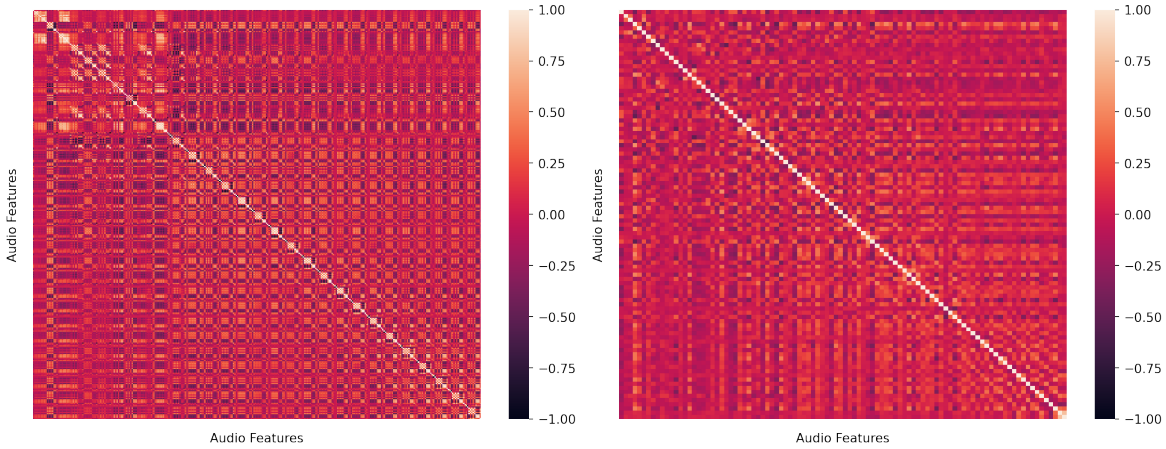
the given task. The process of feature selection is aimed to improve the accuracy of the model and reducing the problem's complexity by removing redundant or irrelevant features.

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.

Correlation-based feature selection

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 4.9a.

A CFS was performed by selecting every pair of features with a Pearson correlation coefficient absolute value of 0.6 or above, then it was removed the feature with the highest average correlation value with all the other features. This process resulted in the elimination of 229 features, leaving 98 features for subsequent analysis. The correlation matrix after the feature selection process is presented in figure 4.9b.



(a) Correlation matrix of all the features. **(b)** Correlation matrix after CFS.

Figure 4.9: Audio features' correlation matrices before and after CFS.

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 CV and compared the mean and standard deviation accuracies of all folds, as well as the total execution time for various classifiers using default parameters from the scikit-learn library [92]. The input given to the models is the set of features obtained after CFS, as shown in Table 4.4.

Table 4.4: Performance of various classifiers in 5-fold CV using the features obtained after CFS.

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

Based on the evaluation results, the RF classifier was chosen for further analysis. This model exhibited the second-best average accuracy across the 5 folds, however, it was much faster than the XGBoost to train. Therefore, RF was the selected model 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 CV with the previously selected RF classifier, using all features except one, and then removing one feature based on the lowest mean accuracy of the 5 folds. This iterative process continues until only one feature remains.

A method was then developed to select the furthest and highest accuracy. This method resembles the standard maximum, but it multiplies the maximum value by a threshold value of 0.99 so that it can find close to the maximum values that have more features removed, creating a balance between accuracy and the number of features. Figure 4.10 displays the mean accuracies obtained at each step and the chosen furthest highest accuracy.

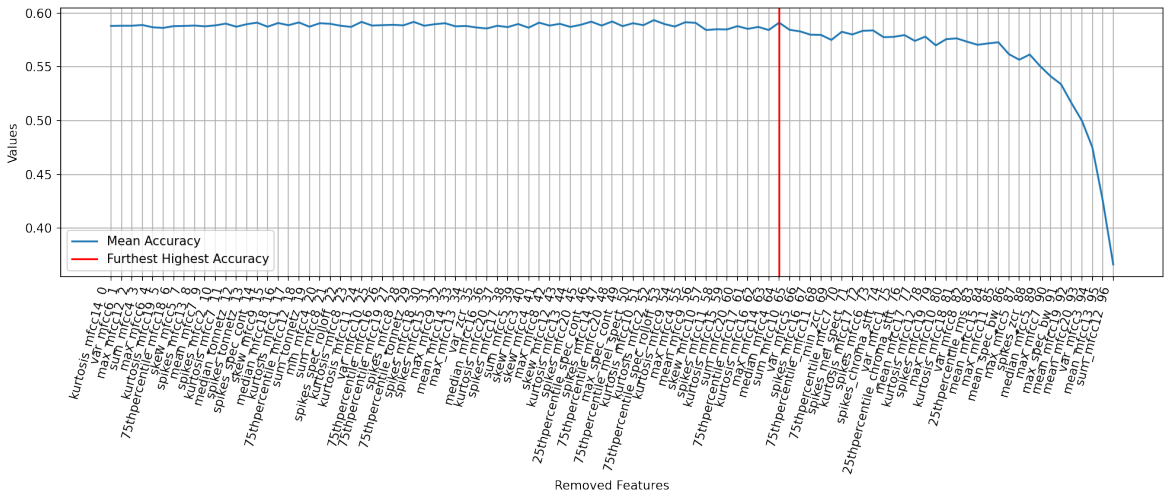


Figure 4.10: Sequential feature selection with backward propagation using the mean accuracy as the selection criteria.

This process led to the elimination of 65 features from the initial set of 98 obtained after CFS, leaving a total of 33 features, as shown in Table 4.5.

Table 4.5: Final set of the 33 selected features.

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

Feature Selection Evaluation

To assess the feature selection quality on the development dataset, we trained and evaluated the predictions of a RF Classifier model with different sets of features, using accuracy as the evaluation metrics. The results are presented in Table 4.6. Additionally, we plotted the confusion matrices of the predictions, which can be found in the appendix .1.5.

Table 4.6: RF 5-fold CV evaluation metrics using different sets of features.

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

The results of the feature selection techniques have shown that while there may be a small loss of accuracy between the initial set of extracted features and the set obtained after CFS, the model can achieve comparable accuracy while using only around 70% of the original set. Moreover, applying backward selection to the remaining 98 features yields optimal results by maintaining the accuracy and reducing the original feature set by approximately 90%.

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

4.3.4 Classifiers Evaluation and Selection

Evaluation Strategy

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.

For the reasons above and because it was the most recurred method in the SOTA research, it was decided to utilize 5-fold stratified CV with an 80-20 train-test split, which provides more fairness to model comparisons. In terms of metrics, we decided to calculate the testing folds averages accuracy, macro-f1 score, precision, recall, and MCC.

Accuracy is a basic metric to evaluate the model's performance that measures the proportion of correctly classified samples. Precision measures the proportion of true positives among all samples predicted as positive by the model. Recall indicates how many of the actual positive samples are correctly identified by the model. The macro-f1 score takes into account both precision and recall to

provide a more balanced measure of performance for imbalanced datasets. It calculates the harmonic mean of precision and recall. MCC is a correlation coefficient between the true labels and the predicted labels. It measures how well the model's predictions match the true labels, and ranges from -1 (perfect disagreement) to +1 (perfect agreement), 0 being no agreement at all, which means the predictions were random. A Confusion Matrix (CM) 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.

Machine Learning Models

To develop the most effective speech emotion recognition model, a variety of automatic machine learning techniques were employed in the study. Two different Automatic Machine Learning (AutoML) techniques were employed to search for the best models. One such technique was the Auto-SKLearn ensemble [103]. This technique searches for the best models and ensembles by exploring a vast space of possible algorithms and hyperparameters using a meta-learning approach. After training and obtaining the ensemble model, the most influential classifiers of the ensemble were identified and taken into account for further exploration individually.

Another AutoML technique, AutoKeras [104], was applied to create a CNN model, which we also used to create another model by combining it with an LSTM. So, in addition to the previous classifiers, two deep-learning models were also evaluated.

Based on this, we explored multiple classifiers for SER and tested several hyperparameters for each model to find the most optimal results, including:

1. RF: A popular decision tree-based ensemble method that trains multiple decision trees on different subsets of the dataset and outputs the mode class predicted by the trees.
2. Balanced RF: A variant of the RF that addresses class imbalance by undersampling the subsets provided to train each tree.
3. XGBoost: A gradient boosting model that uses decision trees as base learners. Each tree is trained using the negative gradient of the loss function concerning the prediction of the previous iteration. This means that each subsequent tree corrects the errors of the previous ones, leading to a more accurate final prediction.
4. AdaBoost: Short for Adaptive Boosting, it is a boosting algorithm that combines multiple weak classifiers into a strong classifier by iteratively adjusting the weights of incorrectly classified instances. At each iteration, a new weak classifier is trained, and its weight is added to the ensemble based on its accuracy. The final prediction is then made by summing the weighted predictions of all the weak classifiers.
5. Histogram Gradient Boosting: An ensemble model that combines gradient boosting with a histogram-based approximation of decision trees, which improves performance on large datasets. It splits data into histograms based on their values instead of splitting features to create decision trees.
6. SVM: A linear or nonlinear model that finds the best boundary separating data into different classes by maximizing the margin between the classes.
7. Ridge:A linear model with L2 regularization that minimizes the sum of squared residuals between the predicted and actual values.
8. Linear Discriminant Analysis: A statistical approach that finds a linear combination of features to maximize the separation between classes by modeling the distribution of the data in each class.

9. CNN: A neural network commonly used for image classification tasks. It applies filters to the given input through convolutional layers to extract different features, passes these features through pooling layers to reduce dimensionality, and produces the final output label using fully connected layers.
10. CNN + LSTM: A hybrid deep learning architecture that combines the strengths of convolutional and recurrent neural networks, allowing for both local and temporal feature extraction from the data.

Results and Conclusions

Having chosen a set of classifiers, we then applied our evaluation strategy on the IEMOCAP dataset. The results obtained from the tested models were compiled and exhibited in Table 4.7, and the confusion matrices of each model are in the appendix .1.6. Upon analyzing these results, XGBoost is the best candidate, reaching an average accuracy of 60.69% while utilizing only 33 audio features, making it a relatively simple model. It also obtained the highest values for macro F1 score, precision, and MCC. In terms of prediction time, it is only slower than the linear models, being the third fastest at making predictions, which is an essential factor for real-time scenarios.

Table 4.7: Tested models' 5-fold stratified CV performance on IEMOCAP.

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

Chosen Model: eXtreme Gradient Boosting. The chosen model for our study is XGBoost, a highly accurate machine learning model renowned for its utilization of gradient-boosted decision tree ensembles. As an improvement over the Gradient Boosting Machine algorithm, XGBoost incorporates regularization techniques, specifically a combination of L1 and L2 regularization, to mitigate the risk of overfitting.

In the XGBoost model, decision trees are sequentially built using a boosting technique. Each subsequent tree aims to rectify the errors made by the previous tree, with a particular emphasis on addressing misclassified data points. This iterative process assigns higher weights to misclassified points to prioritize their correct classification in subsequent iterations. The model optimizes its performance through gradient descent, which minimizes a loss function measuring the discrepancy between predicted and actual values.

By amalgamating multiple decision trees in an ensemble, leveraging boosting, and employing regularization techniques, XGBoost effectively harnesses the collective knowledge of individual trees, resulting in highly accurate predictions.

eXtreme Gradient Boosting Implementation. The Python code for the model was implemented using the XGBoost library [105], and is presented on the Code Snippet 3. To obtain the parameters for the model, we used a Grid Search hyperparameter optimizer function, which performs an exhaustive search over every combination of a specified set of parameters.

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

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

XGBoost has a number of parameters that can be tuned to improve the performance of the algorithm. The most important parameters are:

1. *max_depth*: the maximum depth per decision tree. A deeper tree might increase the performance, but also the complexity and chances to overfit.
2. *learning_rate*: determines the step size at each iteration that the model optimizes toward its objective.
3. *n_estimators*: the number of trees to be boosted in the ensemble.
4. *subsample*: the ratio of the training instances to be sampled for each tree.
5. *colsample_bytree* and *colsample_bylevel*: a family of parameters for specifying the subsampling method of columns of the trees.

eXtreme Gradient Boosting Advantages. Overall, the XGBoost is a versatile and powerful algorithm that can be applied to a wide range of machine-learning problems and provides several advantages:

- Speed: XGBoost is faster than many other popular machine learning algorithms, especially when compared to traditional gradient boosting implementations.
- Performance: it has a strong track record of producing high-quality results in various machine-learning tasks.
- Scalability: XGBoost supports parallel processing, which makes it possible to train models quickly on large datasets.
- Handling Missing Values: It has an in-built routine to handle missing values. XGBoost tries different things as it encounters a missing value on each node and learns which path to take for missing values in the future.
- Regularization: The algorithm includes L1 and L2 regularization that helps to avoid overfitting and improve the generalization ability of the model.
- Probabilistic Predictions: Since this is a decision tree-based model, it can output the probabilities for each class, calculated as the fraction of trees that vote for each class. This provides more information than just the final predicted class, which allows tuning a threshold value for classification.
- Interpretability: XGBoost provides feature importances, allowing for a better understanding of which variables are most important in making predictions.
- Customizability: XGBoost has a wide range of hyperparameters that can be adjusted to optimize performance, making it highly customizable.

4.4 Deep Learning-Based SER

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

4.4.1 Deep Learning Features

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

The 2D DL 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. For the 3D models that performed classification based on the PNG images, the full signal is utilized.

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

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

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

These three used features are displayed in figure 4.11.

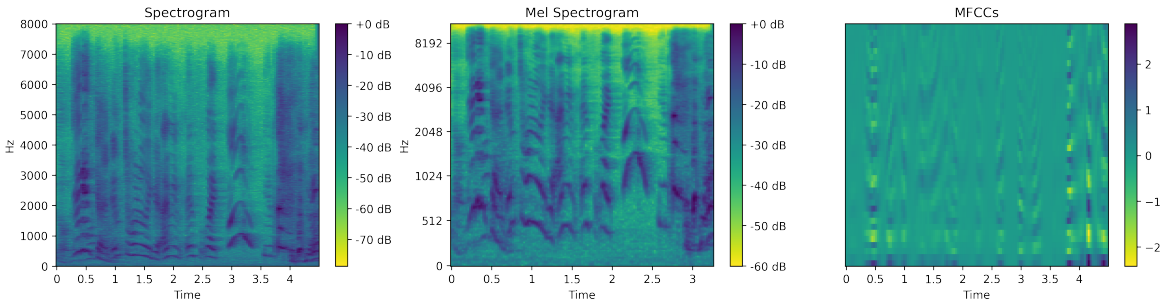


Figure 4.11: Graphical representations of the features used as input for the DL classifiers.

4.4.2 Classifiers Evaluation and Selection

Evaluation Strategy

To evaluate the performance of the classifiers, 5-fold CV was performed using the stratified K-fold strategy to preserve the percentage of samples for each class, following the same strategy for the traditional approach. 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 .2.1. These metrics were computed using the unweighted average across the 5 folds, to give an overall performance overview of the classifier.

Classification of Numeric Data

In order to classify the numeric data, we employed two DL models that take as input the 2D matrices with the corresponding dimensions of the features previously described.

The first model is a deep CNN based on the architecture proposed by Mustaqeem and Kwon [106]. It comprises 7 convolutional layers followed by 2 fully connected layers to perform the classification task. To prevent overfitting, the model includes regularization techniques such as batch normalization after each convolutional layer and a total of three dropout layers. The authors used spectrograms as input for the model and achieved an unweighted average accuracy of 72% by performing 5-fold CV on the IEMOCAP dataset.

The second model is a combination of a CNN and a RNN in a single architecture, based on the work of Ma, Wu, Jia, *et al.* [58] published in 2018. This model consists of 2 convolutional layers, each followed by batch normalization. The output from these layers is then given to a bidirectional LSTM layer followed by a dropout layer and finally, a dense layer with 4 units to classify the emotions. The authors used log spectrograms as input for the model and achieved an unweighted average accuracy of 64.22% by performing 5-fold CV on the IEMOCAP dataset.

Both models utilize an Adam optimizer with a learning rate of 0.001, a sparse categorical cross-entropy loss, and accuracy as the evaluation metric.

Image Classification

In our study of DL-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 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 improve the accuracy and generalization ability of the models, as it allows them to recognize patterns and shapes that are common across a wide range of images.

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

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

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

Results and Conclusions

The experiments were conducted using transfer learning with pre-trained models on a Tesla P4 GPU through the Google Colab service. Table 4.8 summarizes the results of the experiments conducted with all classifiers evaluated.

Among the classifiers evaluated, the ResNet50 model stood out as the most effective, ranking in the top 3 in most metrics except prediction time, as it is a heavier model compared to Xception and VGG16 models.

In terms of the input feature used, while the spectrogram image feature attained the highest average accuracy with the Resnet50 model, the mel spectrogram feature achieved the best overall performance across the other evaluated 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 CV folds.

Despite this, we decided to utilize the spectrogram image as our input feature. This decision was based on the fact that the mel spectrogram is a variation of the spectrogram that uses a mel-scale to convert frequency units to a logarithmic scale that better approximates human auditory perception. This conversion can result in the loss of some information that is only present in the original spectrogram, particularly in the higher frequencies. Therefore, we reasoned that by using the spectrogram image as our input feature, we may be able to preserve some important information that is lost in the mel spectrogram.

Therefore, based on the results obtained, we conclude that the Resnet50 model with spectrogram images as input is the best candidate for the DL model for the SER task. It is also essential to acknowledge that DL can yield even better results with more extensive models' hyperparameter tuning, which would demand more computation time and power for their development and evaluation.

Table 4.8: DL classification models performance on IEMOCAP.

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

Chosen Model: ResNet50. ResNet50 is a powerful deep neural network model primarily used for image classification tasks. It is based on a residual network architecture that allows for training very deep neural networks by addressing the vanishing gradient problem. With 50 convolutional layers, ResNet50 leverages skip connections to enable the direct flow of information between layers, enabling the network to learn more complex representations of the input data.

ResNet50 has achieved SOTA performance in various image classification tasks, demonstrating its effectiveness in tasks such as object recognition and image segmentation. Training a ResNet50 requires significant computational resources due to its large number of parameters, however, pre-trained

ResNet50 models are available in popular DL frameworks such as TensorFlow and PyTorch. These pre-trained models serve as starting points for specific image classification tasks and are fine-tuned on smaller datasets, reducing both training time and computational requirements.

