# COMPARING AUDIO FEATURES FOR SPEECH EMOTION RECOGNITION USING MACHINE LEARNING ALGORITHMS

FATMA GÜMÜŞ

MEF UNIVERSITY

JANUARY 2022

**MEF UNIVERSITY**
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING
GRADUATE PROGRAM IN INFORMATION TECHNOLOGIES

M.Sc. THESIS

# COMPARING AUDIO FEATURES FOR SPEECH EMOTION RECOGNITION USING MACHINE LEARNING ALGORITHMS

FATMA GÜMÜŞ
ORCID ID: 0000-0003-2014-4717

TUNA ÇAKAR

JANUARY 2022

**MEF UNIVERSITY**

This is to certify that we have read the thesis and it has been judged to be successful, in scope and in quality, at the defense exam and accepted by our jury as a Master's Degree in Information Technologies.

Jury Members:

| | VERDICT | SIGNATURE |
|---|---|---|
| Asst. Prof. Dr. Tuna ÇAKAR (Thesis Advisor) | _____ | _____ |
| Prof. Dr. Adem KARAHOCA | _____ | _____ |
| Asst. Prof. Dr. Hüseyin ÖZKAN | _____ | _____ |

This study has been approved in partial fulfillment of the requirements for the Master's Degree in Information Technologies by the MEF Universtiy Graduate School of Science and Engineering.

Date: Seal/Signature:

## ACADEMIC HONESTY PLEDGE

I declare that all the information in this study, is collected and presented in accordance with academic rules and ethical principles, and that all information and documents that are not original in the study are referenced in accordance with the citation standards, within the framework required by the rules and principles.

Name and Surname: Fatma GÜMÜŞ
Signature:

## ABSTRACT

COMPARING AUDIO FEATURES FOR SPEECH EMOTION RECOGNITION USING MACHINE LEARNING ALGORITHMS

Fatma Gümüş

M.Sc. in Department of Information Technologies

Thesis Advisor: Asst. Prof. Dr. Tuna Çakar

January 2022, 106 Pages

Voice is an integral part of our lives. The demand for voice technology in both art and human-machine interaction systems has recently been increased. More information can be transferred quickly by voice. Speech is a natural way of communicating and as a result of this, it is primarily preferred for contacting users in technological areas. Our voice conveys both linguistic and paralinguistic messages in the course of speaking. The paralinguistic part, for example, rhythm and pitch, provides emotional cues to the speaker. Emotions consist of cognitive, physiological and behavioural changes and all these phenomena are interrelated. Generally, an emotion is a state that affects the thoughts and is capable of determining behaviour. Emotion also creates physical and psychological changes. Speech Emotion Recognition topic examines the question 'How is it said?' and an algorithm detects the emotional state of the speaker from an audio record.

Within the scope of this study, machine learning models are developed with classification methods to resolve the problem of speech emotion recognition. Voice consists of a lot of characteristics. However, the optimal audio feature set related to the emotional state cannot be determined yet. The main aim in this study is obtaining the most distinctive emotional features. For this purpose, in order to compare audio features based on different domains Root Mean Square Energy (RMSE), Zero Crossing Rate (ZCR), Chroma and Mel Frequency Cepstral Coefficients (MFCC) features are examined for emotion recognition. A pre-trained model namely wav2vec

Large which has been developed more recently is used to create the inputs also. Support Vector Machine, Multi-Layer Perceptron and Convolutional Neural Network techniques are utilized for developing learning models for comparing traditional features and the pre-trained model representations. In this paper emotions namely, Happy, Calm, Angry, Boredom, Disgust, Fear, Neutral, Sad and Surprise are classified, and furthermore, the models are trained and tested with English and German speech datasets.

When the classification results are examined, it is concluded that the most successful predictions are obtained with the pre-trained representations. The weighted accuracy ratio is 91% for both Convolutional Neural Network and Multilayer Perceptrons models while this ratio is 87% for the Support Vector Machine models. Among the emotional states, Fear has the highest recognition ratio with 95% f-score with Convolutional Neural Network technique which uses a pre-trained model.