ResNet50 Implementation. The code snippet 4 implements in Python our transfer learning approach using a ResNet50 model, pre-trained on the ImageNet dataset with a total of 23,719,108 parameters. However, in this implementation, only 131,396 parameters are trainable, which corresponds to the new layers added.

First, the ResNet50 model is loaded using the Keras API from TensorFlow, with the "*weights*" parameter set to "*imagenet*" to load the pre-trained weights, "*include_top*" set to False to remove the last classification layer, and "*pooling*" set to "*avg*" to obtain a global average pooling layer as the output. The pre-trained model is then set to be non-trainable.

As mentioned previously, the input layer is defined with the spectrogram image shape, and the output of the pre-trained model is fed into a new dense layer with 64 units and ReLU activation. A dropout layer with a rate of 0.5 is added to reduce overfitting. Finally, the output layer consists of a dense layer with 4 units and softmax activation, corresponding to the 4 possible emotions to be recognized. The model is then compiled with the Adam optimizer, sparse categorical cross-entropy loss, and accuracy as the evaluation metric.

With this implementation, we can take advantage of the pre-trained model on a large dataset and fine-tune it for our specific SER task.

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

Code Snippet 4: Python code for the selected ResNet50 classifier using the DL-based SER approach.

ResNet50 Advantages. The ResNet50 model pre-trained on the ImageNet dataset has several advantages that make it a suitable choice for fine-tuning the task of SER using audio spectrograms as input. One of the main advantages is its ability to handle the vanishing gradient problem that arises in deep neural networks with many layers, thanks to its residual connections. This makes it possible to train very deep neural networks without degradation in performance.

Additionally, the ResNet50 model's high accuracy in image classification tasks and relatively fast training time compared to other deep neural networks make it well-suited for large-scale tasks. Its architecture also facilitates implementing transfer learning, where the pre-trained model can be fine-tuned on new datasets with relatively little additional training data.

In conclusion, utilizing the ResNet50 model pre-trained on the ImageNet dataset and fine-tuning it for speech emotion recognition using audio spectrograms as input is an effective approach to achieve high accuracy in this task. The model's ability to handle the vanishing gradient problem, high accuracy in image classification, and fast training time make it a powerful and versatile tool for DL applications.

4.5 Classifiers Results

In this section, we present and analyze the results of the best candidate models obtained from both the traditional and DL-based SER approaches.

The top-performing model from the traditional approach is a XGBoost classifier, utilizing 33 audio features as input. This model achieved an accuracy of 60.69% after performing 5-fold CV on the IEMOCAP dataset. On the other hand, the final model obtained from the DL approach is a Resnet50 model, which uses audio spectrogram images as input. This model achieved an accuracy of 58.24% using the same evaluation strategy and dataset.

4.5.1 Models Cross-Dataset Validation

The performance of the final models trained using the IEMOCAP dataset and tested on three different datasets, namely eNTERFACE'05, EMO-DB, and CREMA-D, are presented in Table 4.9. The cross-dataset validation was performed to evaluate the generalization capabilities, robustness, and significance of trained models. By testing the models on datasets with different recording conditions, language, and emotional expression variability, we can assess their ability to generalize to new, unseen data. This is an important step towards creating robust SER models that can be applied in real-world scenarios where the training data may not perfectly match the test data.

Table 4.9: Final models trained on IEMOCAP and evaluated on different datasets.

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

The DL model generally outperformed the traditional model on all datasets, except for the EMO-DB dataset, where they performed similarly. The best overall performance was achieved on the CREMA-D dataset, with the DL model achieving an accuracy of 54.14% and a macro F1 score of 47.71%.

However, both models exhibited poor results on the eNTERFACE'05 and EMO-DB datasets, with accuracy scores of less than 40%, and, in terms of macro f1-scores they are slightly better for the eNTERFACE'05 but still low. A low macro F1 score means that the model's ability to correctly classify instances of different classes is poor. This could be due to several factors, such as class imbalance or the model's inability to capture important features that distinguish between different classes.

In terms of computation time, the traditional models were faster than the DL models, which is evident from the lower time values in the table. This is a crucial factor when employing these models for real-time emotion recognition systems.

Further insights into the obtained results can be obtained through the confusion matrices, as shown in Figure 4.12. We can observe both models predicted the anger emotion multiple times across all three datasets. In the case of the eNTERFACE'05 dataset, which lacks neutral emotions, the DL model predicted the happiness emotion more frequently compared to the traditional model, which is the leading reason for the slightly better results.

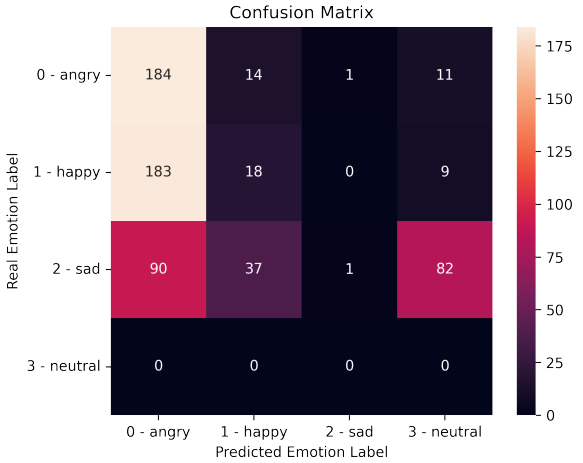
Analyzing the confusion matrix of the EMO-DB dataset, it is clear that the anger emotion was predicted for the majority of audio files, which is likely due to the German language used in the dataset

known for its assertiveness and directness. However, this should not be interpreted as a flaw in the model's performance but rather a reflection of the language's characteristics used in the training and testing datasets. This highlights the English language bias of our SER models and suggests that the results for other languages may not be satisfactory.

Moreover, the traditional model had difficulty recognizing the happiness emotion in the CREMA-D dataset, as in the eNTERFACE'05 dataset, and often confused the neutral emotion with sadness. Although the DL model exhibited the same confusion, it predicted the sadness emotion more often.

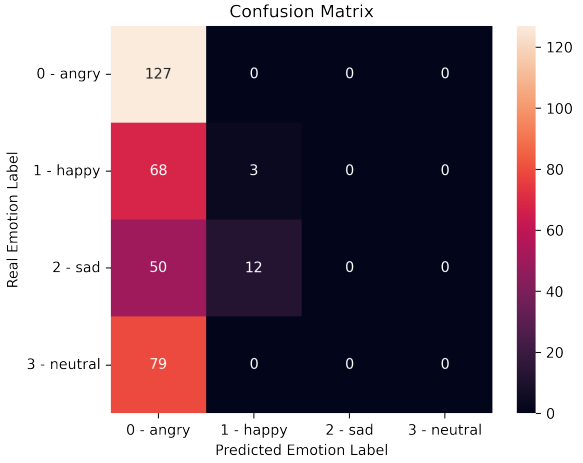
Finally, it is worth noting that the traditional models were faster than the DL models, with lower time values in the table, which is critical when implementing these models for real-time SER systems.

In conclusion, the results suggest that the DL model may be more suitable for the SER task on diverse datasets, owing to its improved feature extraction and generalization abilities. However, the DL model is slower compared to the traditional models, which can significantly impact real-time applications.

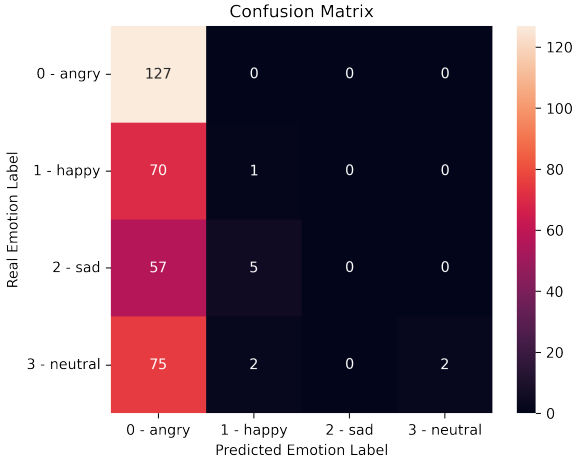


(a) eNTERFACE'05 traditional model confusion matrix.

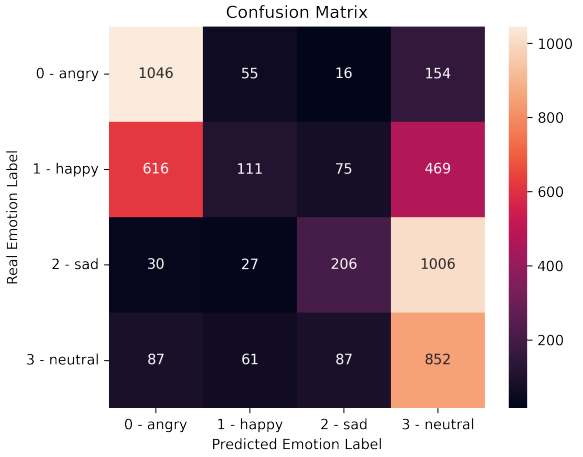
(b) eNTERFACE'05 DL model confusion matrix.



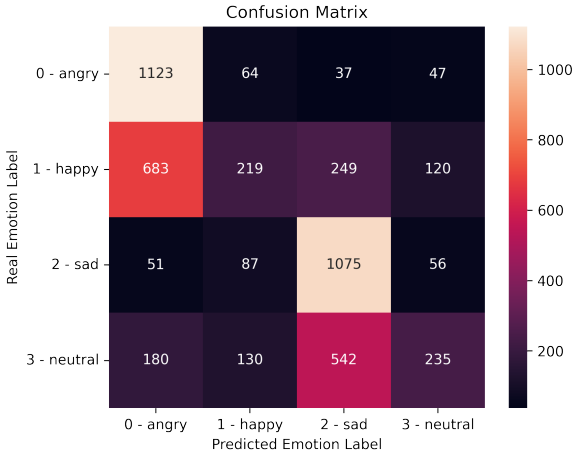
(c) EMO-DB Traditional model confusion matrix.



(d) EMO-DB DL model confusion matrix.



(e) CREMA-D Traditional model confusion matrix.



(f) CREMA-D DL model confusion matrix.

Figure 4.12: Final models confusion matrices on the eNTERFACE'05, EMO-DB and CREMA-D datasets.

4.5.2 SOTA Comparison

In Table 4.10, we summarize the performance of various classification models on the IEMOCAP dataset, including our proposed model (highlighted in blue). The first four rows represent traditional feature-based approaches, while the remaining rows represent DL-based approaches.

It is important to note that the SOTA results are not always directly comparable, as different authors may use different subsets of data and evaluation methodologies. For instance, some authors may not consider the emotion of excitement as happiness, others may only use improvised data and not scripted, or even perform different validation methods such as 10-fold CV. Additionally, the random seeds for the k-fold splits may also impact the final classification performance.

Our proposed model, a XGBoost classifier utilizing 33 audio features, achieved an accuracy of 60.69% on 5-fold CV, outperforming the traditional approaches presented in the first three rows. Another model from the traditional approach, that outperformed ours, utilized a CNN for its feature extraction capabilities; however, we still consider it traditional as the authors still performed feature extraction to obtain the 193-dimensional feature vector for the model's input.

Among the DL-based approaches, our Resnet50 model achieved an accuracy of 58.24%, which is lower than the other SOTA results. The Quaternion CNN achieved the highest accuracy of 70.46%, utilizing a unique approach of encoding the Mel spectrogram in an RGB quaternion domain, which may have contributed to its higher accuracy.

Table 4.10: SOTA SER classification models performance on IEMOCAP.

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

4.6 Discussion

In this chapter, we selected the IEMOCAP dataset for training and testing our models. The dataset contains emotional speech data from 10 actors performing scripted and improvised scenarios, with each recording labeled with four different emotions: anger, happiness, sadness, and neutral. We chose this dataset because of its high-quality recordings, its availability, and its popularity in the SOTA. To evaluate the generalization capability of our models, we also selected three additional datasets to perform cross-dataset evaluation: eNTERFACE'05, EMO-DB, and CREMA-D. These datasets provide a diverse set of challenges for our models, including variations in speech content, recording conditions, and emotional expressions, making them suitable for evaluating the robustness of our models to different scenarios.

Afterwards, we applied a set of audio preprocessing techniques to the audio data, and reached a simple strategy that reduces background noises and trims the silence in the beginning and ending of the audios. This preprocessing step showed that it helps removing irrelevant information and reduce noise. However, the choice of preprocessing techniques may depend on the specific dataset and recording conditions, so we recommended researchers to experiment several preprocessing strategies to determine the most effective approach.

Our findings on the cross-dataset validation support the idea that the DL model is better suited for SER tasks on diverse datasets, due to its improved feature extraction and generalization capabilities of data under different factors. We believe this is mainly because the traditional approach relies on a feature input vector derived from the training dataset, limiting its applicability to other datasets. However, the traditional model is more suitable for real-time applications as it has a faster processing speed than the DL model.

In addition, our XGBoost model achieved competitive results compared to the SOTA approaches presented in the table, despite using only a 1-dimensional 33 audio feature vector. This makes it more computationally efficient than other traditional approaches that require a much larger number of features. Nonetheless, both of our models still have the potential for improvement, especially in the DL approach where fine-tuning was limited due to its computational expense. We also recommend using our models for English-spoken audio recordings with less prominent accents to mitigate language bias since the training dataset contains English audio.

Overall, this chapter offers a comprehensive implementation and thought process, along with conclusions and interpretations of the obtained results, laying a solid foundation for future research in developing more robust and accurate SER models. It highlights the importance of considering the impact of different validation methodologies used in SER research to ensure the models' robustness and applicability to diverse datasets. These findings contribute to the growing body of research on SER and provide valuable insights for researchers and practitioners in the field.

CHAPTER 5

Data Stratification

To gain a deeper understanding of the properties of the data in the IEMOCAP dataset, we conducted data stratification. Our objective was to group the classification labels based on several attributes of the dataset, explore their properties, and potentially identify any limitations these properties may impose on machine learning models in the SER field.

5.1 Recordings Durations

The duration of an audio recording can significantly impact the analysis and modeling of the data. We hypothesize that shorter recordings may not capture enough information to adequately represent the signal of interest, while longer recordings may contain irrelevant or redundant information. For this matter, we stratified the dataset based on the duration of the recordings.

The dataset contains recordings ranging from approximately 0.25 to 34 seconds in duration, and we divided the recordings into three evenly balanced groups, in terms of the number of recordings: short, less than or equal to 2.29 seconds, medium, greater than 2.29 and less than or equal to 4.38 seconds, and long, greater than 4.38 seconds.

Table 5.1 presents the impact of the duration of the recordings on the performance of the chosen traditional XGBoost model. The longest recordings have the highest performance, with an accuracy of 63.77%.

Table 5.1: Traditional model 5-fold cross-validation results on stratified data based on the recordings' duration.

Recordings Duration	Total Data	Accuracy	Macro F1	Precision	Recall	MCC
Short $(]0, 2.29]$ s)	1844	55.69±1.15	54.64	56.59	53.54	0.38
Medium $(]2.29, 4.38]$ s)	1843	56.65±2.28	57.16	57.25	57.26	0.41
Long $(]4.38, 34]$ s)	1844	63.77±0.87	64.16	64.32	64.17	0.51

The obtained classification results suggest that longer recordings are easier for the model to classify accurately, as they provide more information. This allows us to conclude that recordings' duration has a substantial performance impact on the SER task and is a significant factor to consider when building classification models.

5.2 Speaker Gender

Another attribute we used for stratification was the gender of the speaker. The dataset contains a similar amount of recordings from speakers of both genders, having 2649 recordings with female speakers and 2882 with male speakers.

To evaluate the classification performance of different genders, we trained and tested the model on recordings from female speakers, male speakers, and mixed-gender recordings. Table 5.2 shows the 5-fold cross-validation results of the traditional model on each category.

Table 5.2: Traditional model 5-fold cross-validation results on stratified data based on speaker gender recordings.

Training Gender	Testing Gender	Accuracy	Macro F1	Precision	Recall	MCC
Female	Female	59.95±1.45	60.47	60.4	60.77	0.46
Male	Male	59.99±1.28	60.67	61.38	60.2	0.45
Female	Male	50.28±0.64	50.57	50.36	51.1	0.32
Male	Female	49.41±1.43	50.26	53.37	49.24	0.32

From the table, it can be observed that the model performed similarly on both genders, when the testing data's gender is equal to the training data's gender, with accuracies close to 60%. However, when testing on the opposite gender of the training, the model's accuracy dropped significantly to 50.28% and 49.41% for female and male-trained models, respectively. The other metrics showed equivalent behavior, indicating difficulty in correctly identifying the emotion in mixed-gender contexts.

These results suggest that the model's performance is affected by the gender of the speakers in the training data and therefore, it is important that the model is provided with a gender-balanced set of training data to reduce gender bias.

5.3 Discrete Emotions

This section examines the classification performance of different emotional categories in the IEMOCAP dataset. The dataset contains four emotional categories: anger, happiness, sadness, and neutral. We performed data stratification based on these labels, as well as grouping them based on emotional planes, resulting in 15 groups of data. Table 5.3 displays the traditional model's 5-fold cross-validation results on stratified data based on discrete emotions. The last two rows of the three and two label cases combine the labels based on their distribution in the arousal and valence planes, respectively.

In the three-label case, the best performance was achieved when classifying between anger (high arousal and dominance), neutral, and sadness (low arousal and dominance), with an accuracy and macro F1 score of 75%. The lowest results were obtained with emotions far apart in the valence plane, such as anger or sadness (low valence), neutral, and happiness (high valence). The last two rows of the three-label case support the same conclusion.

In the two-label case, where the emotional categories were paired, the best performance was achieved when classifying between angry and sad, with an accuracy and macro F1 score of 92%. However, this group had the least amount of total data. The worst performance was achieved when classifying between neutral and happy, with an accuracy and macro F1 score of around 73%, which had more total data to classify. Similar to the three-label case, the model achieved better results when classifying labels far apart on the arousal or dominance planes, and worse in the valence plane.

Table 5.3: Traditional model 5-fold cross-validation results on stratified data based on the discrete emotions.

Labels	Total Data	Accuracy	Macro F1	Precision	Recall	MCC
Four Labels						
Angry, Happy, Sad, Neutral	5531	60.69±1.37	61.32	61.66	61.19	0.47
Three Labels						
Angry, Neutral, Sad	3895	75.17±0.99	75.30	76.21	74.62	0.62
Angry, Happy, Sad	3823	70.99±1.55	71.37	71.38	71.42	0.56
Sad, Neutral, Happy	4428	65.74±1.85	65.85	65.94	65.86	0.48
Angry, Neutral, Happy	4447	64.29±1.13	63.81	64.41	63.60	0.45
Sad, Neutral, Happy+Angry	5531	69.32±1.24	67.29	67.44	67.2	0.50
Angry+Sad, Neutral, Happy	5531	60.64±1.36	59.72	60.03	59.73	0.40
Two Labels						
Sad, Angry	2187	92.00±1.30	92.00	92.00	92.00	0.84
Angry, Neutral	2811	84.53±1.26	83.49	84.25	82.98	0.67
Sad, Happy	2720	83.79±0.85	83.10	83.08	83.11	0.66
Sad, Neutral	2792	81.02±1.16	79.70	80.29	79.29	0.60
Angry, Happy	2739	75.58±0.93	74.15	74.78	73.79	0.49
Happy, Neutral	3344	73.24±1.06	73.15	73.34	73.15	0.46
Sad+Neutral, Happy+Angry	5531	77.24±1.04	77.21	77.31	77.21	0.55
Angry+Sad, Neutral+Happy	5531	73.82±0.77	71.70	72.90	71.21	0.44

It is important to consider that the overall dataset size influences classification outcomes. When the amount of data is reduced, there is a decrease in emotional diversity, which can potentially facilitate the model's comprehension. However, the results suggest that a classification model is better at distinguishing emotional categories that are distant in the arousal or dominance planes than in the valence plane. For instance, anger and happiness (high arousal and dominance) are easier for a model to distinguish between sadness (lower arousal and dominance), while the model has a harder time classifying anger and sadness (low valence) between happiness (high valence).

Based on our analysis, we suggest that the subjectivity of emotions, along with the distribution on emotional planes, may contribute to the difficulty in accurately classifying certain emotional categories, especially happiness and neutral. Hence, it is imperative to develop emotional datasets with a high degree of confidence, utilizing robust labeling techniques that consider the distribution of emotions across different emotional planes. This approach can help enhance the performance of machine learning models in the field of SER.

5.4 Dimensional Emotions

In addition to the categorical classification, we also explored the dimensional annotations of the dataset in terms of valence (ranging from unpleasant to pleasant), arousal (ranging from calm to excited), and dominance (ranging from submissive to dominant). These dimensions were rated on an integer scale from 1 to 5 by human judges.

Initially, we aimed to investigate the level of difficulty associated with classifying each emotional dimension, so we utilized traditional approach features to assess the performance of a RF Regressor model from sklearn for each dimension. Each dimension underwent 5-fold cross-validation. We used Mean-Absolute-Error (MAE), Root-Mean-Squared-Error (RMSE), and R^2 as evaluation metrics for the model.

Table 5.4 shows the results of our experiments on the regression task. The arousal and dominance dimensions are easier to classify than valence. One possible reason for this is that valence is a more complex and abstract concept that may be more difficult to recognize and classify in speech. Additionally, the annotations for valence may be more subjective and varied among annotators than the others, leading to less reliable labels and lower classification performance.

Table 5.4: RF Regressor 5-fold cross-validation using dimensional emotions as labels.

Regression Labels	MAE	RMSE	R^2
Valence	0.700	0.859	0.182
Arousal	0.414	0.522	0.520
Dominance	0.534	0.657	0.362

We conducted a comparison between the discrete and dimensional annotations by calculating the means of each dimensional label based on the discrete one. These means, known as dimensional centroids, are presented in Table 5.5 and in a 2D plane 5.1. To determine if the centroids are accurate or too distant from the expected values, we compared them with the SOTA Russell and Mehrabian's VAD model [16].

Table 5.5: IEMOCAP dimensional centroids and comparison to the VAD model.

Emotion	Centroids		
	Arousal	Valence	Dominance
VAD Anger	4.34	2.14	3.68
VAD Happiness	3.96	4.52	3.70
VAD Neutral	3.00	3.00	3.00
VAD Sadness	3.54	1.74	2.34
Anger	3.64 (−0.70)	1.91 (−0.23)	3.95 (+0.27)
Happiness	3.41 (−0.55)	3.95 (−0.57)	3.23 (−0.47)
Neutral	2.73 (−0.27)	2.97 (−0.03)	2.83 (−0.17)
Sadness	2.56 (−0.98)	2.25 (+0.51)	2.83 (+0.49)

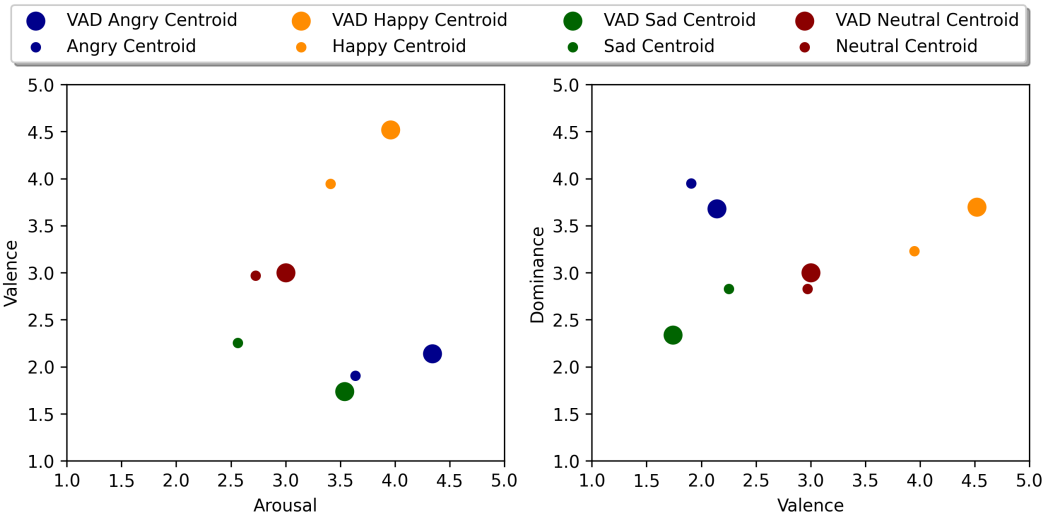


Figure 5.1: 2D representation of the IEMOCAP and VAD model dimensional centroids.

The results indicate that the centroids of each discrete emotion have some differences to the VAD model. For the arousal dimension, it was noted that all emotions have lower values, with sadness showing a greater deviation. In terms of valence, the annotations are more comparable to the VAD model, but sadness contradicts the lower trend of the other two emotions. The annotations on the dominance dimension also vary the deviation trends.

Based on this, we decided to remove any data that contains dimensional annotations far from the VAD model centroids for each discrete emotion. Moreover, to complement our numerical observations, we employed scatter and violin plots to explore potential restrictions that could enhance the separation and distinction of emotions along different dimensions. The visualizations generated from this analysis can be found in Section .3.1 of the appendix. Table 5.6 demonstrates this conflicts removal process between each emotion's categories and the respective dimensional annotations.

Table 5.6: Maintained dimensional annotations range for each emotion category.

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

With this process, 1136 audio files were removed, and 4395 audio files were retained. The new VAD centroids calculated, as shown in table 5.7, were closer to the numeric values of the VAD emotion centroids. Figure 5.2 depicts a 2D visualization of the centroids of the VAD model, along with the data with and without dimensional conflicts. The visualization indicates that the centroids are now closer to the VAD emotion centroids after the conflicts removal process, which is a possible indication that the conflict-free data is more suitable for the SER task.

Table 5.7: IEMOCAP dimensional centroids and comparison to the VAD model after the conflicts removal process.

Emotion	Centroids		
	Arousal	Valence	Dominance
Anger	3.78 (−0.56)	1.86 (−0.28)	4.02 (+0.34)
Happiness	3.49 (−0.47)	4.00 (−0.52)	3.39 (−0.31)
Neutral	2.75 (−0.25)	3.03 (−0.03)	2.96 (−0.04)
Sadness	2.47 (−1.07)	2.05 (+0.31)	2.61 (+0.27)

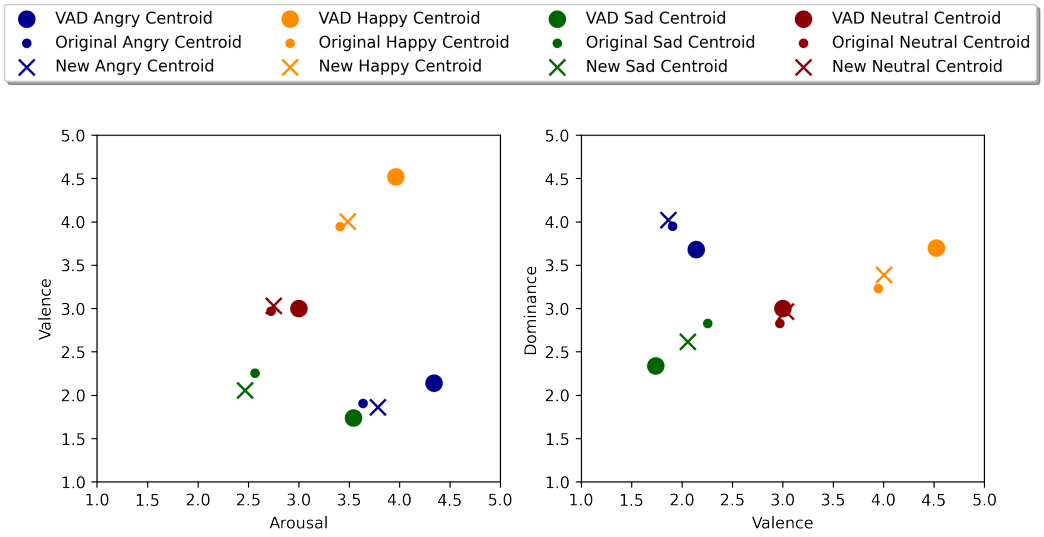


Figure 5.2: 2D visualization of the entire IEMOCAP data, along with the conflict-free data and the VAD model's dimensional centroids.

To evaluate further if this removal of data translates to cleaner data and, therefore, to a more effective emotion recognition process, a 5-fold cross-validation was performed using that conflict-free data and compared against the results obtained using the complete dataset. The results are shown in Table 5.8, and they indicate that the conflict-free dataset outperforms the complete dataset in all metrics. While this outcome supports our hypothesis, the observed improvements are not significant enough to attribute them solely to the removal of data based on dimensional annotations. Other random variables, such as variations in emotion distributions across folds or the decrease in the total number of files, could also have contributed to the observed improvements. Hence, further investigation is required to establish the actual impact of this process.

Table 5.8: Traditional model 5-fold cross-validation results with the conflicts removal process.

Data	Total Data	Accuracy	Macro F1	Precision	Recall	MCC
All Data	5531	60.69±1.37	61.32	61.66	61.19	0.470
Conflict-Free Data	4456	62.01±1.74	63.00	63.36	62.76	0.480

5.5 Discussion

The data stratification process enabled us to uncover the shortcomings of the IEMOCAP dataset and the model we developed. Having obtained valuable insights from our observations and conclusions on the subsections about duration, gender, and dimensional annotations, we aim to utilize the training data in our model to enhance its reliability and performance. In light of this, we have decided to train both traditional and DL models on data that aligns with our objectives.

To achieve this, to the conflict-free data obtained from the dimensional conflict removal process, we applied an additional file duration condition. Specifically, we only retained audio files with a duration exceeding 1 second, as it was necessary to ensure the presence of sufficient emotional data for the model to learn effectively. As a result of this process, we obtained a total of 4200 audio files, with a nearly balanced gender distribution, comprising 52.9% male and 47.1% female speakers.

Table 5.9 presents the 5-fold cross-validation results of the traditional model trained on the IEMOCAP dataset with different sets of selected data. The first row shows the results obtained with

all 5531 files of the dataset. The second row shows the results after applying the dimensional conflict removal process, resulting in 4395 files. The third row shows the results of the conflict-free data with the previously mentioned additional duration condition, resulting in 4200 audio files. The results demonstrate that the performance of the traditional model improves as the data selection process becomes more rigorous. The dimensional conflict removal process resulted in an increase in accuracy from 60.69% to 62.01%, while the additional duration condition further improved the accuracy to 62.89%, the other metrics exhibit similar improvements, indicating that the model's classification ability improved with the use of cleaner and more reliable data.

Table 5.9: Traditional model 5-fold cross-validation results in different sets of data.

Data	Total Data	Accuracy	Macro F1	Precision	Recall	MCC
All	5531	60.69±1.37	61.32	61.66	61.19	0.470
Conflict-Free	4456	62.01±1.74	63.00	63.36	62.76	0.480
Conflict-Free With Duration Condition	3347	62.89±0.67	63.76	63.66	63.90	0.500

Removing conflicts from the dataset results in better performance metrics for our models, however, this approach reduces the amount of data which may be the reason for these improvements through an overfit in a small train dataset. To verify if the model does improve, we decided to train and save both traditional and DL models with the data that met our set of conditions. These saved models are called stratified models. We then repeated the process made in the previous section and evaluated them on three different datasets, namely eNTERFACE'05, CREMA-D, and EMO-DB.

As shown in Table 5.10, the stratified models outperformed the previous models trained on all of the IEMOCAP data on most metrics, which demonstrates that our stratification study resulted in improved cross-dataset performance of the models.

Table 5.10: Final models trained on IEMOCAP and evaluated on different datasets.

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

CHAPTER 6

Speech Emotion Recognition Pipeline

In this section, we present an audio pipeline for performing SER on any systems, such as video conferences, and can be used both online and offline time.

6.1 Architecture

Our developed pipeline includes several stages, each of which plays an essential role in properly identifying emotional content from the audio signal. Figure 6.1 demonstrates the architecture of our developed pipeline.

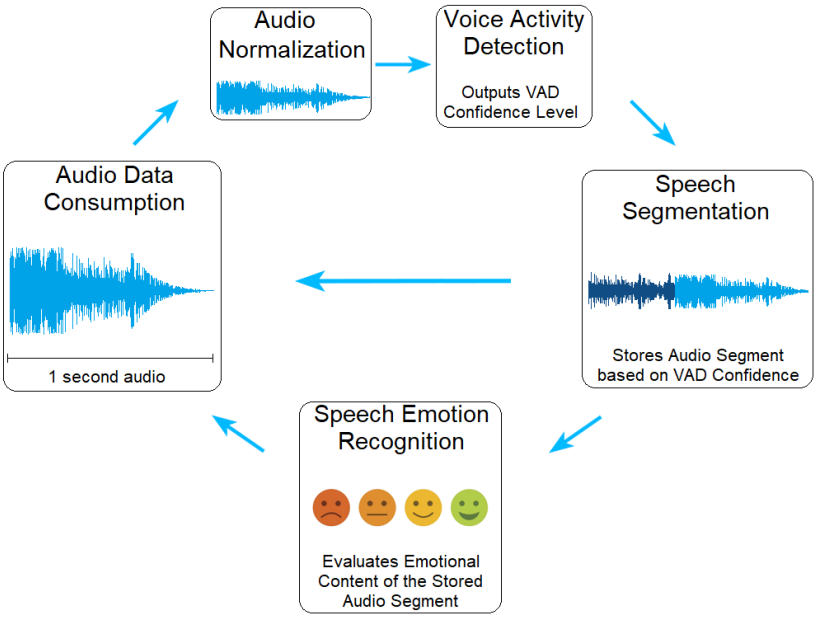


Figure 6.1: Developed SER pipeline architecture.

6.1.1 Data Consumption

The first step of the pipeline is to continuously consume the binary audio data of a video conference participant.

According to a 2022 article on pause duration of English speech [107], the mean pause duration for commas ranged from 0.51 to 0.78 seconds, while the mean pause duration for periods ranged from 1.40 to 1.43 seconds. Since we want to capture sentences/utterances, we decided to feed the pipeline with audio data corresponding to 1 second, which allows for smaller pauses such as breathing or commas during a segment and at the same time, separates sentences.

6.1.2 Normalization

The next step is converting the consumed binary data to an array of floats, and afterward, normalizing the audio signal. The normalization consists of, when necessary, resampling the audio to a sampling rate of 16000 Hz and converting the signal to mono by averaging samples across the channels, recurring to the Librosa toolkit [93]. These operations are necessary so that the audio can be accepted and interpreted by the machine learning models that will follow and analyze it.

6.1.3 Voice Activity Detection

The third step of the pipeline is to detect voiced speech of the previously consumed second of audio, using a VAD model.

We chose the sileroVAD tool [100], as it is an open-source algorithm that provides accurate results in real time. It is specifically designed for low-resource environments and can work efficiently on devices with limited processing power.

SileroVAD is based on a deep learning model trained on a large corpus of audio data. This means that it can detect voice activity even in noisy environments or when the speaker's voice is weak. This is crucial for our pipeline since preprocessing on unvoiced data would be a waste of resources, as that segment will be discarded due to not having enough speech content for emotional classification. This allows us to only do that audio preprocessing before a spoken audio segment is detected to be classified.

This VAD algorithm also returns a confidence level associated with the detection of voice activity, which allows us to fine-tune the minimum value of confidence to consider a voiced segment. For our experiments, we utilized a minimum value of 0.6.

6.1.4 Speech Segmentation and Emotional Recognition

Emotions are usually short-lived, and the speech remains invariant for a brief period. Speech segmentation is the process in which the continuous speech signals are partitioned into short-length segments while maintaining the emotional information suitable for being inputted to SER models.

The strategy for our segmentation is to create speech segments with a minimum of 1 second of duration and at most 8 seconds, which provides a way to detect emotions in real-time as a person speaks one or more complete sentences. The pipeline consumes 1 second of audio, it stores the segment if there is enough confidence that it detected voice activity, if it does not pass the threshold and it has previously saved any audio segment, it feeds it to a SER model. When the saved segment reaches a maximum of 8 seconds, it also performs the classification. It is important to note that these minimum and maximum duration values can be altered to fit different situations and needs, and also, the audio preprocessing techniques (noise reduction and trimming) are executed after the pipeline generates the audio segments to be used by the SER models.