Keywords: Emotion recognition; audio features; classifiers

Numeric Code of the Field: 92404

# ÖZET

Konuşmadan Duygu Çıkarımı için Makine Öğrenimi Algoritmaları Kullanılarak Ses Özelliklerinin Karşılaştırılması

Fatma Gümüş

Bilişim Teknolojileri Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Tuna Çakar

Ocak 2022, 106 sayfa

Ses, hayatımızın tamamlayıcı bir parçasıdır. Son yıllarda ses teknolojisine olan talep sanat ve insan-makine etkileşimi sistemlerinde artmaktadır. Ses ile daha fazla bilgi daha hızlı bir şekilde aktarılabilmektedir. Konuşma iletişim kurmanın doğal bir yoludur ve bunun sonucu olarak teknolojik alanlarda kullanıcı ile temas kurmada öncelikli tercih edilir. Sesimiz konuşma sırasında hem dilsel hem de dilsel olmayan bilgileri taşır. Dilsel olmayan ritim, perde gibi bilgiler konuşmacının duygu durumu ile ilgili ipucu sağlar. Duygular bilişsel, fizyolojik ve davranışsal değişikliklerden oluşur ve tüm bu fenomenler birbirleriyle ilişkilidir. Genel anlamda duygu, düşünceleri etkileyen, davranışları belirleyebilen, fiziksel ve psikolojik değişiklikleri oluşturan durum olarak açıklanabilir. Konuşmadan Duygu Çıkarımı konusu 'Nasıl söyledi?' sorusunun cevabını inceler ve kayıt edilmiş bir sesten algoritma yardımı ile duyguyu belirlemeye çalışır.

Bu çalışmada, Konuşmadan Duygu Çıkarımı problemine makine öğrenimi türlerinden sınıflandırma yöntemi ile çözüm aranmıştır. Ses çok fazla sayıda karakteristikten oluşmaktadır, bu karakteristiklerin duygu ile ilişkili olan optimize seti henüz tespit edilememiştir. Bu özellikleri karşılaştırmak ve en ayırt edici özelliği belirleyebilmek için sesin farklı boyutlardaki Root Mean Square Energy (RMSE), Zero Crossing Rate (ZCR), Chroma ve Mel Frequency Cepstral Coefficients (MFCC) özellikleri duygu tahmini için incelenmiştir. Daha yakın zamanlarda geliştirilmeye başlanan ön eğitimli model ile girdilerin oluşturulabilmesi için wav2vec Large modeli de kullanılmıştır. Geleneksel yöntemler ile elde edilen öznitelik ve ön eğitimli model girdilerinin duygu tahmini karşılaştırması için Destek Vektör Makineleri, Çok Katmanlı Algılayıcılar ve Evrişimsel Sinir Ağı algoritmaları ile modeller geliştirilmiştir. Çalışmada Mutlu, Sakin, Kızgın, Can Sıkıntısı, İğrenme, Korku, Nötr, Üzüntü ve Şaşkın duyguları sınıflandırılmaya çalışılmış, İngilizce ve Almanca konuşma setleri kullanılarak modeller eğitilmiş ve test edilmiştir.

Sınıflandırma sonuçları incelendiğinde en başarılı tahminlerin ön eğitimli modeller ile elde edildiği görülmektedir. Evrişimsel Sinir Ağı ve Çok Katmanlı Algılayıcılar için %91 ağırlıklı doğruluk oranı ön eğitimli modeller için ortak iken bu oran Destek Vektör Makineleri'nde %87'dir. Duygular arasında ise en iyi tahmin ön eğitimli model kullanılan Evrişimsel Sinir Ağı yöntemiyle Korku duygusu için %95 f-skor ile elde edilmiştir.