6.2 Pipeline Validation

To validate the effectiveness of our proposed pipeline, we conducted a few tests using the IEMOCAP session files. These files involve two speakers, one female, and one male, interacting with each other in scripted and improvised scenarios to elicit emotions.

Figure 6.2 depicts an audio signal from a session that was manually annotated to label the frames with emotions for both genders and the combination of both speakers. Additionally, the figure displays the voiced segments that our pipeline detected and performed SER.

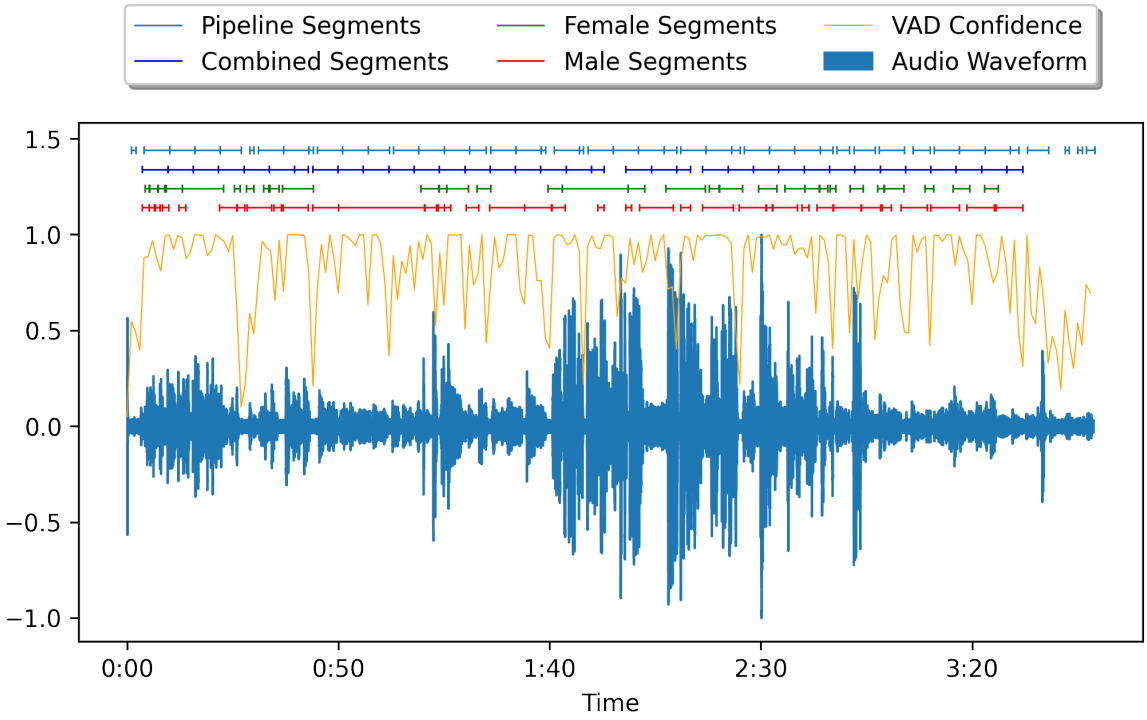


Figure 6.2: IEMOCAP session annotated and detected from our developed pipeline segments.

The efficacy of the pipeline was confirmed by its ability to detect voiced segments in the same intervals as those of the combined segments. To further test the pipeline, we passed all 11 hours of IEMOCAP audio data through it and compared the predicted emotions with the dataset annotations. For this test, it was used an AMD Ryzen 5 5600X 6-core processor, and for classifying the detected segments, the traditional model was selected.

Table 6.1 shows the emotions annotated in the IEMOCAP dataset, along with the predicted emotions resulting from our pipeline.

Table 6.1: Annotated emotions of the IEMOCAP dataset and the pipeline's predicted segments.

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

6.3 Discussion

The pipeline was capable of processing the entire 11-hour audio data in just 15 minutes, including the time the model took to make predictions, which demonstrates that the pipeline is capable of detecting emotions in real-time, by continuously processing audio segments as they arrive.

In total, the dataset comprises 10,039 annotated segments, out of which the pipeline detected 7,698 segments. This is in line with our expectations, as the pipeline only considers segments with at least 1 second of audio data, and the dataset contains shorter audio clips. As for the distribution of predicted emotions, we anticipated that each emotion would have a similar number of predictions of the 4 emotions it was trained, given that it was trained using utterances from the same dataset.

The development of the pipeline represents a significant step forward in improving the user experience in video conference systems. With the pipeline integrated into the system, participants will have access to real-time emotional feedback. This feedback can help participants understand their emotional states, identify potential areas of conflict or misunderstanding, and make more informed decisions about how to interact with others at the conference. It is also important to note that certain configuration values may need adjustments on different use case scenarios, such as the minimum and maximum duration of the segments, the voice activity confidence level, and also, since the classifiers can return the probabilities of each emotion, a minimum level of confidence for the segment's predicted emotion.

In conclusion, the integration of the pipeline into video conference systems, when employed with an efficient SER classifier, has the potential to revolutionize communications in virtual environments, and, to can create more positive and productive interactions between people.

CHAPTER 7

Discussion and Conclusions

In this final chapter, we present an analysis of the results obtained throughout this dissertation research, addressing the contributions made to the field of SER and highlighting their novelty and significance. We provide a summary of the main findings and discuss their implications. Finally, we suggest possible directions for future research to complement and advance our work.

7.1 Discussion

Throughout the previous chapters, we have presented a detailed discussion of the results obtained from our experiments. In this section, we will provide a more concise analysis and examine the broader implications of our research.

One of the main challenges in this field is the lack of consistency in the labeling process of available datasets. To address this issue, we utilized the widely used and validated IEMOCAP dataset, allowing for a more accurate comparison of model performance. Our audio preprocessing operations successfully reduced noise and eliminated silent frames.

We evaluated two approaches to emotion recognition: traditional and DL-based SER, focusing on the effectiveness of feature extraction methods and model performance. Our study led us to select an XGBoost classifier based on a set of 33 hand-crafted audio features, and a fine-tuned ResNet50 model using audio spectrogram images. The XGBoost classifier achieved the second-highest accuracy among SOTA articles employing the traditional approach in the IEMOCAP dataset, with 60.69% accuracy in 5-fold CV, while using much less audio features. The ResNet50 model produced comparable results to DL approaches, achieving 58.24% accuracy. The XGBoost classifier demonstrated faster prediction times, making it suitable for real-time classifications, while the DL model showed superior cross-dataset validation results, indicating greater generalization capabilities.

The importance of proper data stratification was highlighted, addressing potential biases and accounting for the subjective nature of emotions in labeled data. Data stratification improved model performance within the IEMOCAP dataset and cross-dataset validation.

The developed SER pipeline is versatile and scalable, incorporating a VAD tool for creating voiced audio segments and supporting both traditional and DL-based SER models for classification. The pipeline was tested on the entire raw IEMOCAP dataset, successfully detecting audio segments and predicting emotion labels. This confirms the pipeline's robustness in detecting emotions from speech signals, enabling its use in various scenarios.

7.2 Contributions and Developed Tools

This study has made several contributions toward improving accuracy and efficiency in emotion recognition systems.

Firstly, we provide a comprehensive and up-to-date review of recent advances in SER, covering traditional approaches, DL-based methods, and multimodal emotion recognition. We emphasize the need for larger and more diverse datasets and address the challenges of variability in emotions across cultures and contexts.

Furthermore, we identify a small set of audio features applicable to any model, contributing to traditional SER approaches. Additionally, we demonstrate the effectiveness of transfer learning techniques in DL-based SER.

Our data stratification study of the IEMOCAP dataset establishes conditions for obtaining higher-quality data suitable for diverse environments, providing guidance for future research.

We have also developed two Python classes that automate feature extraction and emotion classification [108]. These classes allow users to choose between the XGBoost and ResNet50 models trained on either the entire IEMOCAP dataset or the stratified data, returning predicted emotions or probabilities.

Finally, we present an SER pipeline capable of analyzing audio streams in real-time or offline, utilizing a VAD model and Python classes for emotion labeling. This pipeline can be seamlessly integrated into streaming services, enhancing user experience with real-time emotional feedback.

7.3 Conclusion

This dissertation explores and develops SER classifiers using various approaches and techniques, making significant contributions to the field.

The research begins with a thorough review of SOTA studies, followed by an analysis of different SER approaches. The methodology covers requirements gathering, dataset selection, audio feature engineering, and model implementation, ensuring reliable and accurate models.

By addressing key aspects such as dataset quality, feature engineering, and model selection, we contribute to the SER field by providing valuable insights into the application of traditional and DL techniques. Our findings emphasize the importance of these factors in achieving accurate and reliable results.

In conclusion, this dissertation advances our understanding of SER and provides a framework for developing reliable models. The research findings and methodology contribute to the existing body of knowledge in emotion recognition.

7.4 Future Work

Several areas of future research can advance our work on SER. Incorporating additional features, such as physiological signals or facial expressions, can create more robust and accurate models. Developing multimodal models capable of handling noisy and non-stationary acoustic environments is another potential area for improvement. Contextually aware models that consider situational factors and user history can provide more personalized emotion recognition services.

Further investigation into audio preprocessing techniques is necessary, as they play a critical role in the success of SER models. Studying the effects of audio preprocessing techniques on classification accuracy using available datasets and developed models would provide insights into optimal preprocessing pipelines.

Fine-tuning the developed models, particularly the DL model, can potentially lead to even better performance and accuracy, although it requires computationally expensive operations. Adjusting hyperparameters and increasing the number of trainable parameters can further improve performance.

Integrating emotion recognition technology into real-world applications, such as virtual assistants, smart homes, monitoring systems, and call centers, is promising for future research. Validating the developed SER tools through integration into various applications and designing emotionally intelligent and responsive interfaces can enhance the end-user experience. Continued research and development in these areas can lead to significant advances in SER technology and its applications.

CHAPTER 8

Appendix

.1 Traditional Feature-Based SER

.1.1 Audio Features Visualization

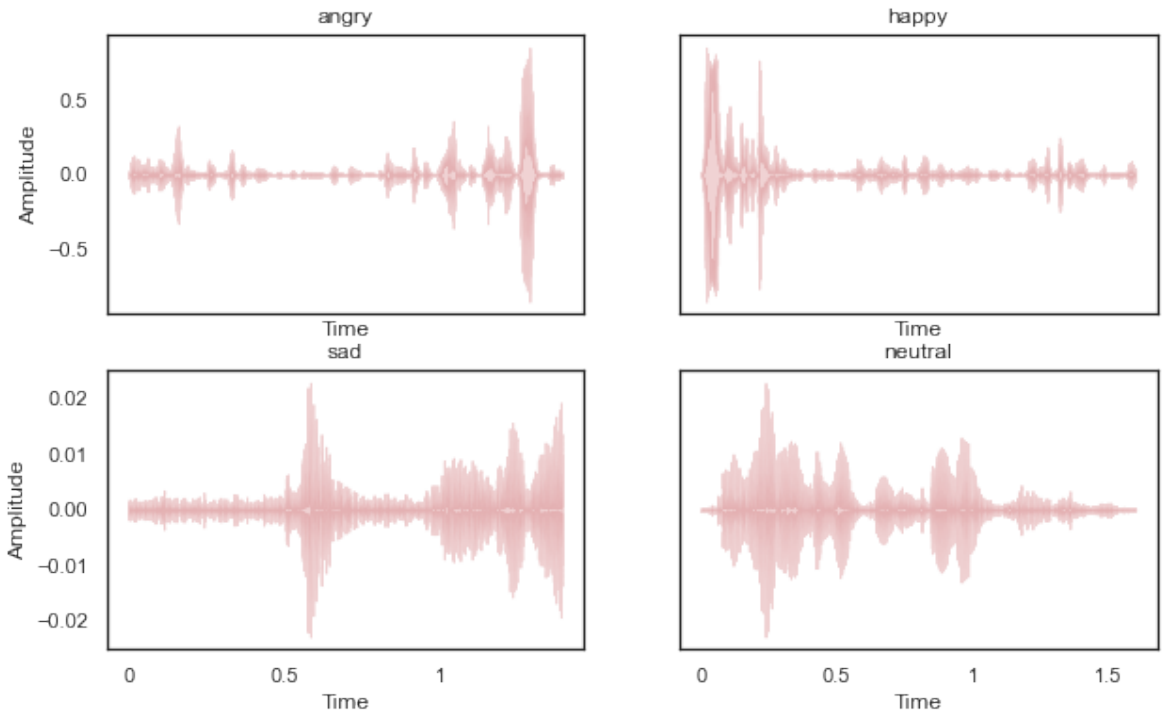


Figure .1: Audio Signal wave plots of one audio segment for all emotions.

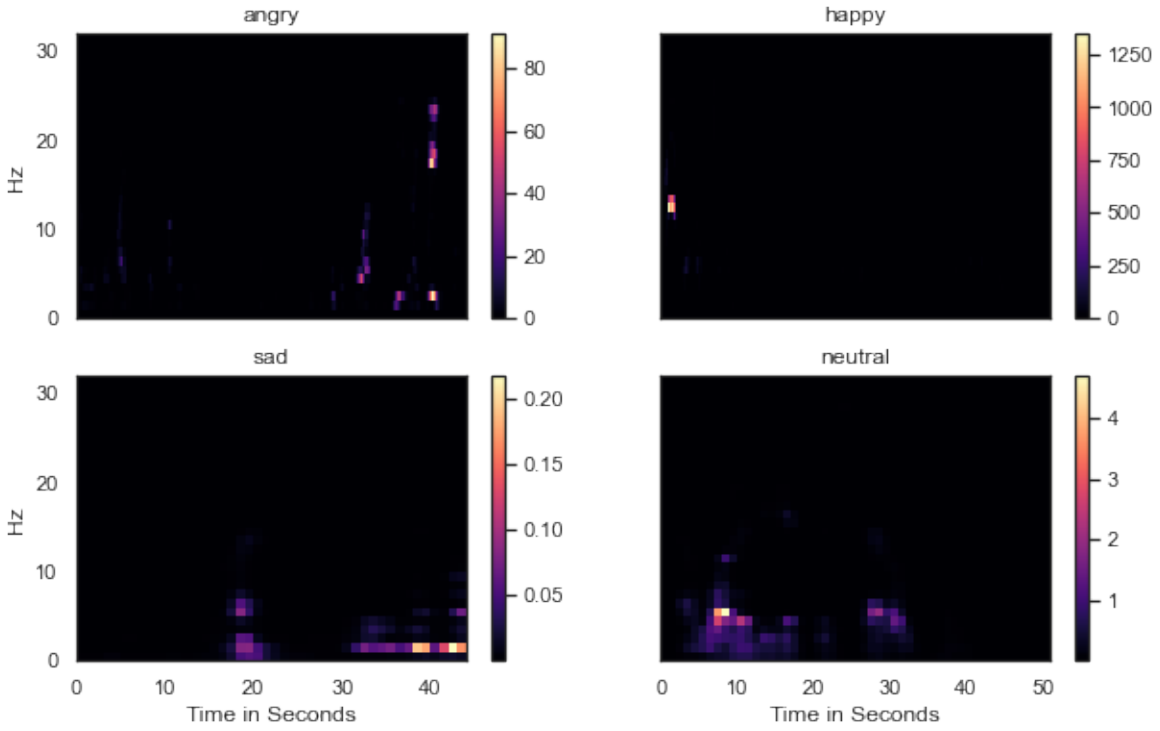


Figure .2: Log mel magnitude spectrograms of one audio segment for all emotions.

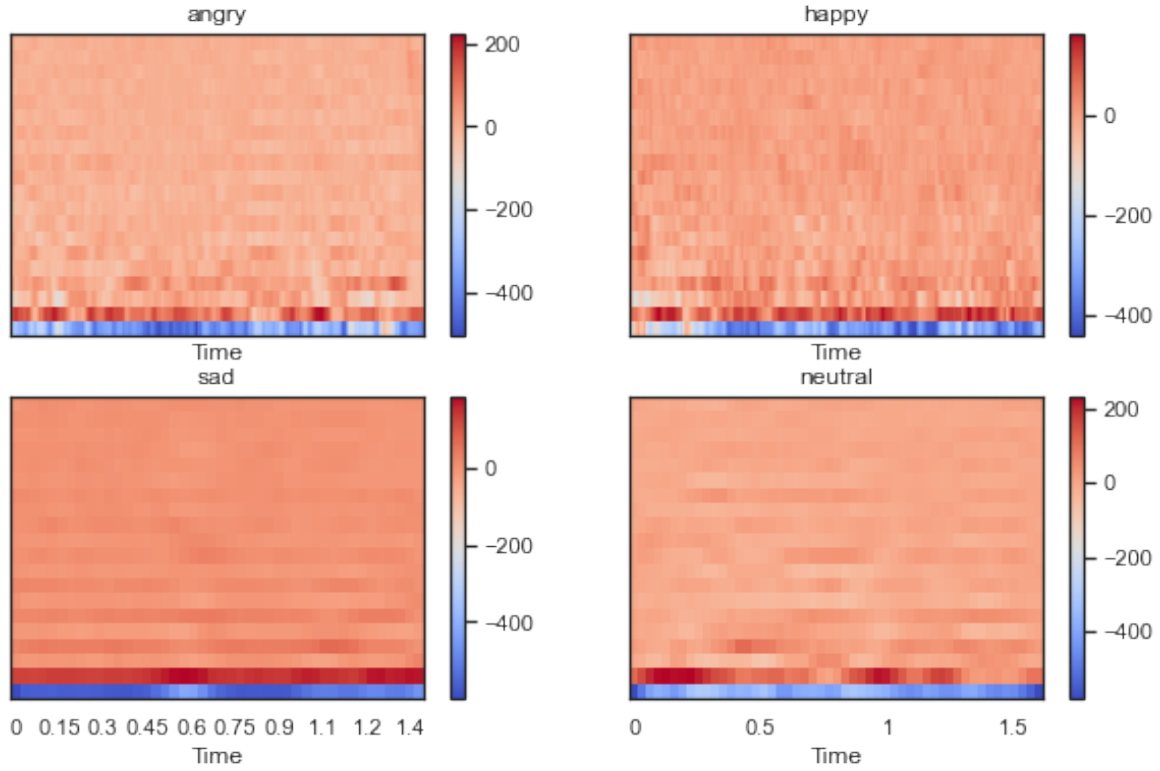


Figure .3: MCC spectrogram of one audio segment for all emotions.

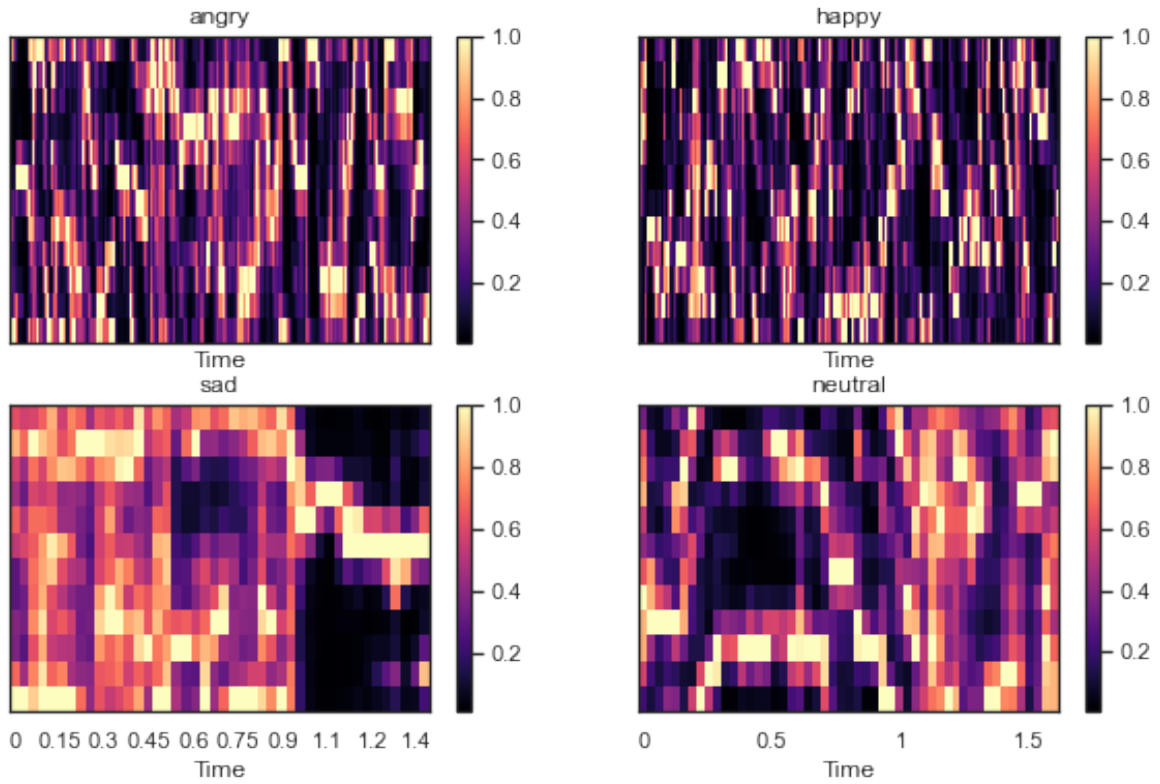


Figure .4: Chromogram spectrograms of one audio segment for all emotions.

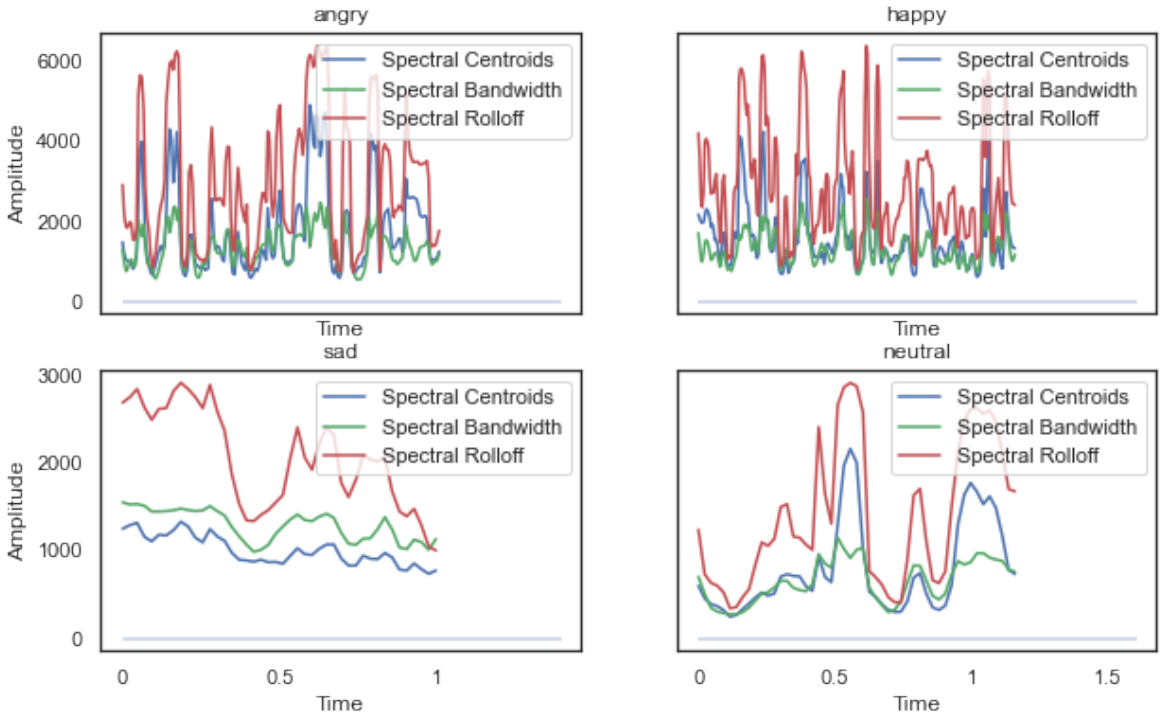


Figure .5: Spectral wave plots of one audio segment for all emotions.

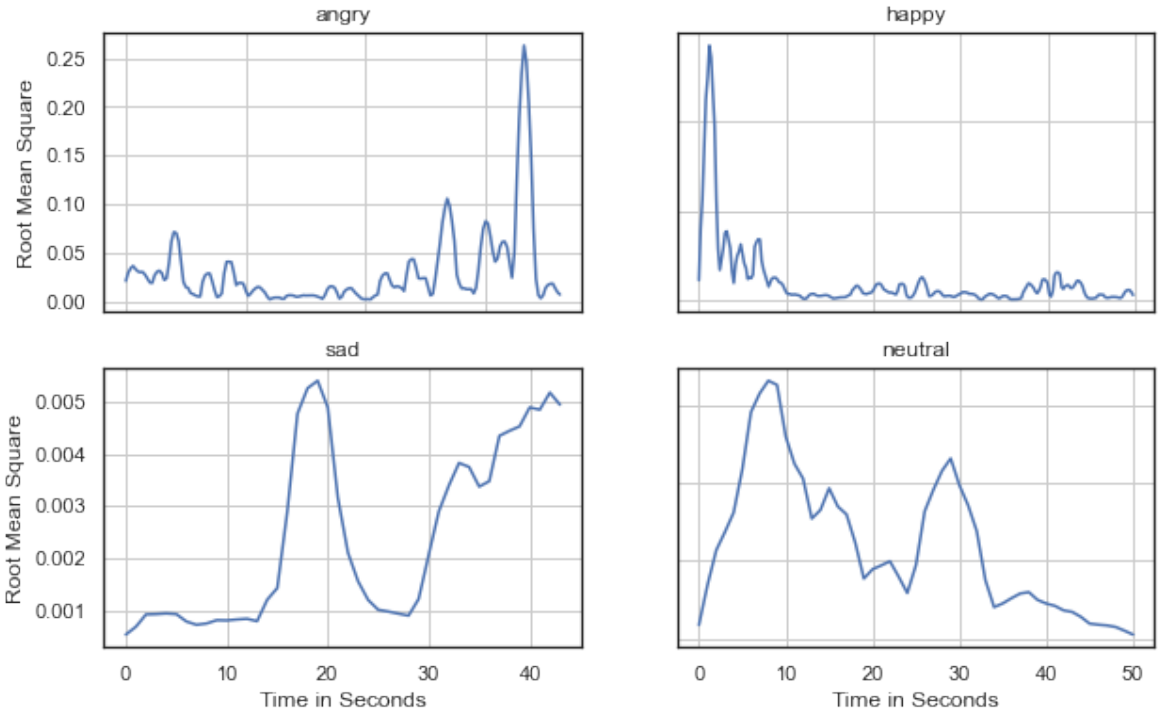


Figure .6: Root-Mean-Square energy wave plots of one audio segment for all emotions.

.1.2 Features Mean Values Overview

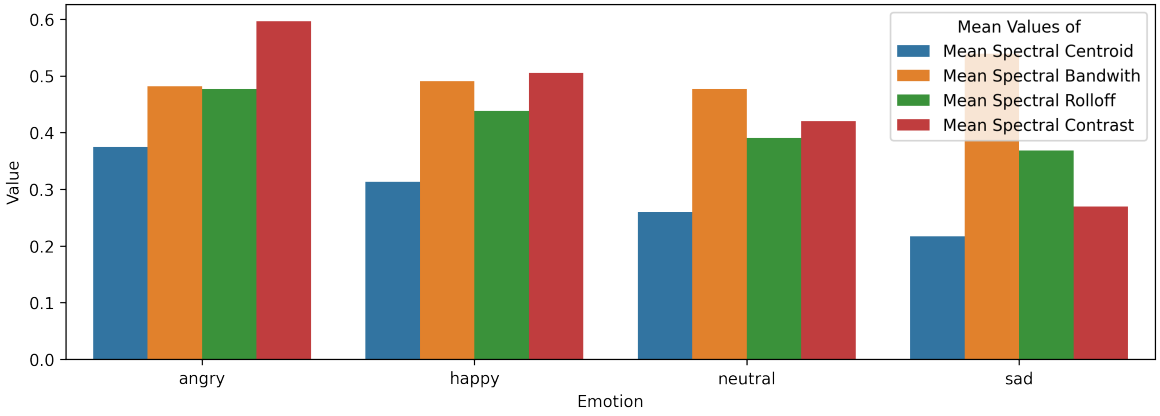


Figure .7: Bar plot with the mean values of the mean spectral centroid, bandwidth, roll-off, and contrast features.

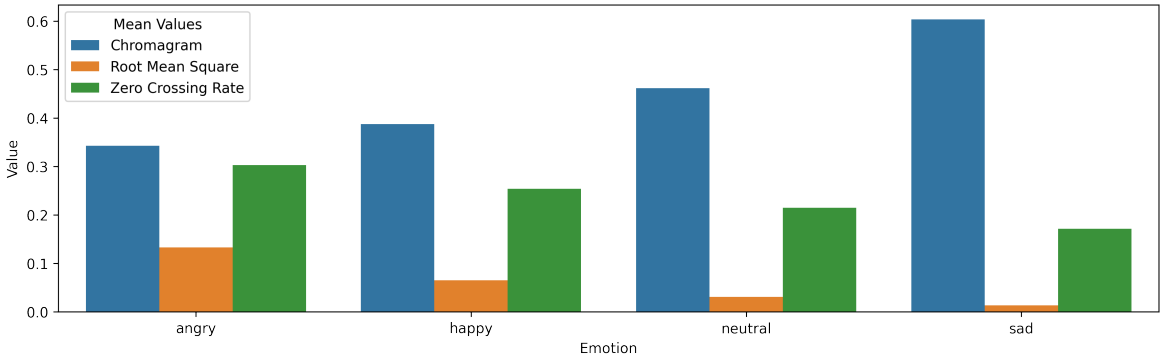


Figure .8: Bar plot with the mean values of the mean chromogram, root-mean-square and zero crossing rate features.

.1.3 Wave Plots with Surrounding Areas

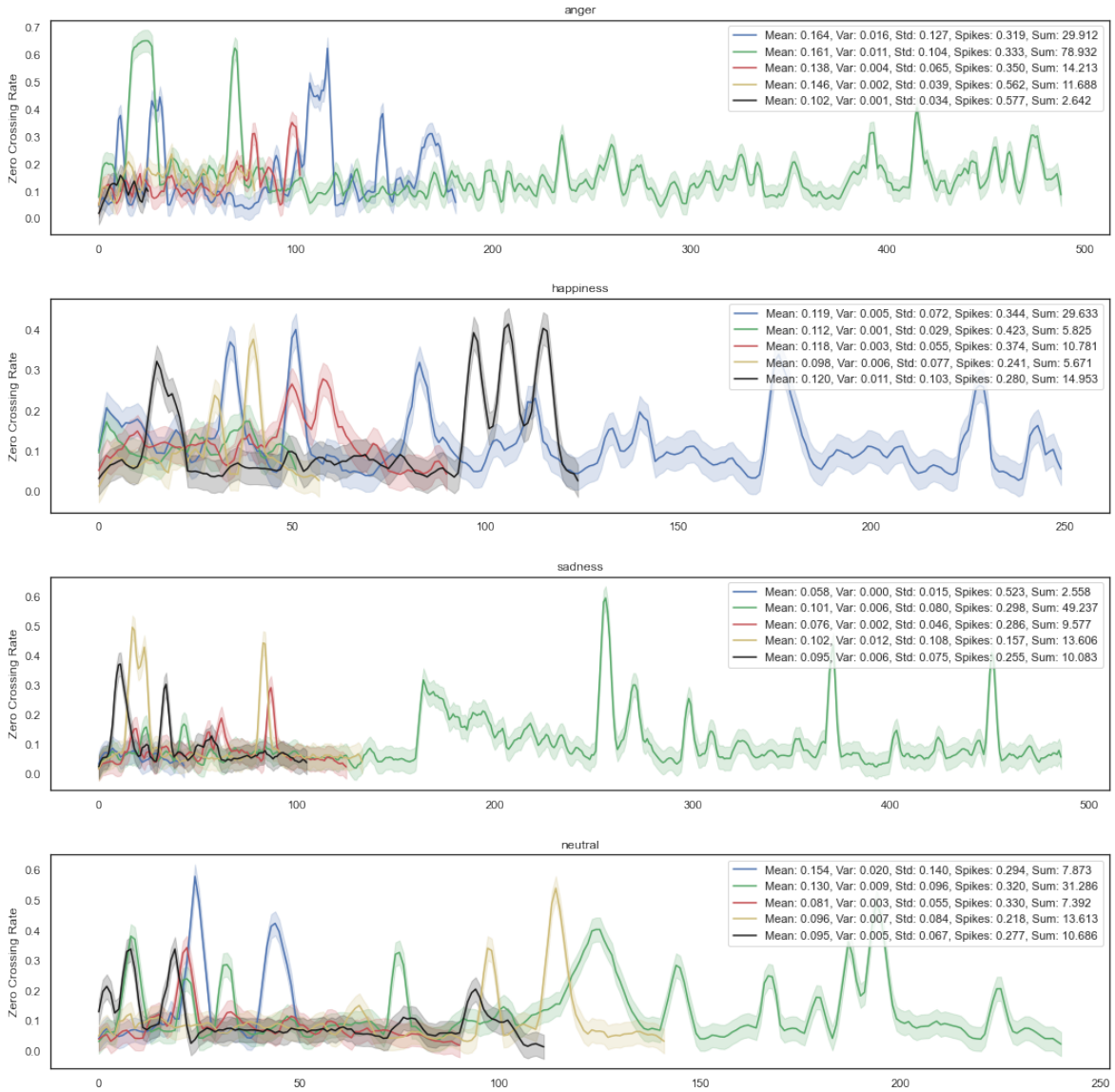


Figure .9: Zero crossing rate wave plots with a surrounding area of five male subjects and the same sentence for all emotions.

.1.4 Variation Plots

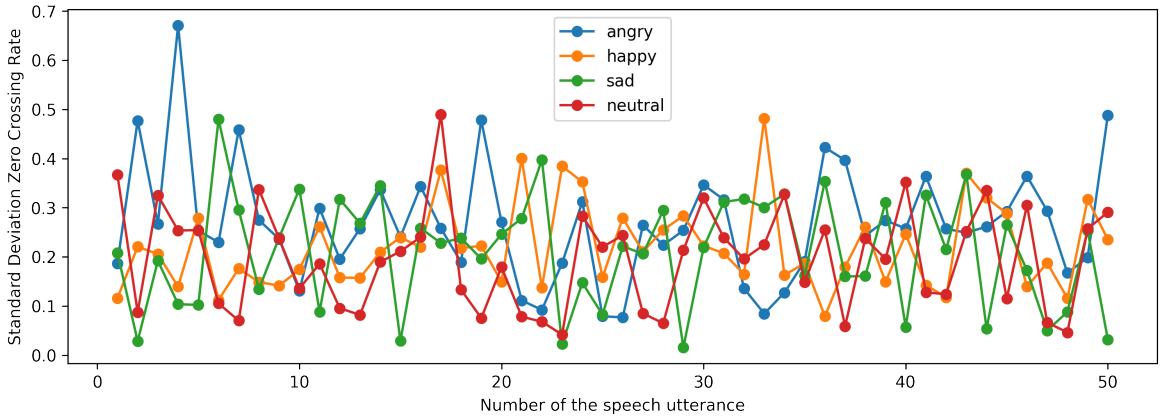


Figure .10: Zero crossing rate standard deviation values variation plot along 50 audios of speech utterances for all emotions.