Anahtar Kelimeler: Duygu tanıma; ses özellikleri; sınıflandırıcılar

Bilim Dalı Sayısal Kodu: 92404

# TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## ABBREVIATIONS

**AESDD** **:** Acted Emotional Speech Dynamic Database

**CNN** : Convolutional Neural Network

**EMODB** : Berlin Database of Emotional Speech

**HMM** : Hidden Markov Model

**IEMOCAP** : The Interactive Emotional Dyadic Motion Capture

**LFPC** : Log frequency power coefficients

**LSTM** : Long Short-Term Memory

**MFCC** : Mel-Frequency Cepstral Coefficients

**MLPC** : Multi-Layer Perceptron Classifier

**RAVDESS** **:** The Ryerson Audio Visual Database Emotional Speech Song

**ReLU** **:** Rectified Linear Unit

**RMSE** : Root Mean Squared Energy

**RNN** : Recurrent Neural Networks

**SER** : Speech Emotion Recognition

**SHLA** : Shared Hidden Layer Autoencoder

**SVM** : Support Vector Machine

**STFT** **:** Short Time Fourier Transform

**TESS** **:** Toronto Emotional Speech Set

**ZCR** : Zero Crossing Rate

## INTRODUCTION

Voice is an integral part of our world and has a significant impact on us. In recent years, the usage of voice technology in art and Human Machine Interaction systems has been more widespread and popular. Intelligent personal assistants, voice commerces, encryption for security processes are some sample application areas. A great number of users utilize and experience them. Fundamentally, they save time and make daily life easier.

Speech related domain in computer science is named as Speech Processing and involves Speech Recognition, Speech Emotion Recognition and Speaker Identification topics. Speech Recognition deals with “What is said?”, Speech Emotion Recognition deals with “How it is said?” lastly, Speaker Identification deals with “Who says it?” (Swain et al., 2018)

Speech is one of the basic, efficient and powerful communication methods. At the beginning of the 20th century, electroacoustic analysis was used for determining emotions in psychology (Scherer, 2003). In academics, Speech Emotion Recognition has become one of the most wondered and investigated research areas (Jain et al., 2020). This research program aims to determine the emotional state of the speaker based on speech signals. Significant studies have been undertaken during the last two decades to identify emotions from speech by using machine learning. However, it is still a challenging task because emotions change from one to another and there are environmental factors which have significant effects on emotions. Furthermore, sound consists of numerous parameters and there are various anatomical characteristics to take into consideration. Determining an appropriate audio feature set for emotion recognition is still a critical decision point for an emotion recognition system.

Machine Learning is designed to emulate human intelligence and utilizes statistics or neural networks to make decisions without specific programs. To produce valuable results, it should have a learning ability. For building a machine learning model, example datasets and parameters are given generally. There are an enormous

variety of statistical techniques and neural network approaches for creating models.

Quite a lot of approaches for SER have been applied successfully till today. In this paper, both hand-crafted features based on time and frequency domains and pre-trained model inputs are used for determining the state-of-the-art audio features for emotion recognition. For superior results three speech databases are combined and three machine learning approaches are performed. Below, there is a schema which shows components of the study:

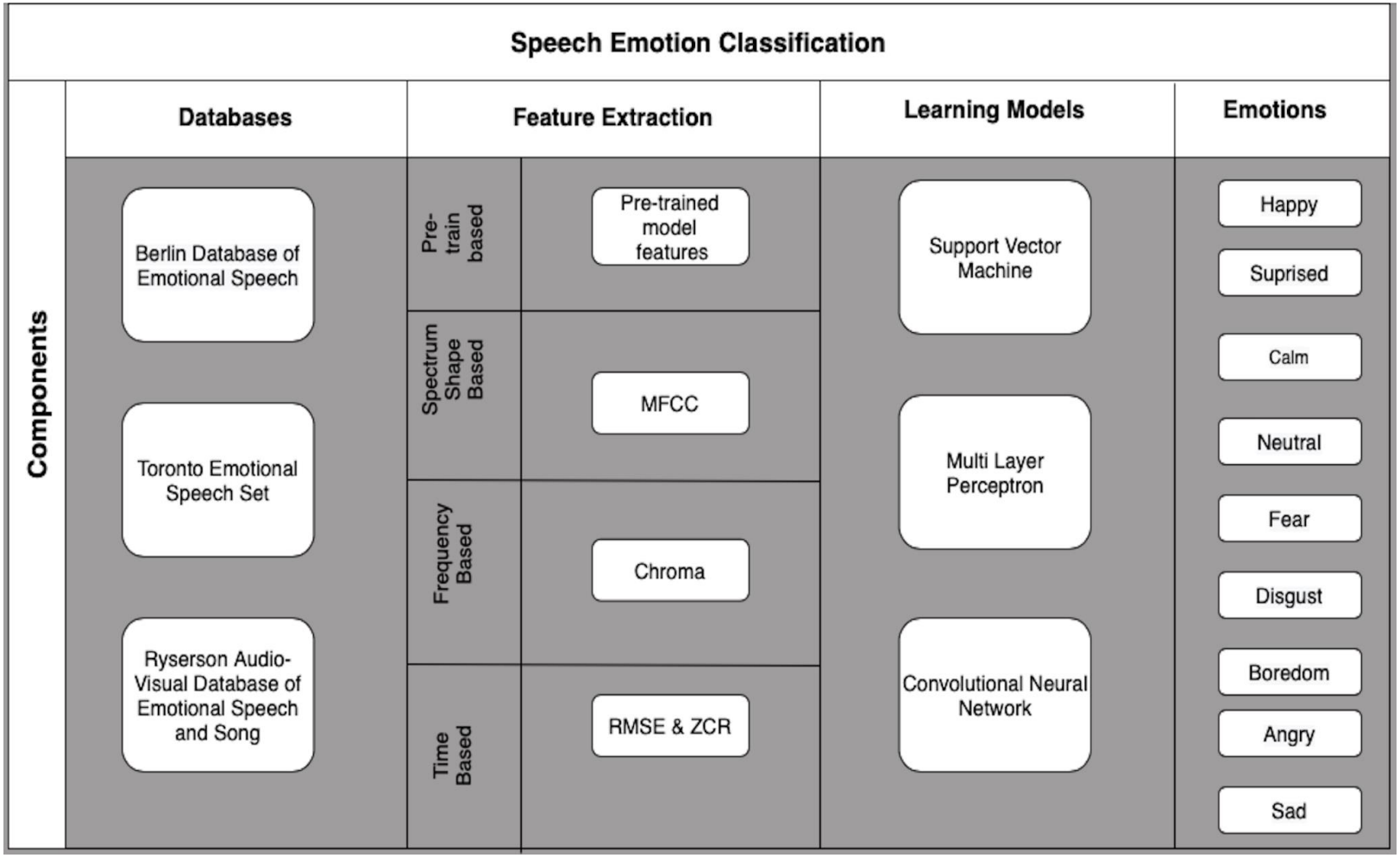



Figure 0.1: Components of the research

## 1. Purpose of Thesis

This paper aims to discover the most discriminative features of audio for emotion recognition from speech. To achieve such a competitive goal, unsupervised learning from raw audios and traditional feature extraction methods are systematically measured and compared. The most popular audio features namely, Mel-Frequency Cepstral Coefficients (MFCC), Chroma, Root Mean Squared Energy (RMS) and Zero Crossing Rate (ZCR) are evaluated for traditional approaches based on different domains of audio. On the other hand, for unsupervised learning a pre-trained model namely wav2vec Large is applied and representations of raw audios are employed as input samples. In the pre-trained model method, no specific feature is selected. In order to diversify and strengthen instances, three emotional speech datasets which are The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto Emotional Speech Set (TESS) and Berlin Database of Emotional Speech (EMODB) are combined. In this way, both speech corpus and emotion types are enriched. Besides, they acted by artists and actresses who are at different ages. To figure out the effect of input on models, both classical and deep learning algorithms are developed. Support Vector Machine preferred from classical algorithms, Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN) preferred as deep learning algorithms as all of them are extensively used.

## 2. Literature Review

Cowie et al. (2001) indicate that emotion has both a broad and narrow sense. In the narrow sense, emotion has a major impact on mental life, which is also called full-blown emotion. On the other hand, in the broad sense, emotion has a smaller impact, which is also called underlying. They studied emotion in a broad sense and mentioned that there is not a set of basic emotions agreed upon, emotion selection requires pragmatic choices and should take various cultural differences into account. According to them, the rate of the vocal cord's vibration determines the fundamental frequency of the acoustic signal. Speech carries both paralinguistic information such as pitch, intensity and linguistic information which includes rules of language and they

are linked. The magnitude of speech attributes increases or decreases from one emotion to another. Moreover, extracting speech attributes from acoustic signals is one of the major tasks of speech emotion recognition. Furthermore, paralinguistic features have a significant effect on recognizing emotion.