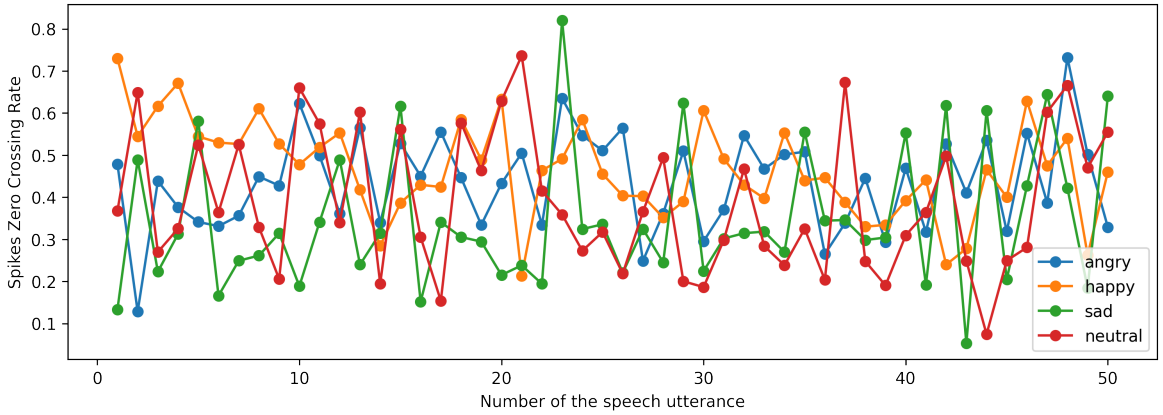


Figure .11: Zero crossing rate spikes values variation plot along 50 audios of speech utterances for all emotions.

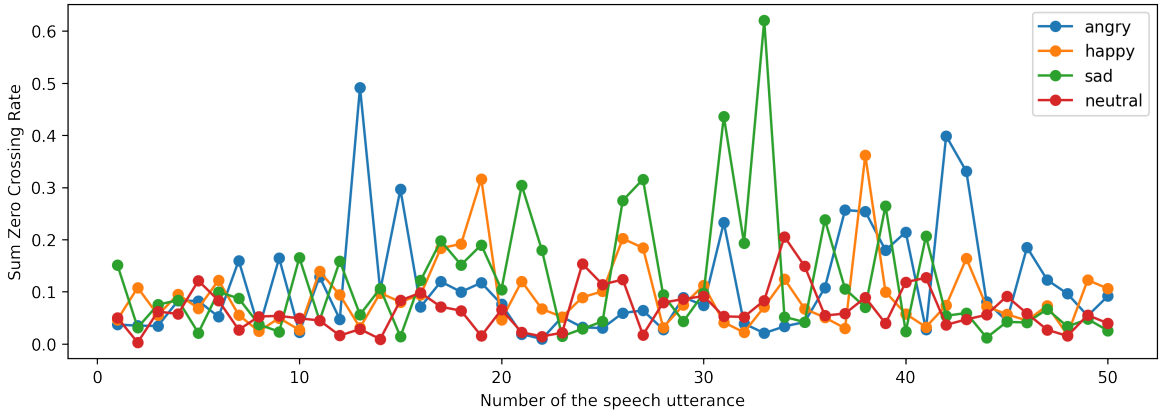
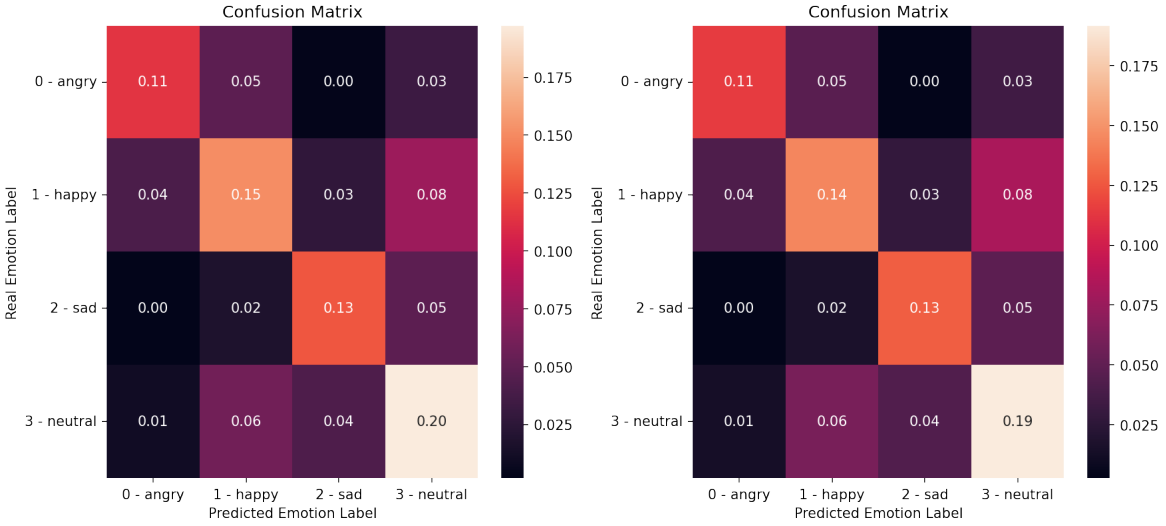


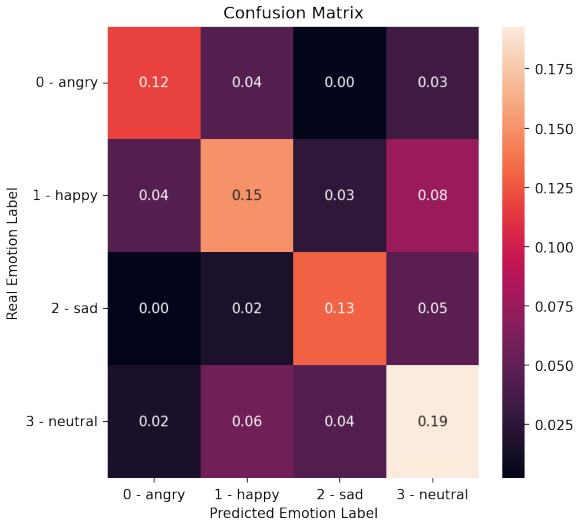
Figure .12: Zero crossing rate sum values variation plot along 50 audios of speech utterances for all emotions.

.1.5 Confusion Matrices



(a) CM using all 327 features.

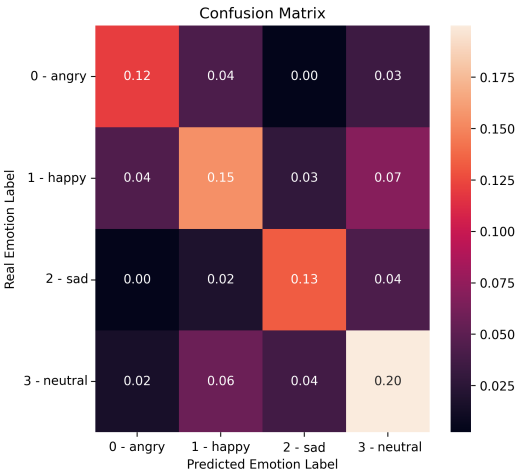
(b) CM using 98 features after CFS.



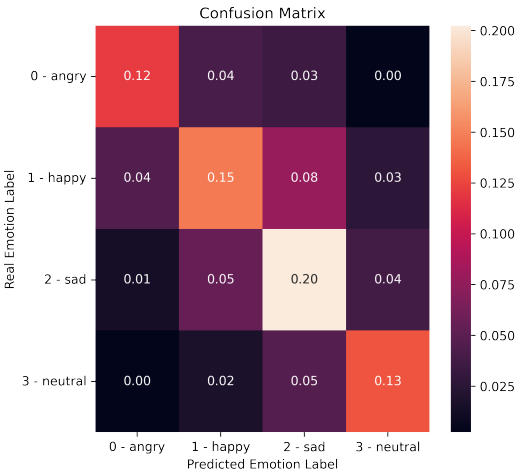
(c) CM using 33 features after CFS and backward selection.

Figure .13: RF 5-fold CV confusion matrices using different sets of features.

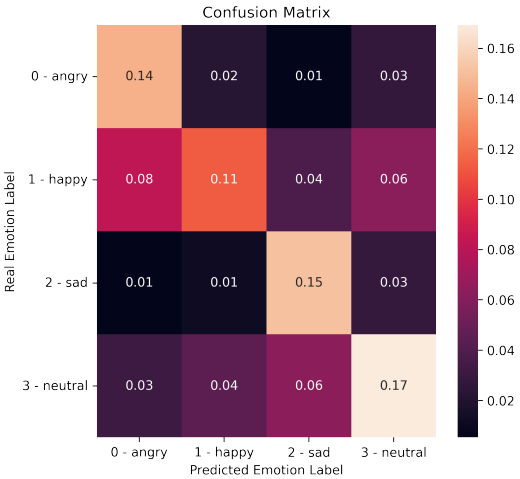
.1.6 Classifiers Evaluation and Selection



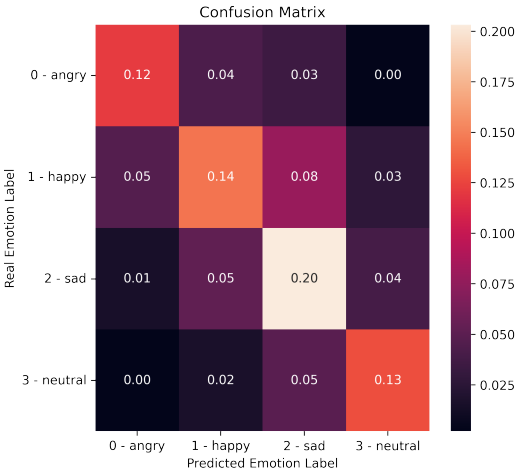
(a) XGBoost CM.



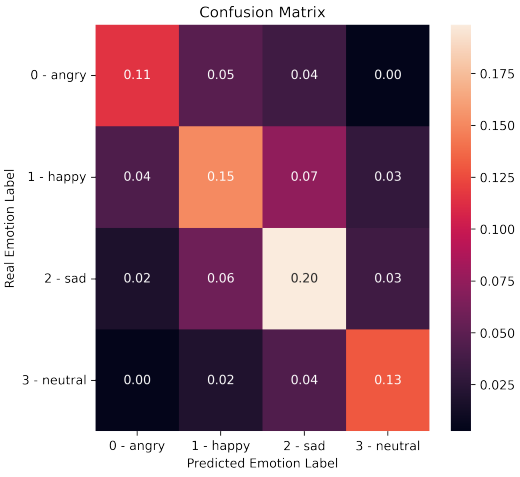
(b) AdaBoost CM.



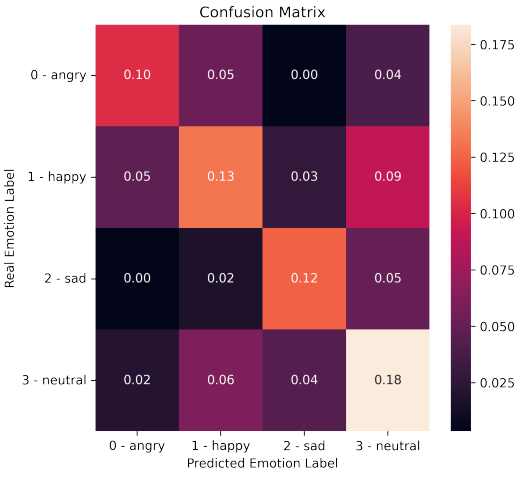
(c) Balanced RF CM.



(d) RF CM.

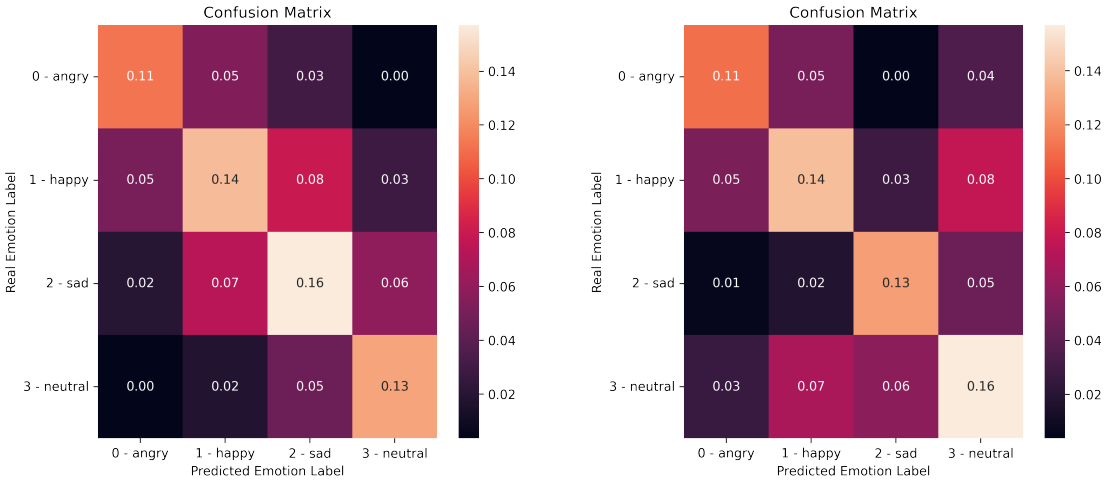


(e) Histogram Gradient Boosting CM.



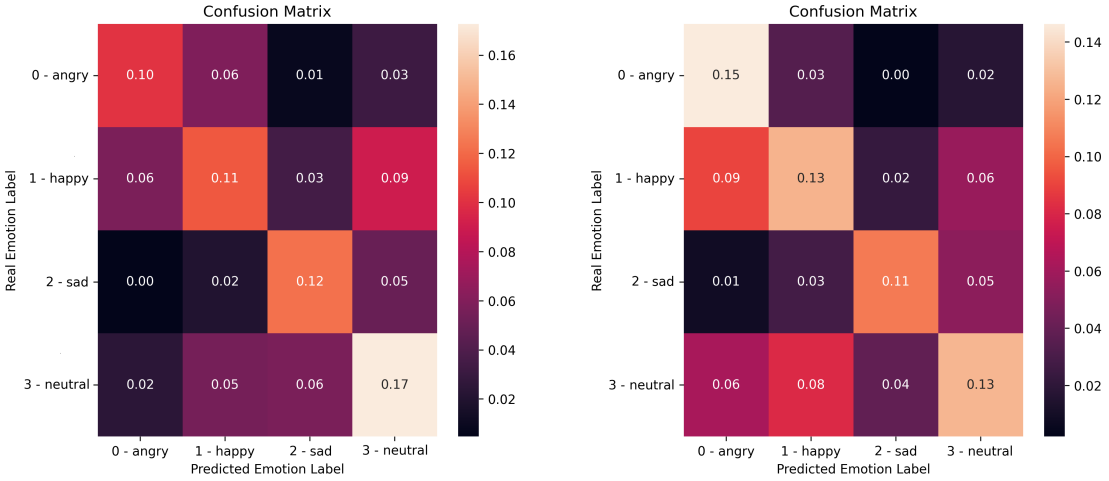
(f) SVM CM.

Figure .14: Tested models' 5-fold stratified CV confusion matrices on IEMOCAP (1).



(a) Linear Discriminant Analysis CM.

(b) Ridge CM.



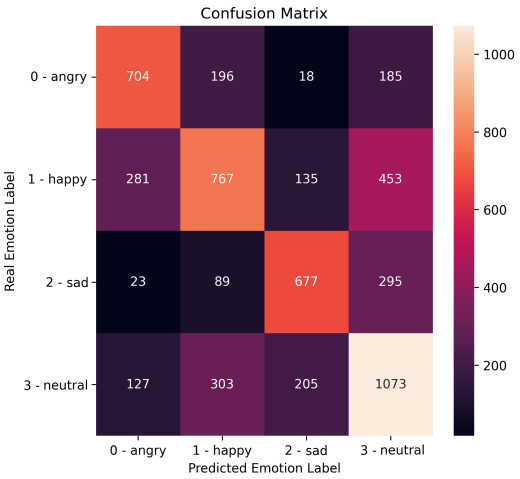
(c) LSTM CM.

(d) CNN CM.

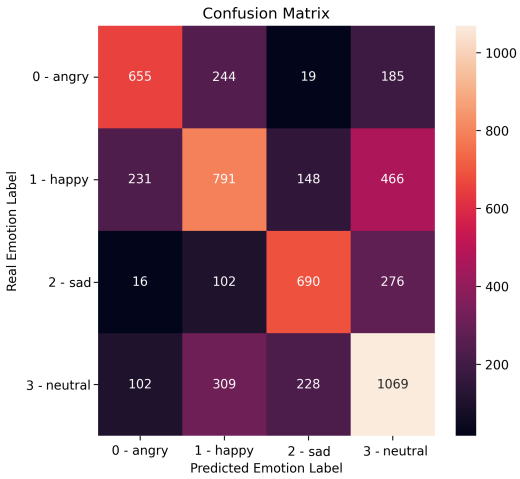
Figure .15: Tested models' 5-fold stratified CV confusion matrices on IEMOCAP (2).

.2 Deep Learning-Based SER

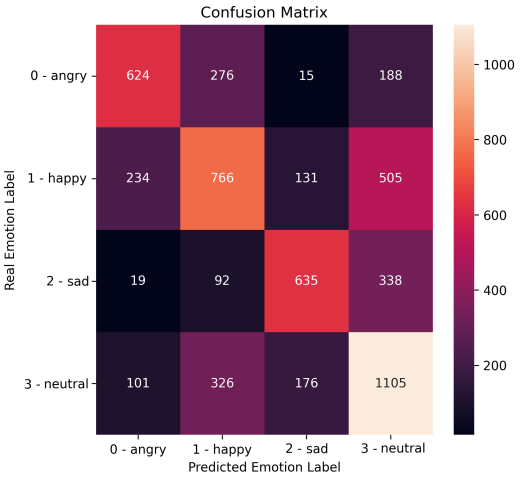
.2.1 Classifiers Evaluation and Selection



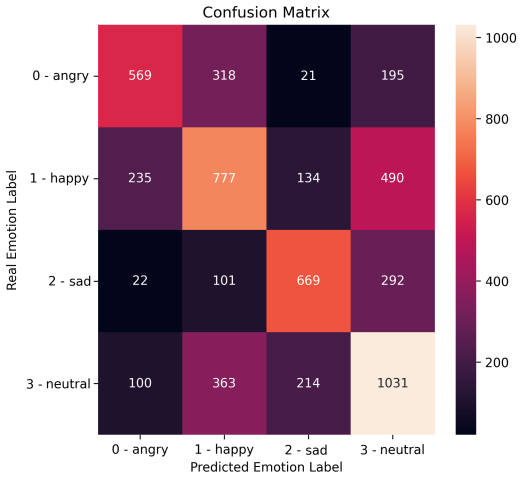
(a) Resnet50 using spectrogram images CM.



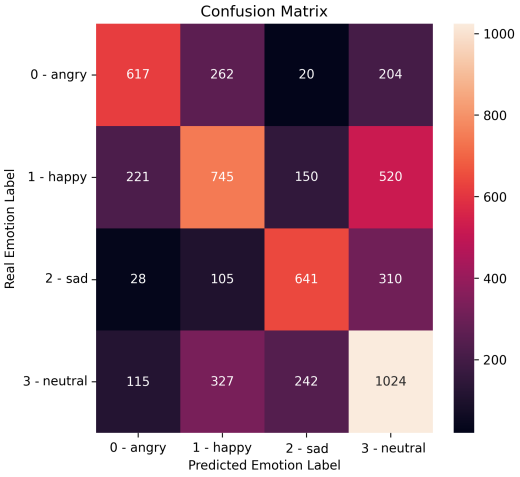
(b) Resnet50 using mel spectrogram images CM.



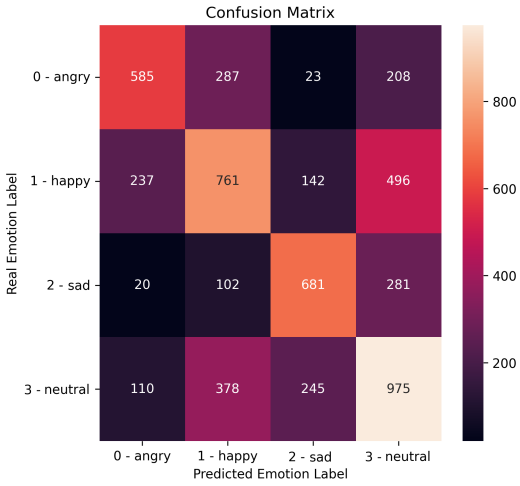
(c) Resnet50 using MFCC images CM.



(d) VGG16 using mel spectrogram images CM.

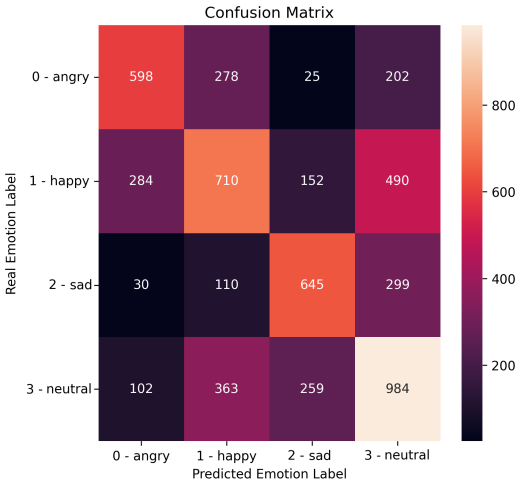


(e) VGG16 using MFCC images CM.

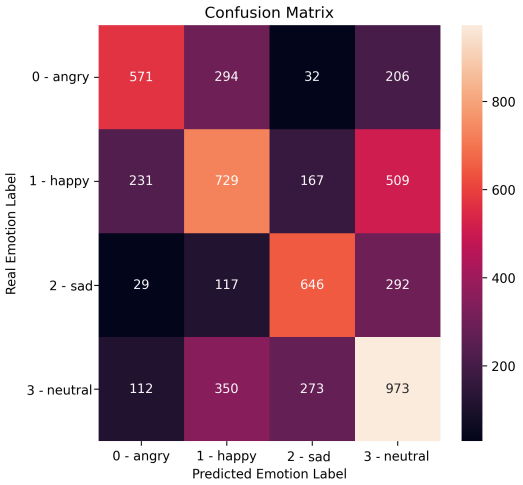


(f) VGG16 using spectrogram images CM.

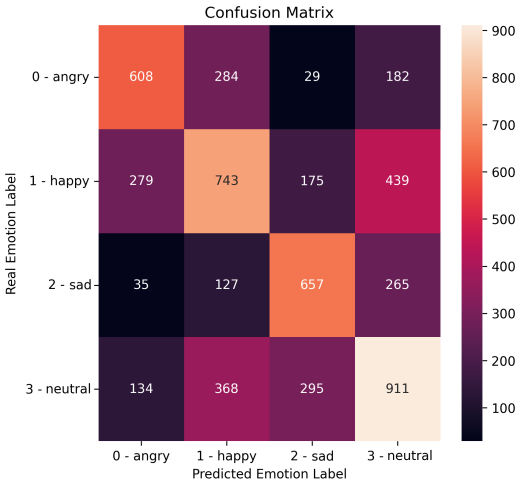
Figure .16: DL classification models confusion matrices on IEMOCAP (1).



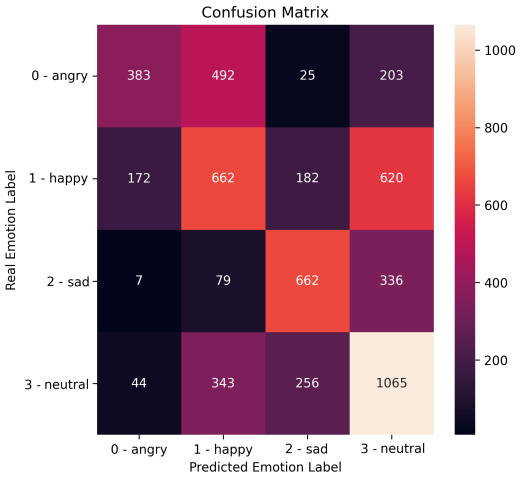
(a) Xception using mel spectrogram images CM.



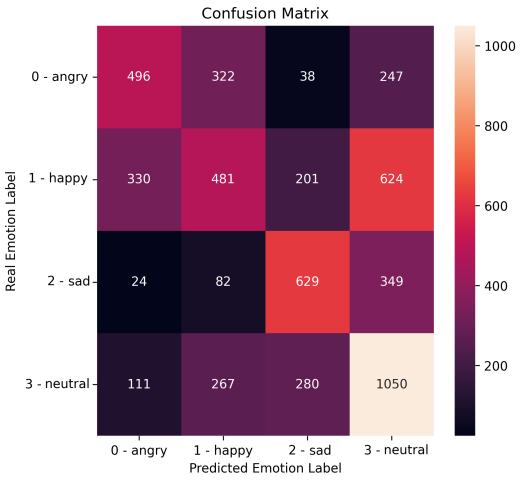
(b) Xception using MFCC images CM.



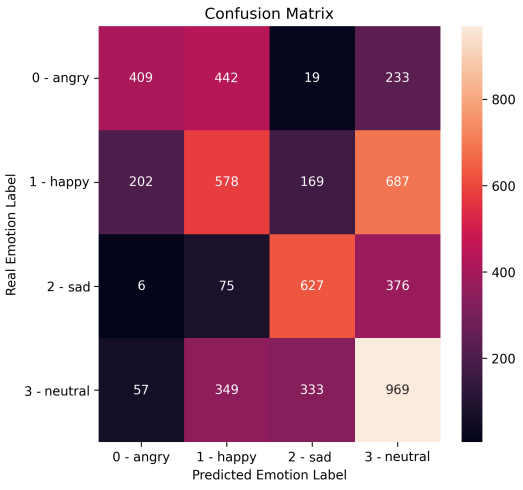
(c) Xception using spectrogram images CM.



(d) 2D-CNN using spectrograms CM.

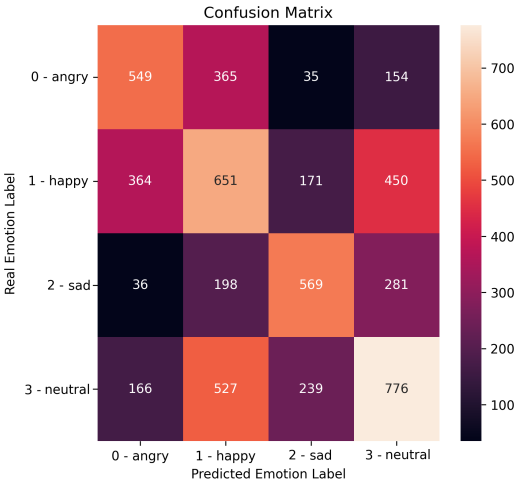


(e) 2D-CNN & RNN using mel spectrograms CM.

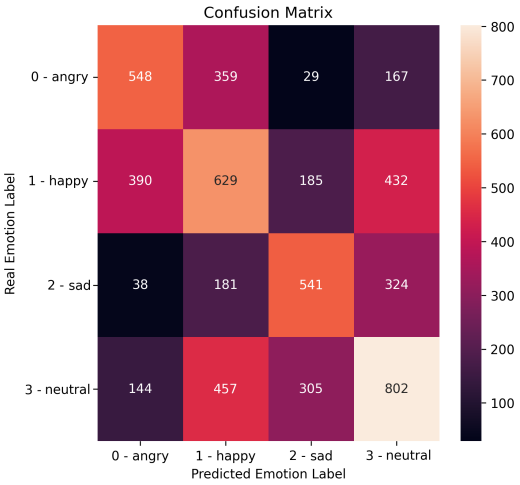


(f) 2D-CNN using MFCC CM.

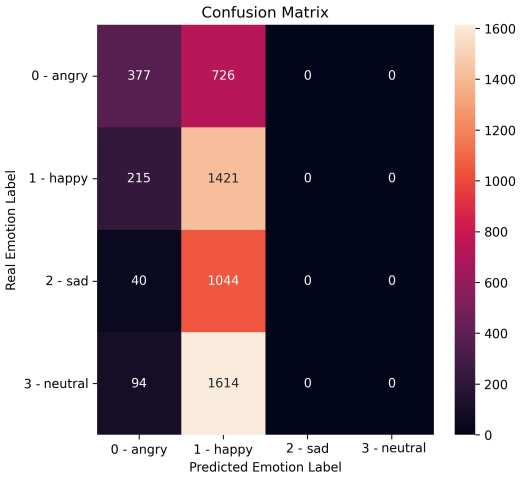
Figure .17: DL classification models confusion matrices on IEMOCAP (2).



(a) 2D-CNN & RNN using spectrograms CM.



(b) 2D-CNN & RNN using MFCC CM.



(c) 2D-CNN using mel spectrograms CM.

Figure .18: DL classification models confusion matrices on IEMOCAP (3).

.3 Data Stratification

.3.1 Dimensional Emotions

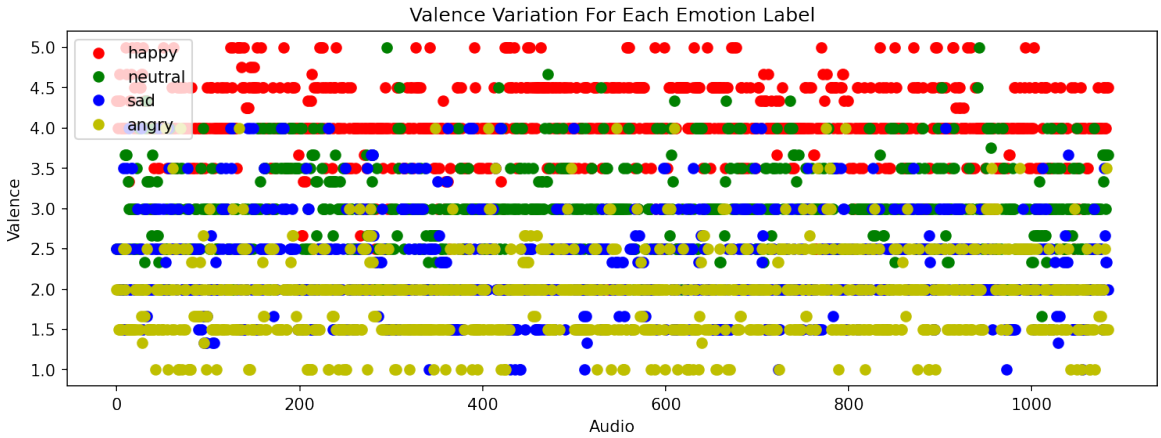


Figure .19: Scatter plot of the annotated emotions in the valence dimension.

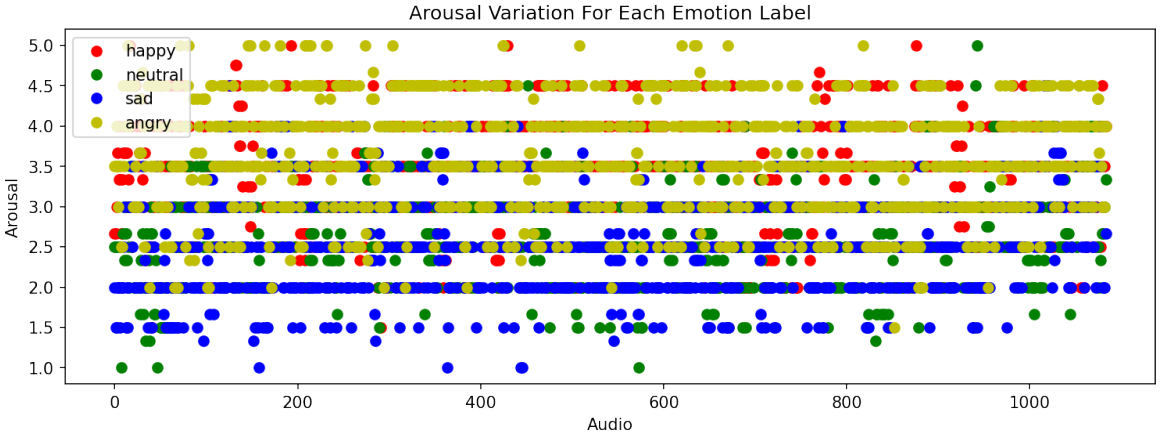


Figure .20: Scatter plot of the annotated emotions in the arousal dimension.

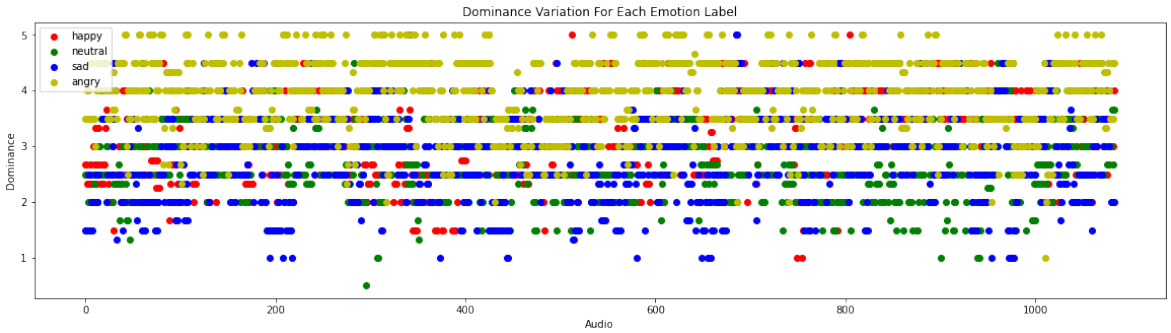


Figure .21: Scatter plot of the annotated emotions in the dominance dimension.

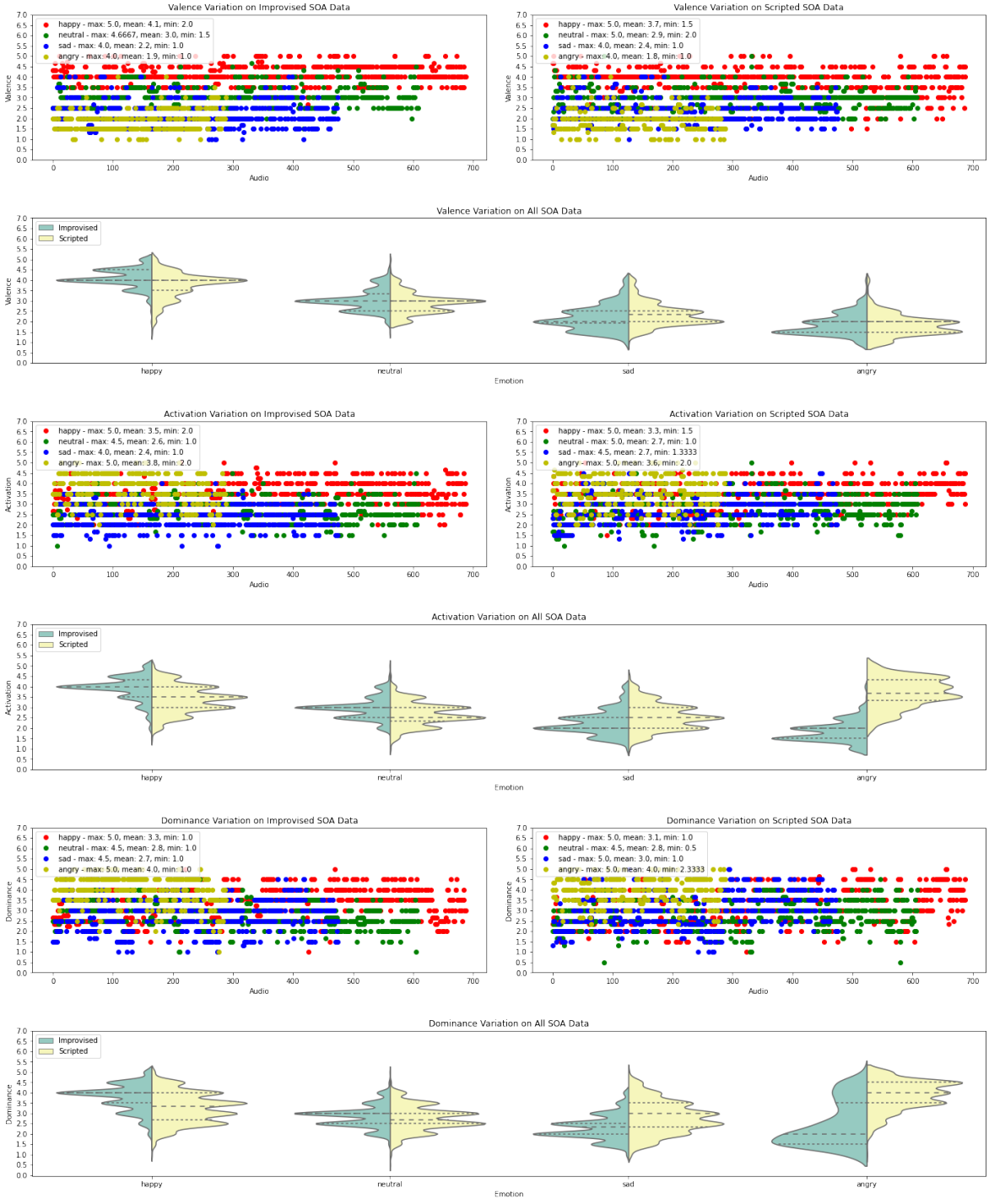


Figure .22: Scatter and violin plots of the emotional content of the IEMOCAP in terms of VAD relative to the all emotions.

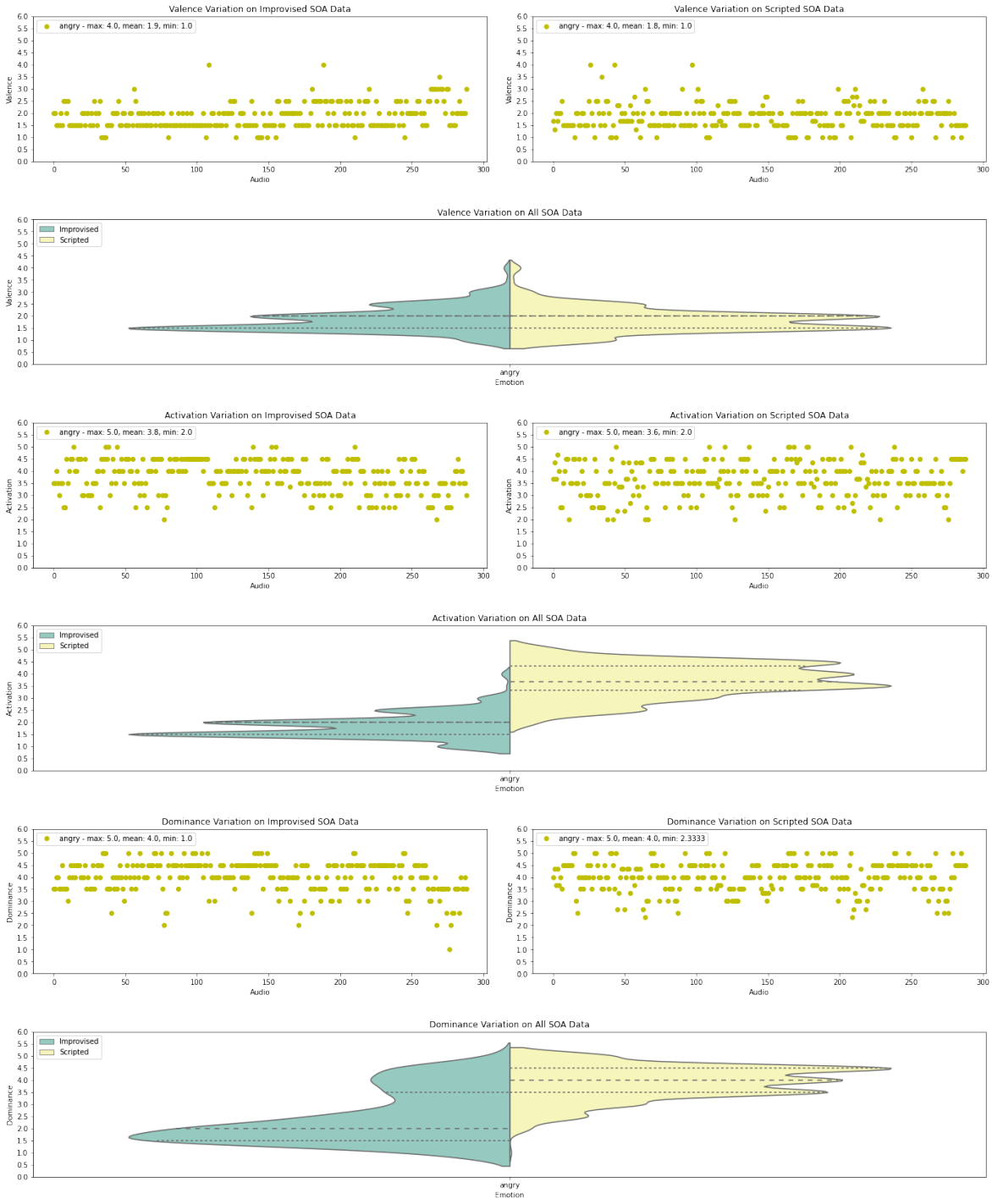


Figure .23: Scatter and violin plots of the emotional content of the IEMOCAP in terms of VAD relative to the angry emotion.

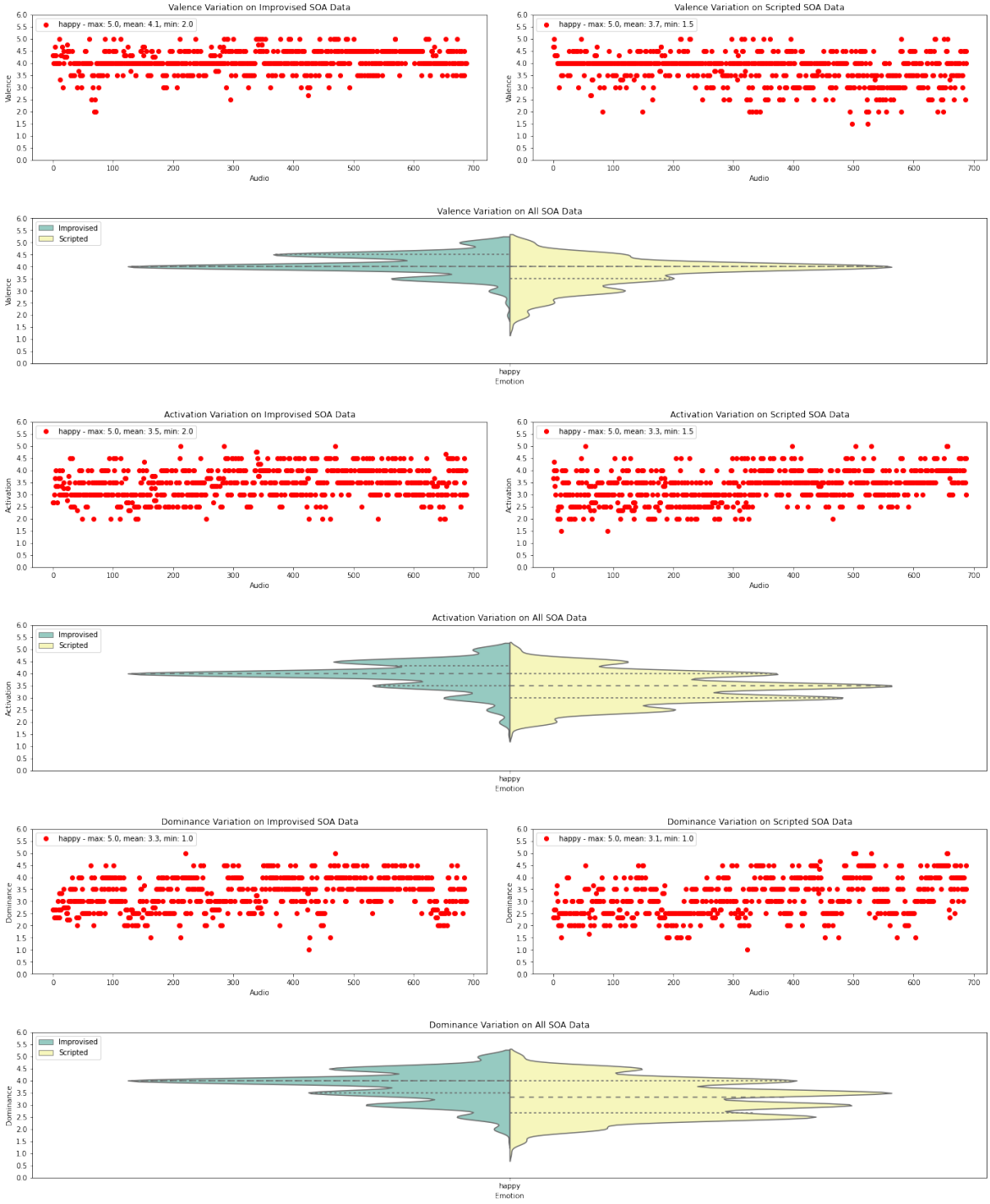


Figure .24: Scatter and violin plots of the emotional content of the IEMOCAP in terms of VAD relative to the happiness emotion.

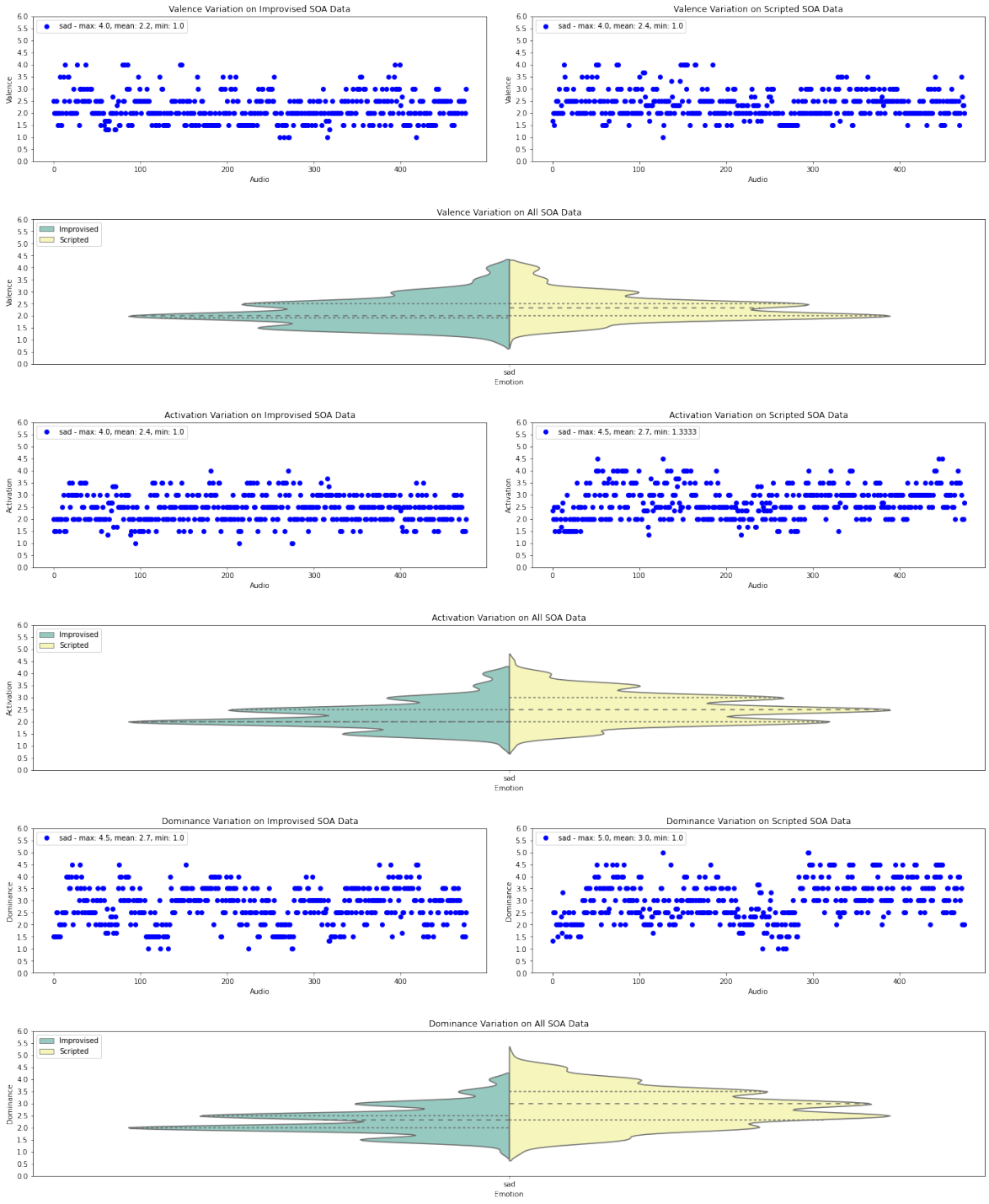


Figure .25: Scatter and violin plots of the emotional content of the IEMOCAP in terms of VAD relative to the sad emotion.

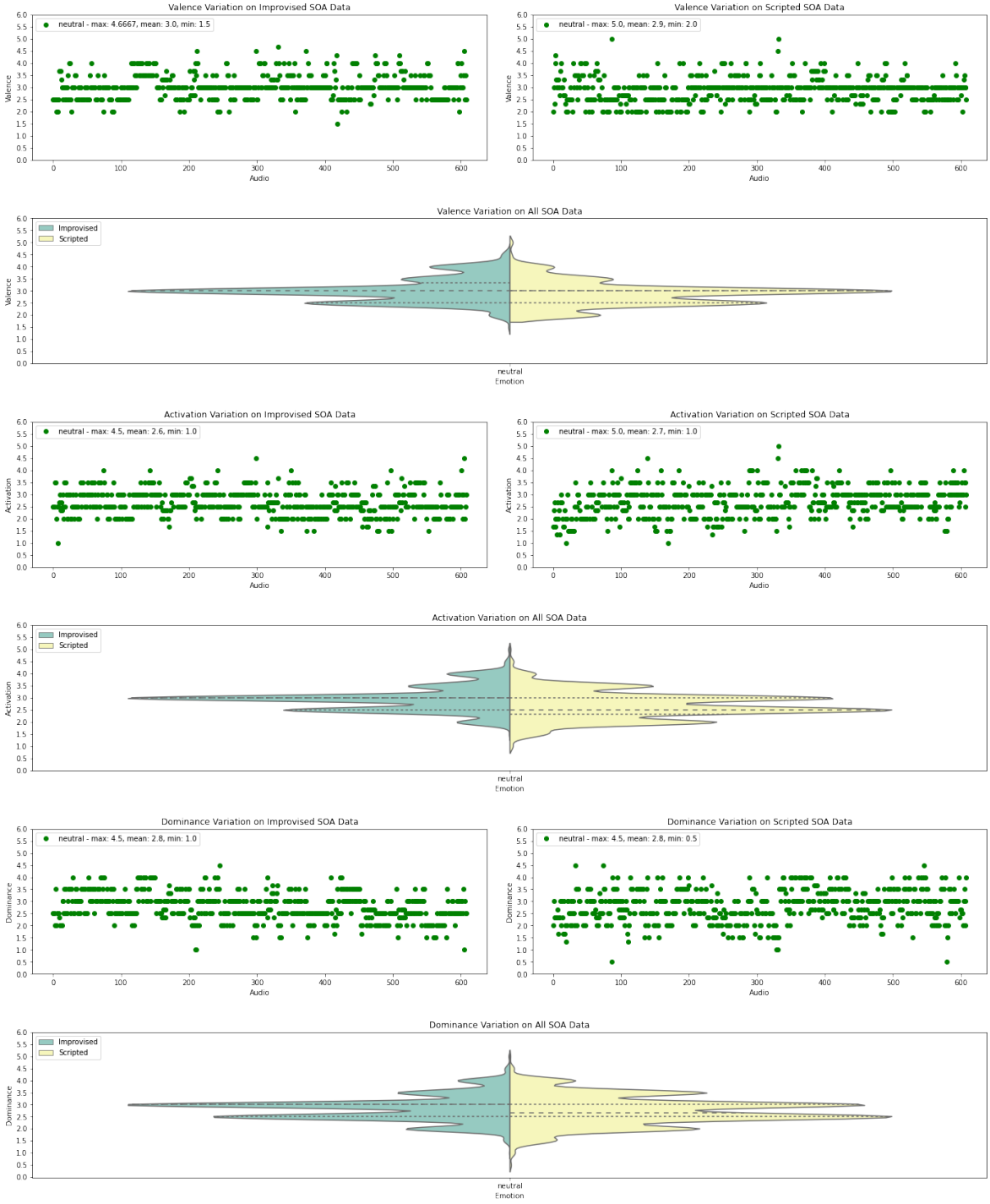


Figure .26: Scatter and violin plots of the emotional content of the IEMOCAP in terms of VAD relative to the neutral emotion.

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 Ballegooy, and R. Englert, «An emotion-aware voice portal», Jan. 2005.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[54] 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.

[55] 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.

[56] 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

[70] X. Yan, H. Xue, S. Jiang, and Z. Liu, «Multimodal sentiment analysis using multi-tensor fusion network with cross-modal modeling», *Applied Artificial Intelligence*, vol. 36, no. 1, Oct. 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*, 2022. DOI: 10.5281/ZENODO.6759664. [Online]. Available: https://zenodo.org/record/6759664.

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

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

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

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

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

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

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

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

[102] M. 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.

[103] C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., *Efficient and Robust Automated Machine Learning*, Curran Associates, Inc., 2015. [Online]. Available: https://papers.neurips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.

[104] H. Jin, F. Chollet, Q. Song, and X. Hu, «Autokeras: An automl library for deep learning», *Journal of Machine Learning Research*, vol. 24, no. 6, pp. 1–6, 2023. [Online]. Available: http://jmlr.org/papers/v24/20-1355.html.

[105] T. Chen and C. Guestrin, «XGBoost: A scalable tree boosting system», in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785.

[106] Mustaqeem and S. Kwon, «A CNN-assisted enhanced audio signal processing for speech emotion recognition», *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019. DOI: 10.3390/s20010183. [Online]. Available: https://doi.org/10.3390/s20010183.

[107] S. Liu, Y. Nakajima, L. Chen, *et al.*, «How pause duration influences impressions of english speech: Comparison between native and non-native speakers», *Frontiers in Psychology*, vol. 13, Feb. 2022. DOI: 10.3389/fpsyg.2022.778018. [Online]. Available: https://doi.org/10.3389/fpsyg.2022.778018.

[108] M. Mário Silva, *Speech Emotion Recognition Classifiers and Pipeline*, version 0.1, Apr. 2023. [Online]. Available: https://github.com/VADER-PROJ/SER_Tools.