Gunes & Pantic (2010) mention in their study that there are three approaches for emotion: The first one is the categorical approach and to this approach there are basic emotions which are recognized universally. In addition to basic emotions, there are discrete emotions affecting mental states which are more determinable in daily life. The second one is the dimensional approach which has valence, arousal, and potency as the three main items for affect variability. Valence is about positive or negative emotion according to happiness; arousal is about excited or apathetic emotions; power is about the degree of control over emotion. According to the magnitude of the emotion, emotional labels may be positioned on a two-dimensional plane in which the x-axis shows valence and the y-axis shows arousal. The last approach is the appraisal-based approach in which emotions have a relation with both our own state and the outside world. This approach does not restrict the emotions neither with numbers nor dimensions, but there are different emotional states for various evaluations. In the study, the dimensional approach is examined for classifying emotions by using facial - bodily expressions and audio signals. Although varied solutions have been produced, there is no agreement on how dimensional emotion can be modelled. Some of the critical factors for the dimensional approach include the affection of baseline problems, choice of observable and non-observable modalities, and determining the duration of the emotion.

Schmitt et al. (2016) used bag-of-audio-words (BoAW) in their study to recognize emotions in speech. BoAW is used for the Natural Language Processing area to eliminate useless words. In the study, a codebook is generated in the training partition by using a random sampling method. After generating the codebook, acoustic low-level descriptors (LLDs) are quantized. Then, a bag is created from frequencies. The classic logarithm that uses the bag-of-words technique is applied to compress the range of the frequencies. The Support Vector Machine (SVM) based regression is used to predict emotional arousal and valence dimensions. Delay, window and codebook

sizes, a number of assignments and SVR complexity parameters are handled. The RECOLA database is used for the research and only Mel-Frequency Cepstral Coefficients (MFCCs) feature is taken into consideration. The study shows that Gaussian and Polynomial kernels do not perform better than Linear kernels with BoAW. Furthermore, larger codebooks are more useful for the prediction of valence than arousal.

Neumann & Vu (2017) use Attentive Convolutional Neural Networks (ACNN) for emotion recognition. They studied comparing different feature types, signal length relation on the system, improvised and scripted speech performance and duration for making a prediction. The model they apply consists of two main parts: a CNN with one convolutional layer that learns the presentation of the audio signal and a max pooling layer. The other part of the model is the attention layer that computes the weighted sum of all the information extracted from different input parts. The attention vector and output from the pooling layer are forwarded to a Softmax layer. The IEMOCAP database is used for all experiments. Experiment results show that prosody features perform worse than cepstral features like logMel and MFCC. It is stated in the study that prosodic features are strongly speaker-dependent and these features' usage is debatable in speaker-independent emotion recognition. A reason for this situation could be due to the prosody feature set’s information level as it contains too little information. MFCC, logMel, and eGeMAPS features generally produce similar results.

Satt et al. (2017) studied on raw spectrograms. They calculate spectrograms from audios then apply deep learning to spectrograms directly without extracting features. In the first stage, each sentence in the database was split for 3 seconds. These new sentences are used for labelling throughout the system. Whole sentences are used in the testing phase. They tried limiting the prediction latency even though it was losing accuracy. For each sentence, a spectrogram is calculated and then normalized. Convolutional Networks and Recurrent (LSTM) Networks are used for classifying. As a pre-process, non-speech noise sounds are removed from log spectrograms based on harmonic filtering. Clean spectrograms test data has the highest recognition rate of

68.8% weighted accuracy. The emotions namely, Anger, Happiness, Neutral and Sadness are classified and IEMOCAP database is used in the study.

Lim et al. (2016) research for CNNs, RNNs and time distributed CNNs-based SER methods. CNNs-based SER method both achieves signal on a 2D signal representation by using Short Time Fourier Transform (STFT) and connects CNNs and RNNs without using any hand-crafted features. RNN uses sequential data information and shares the same parameters at each layer whereas CNN accepts all input signals different from each other. When classification accuracy with CNNs and sequential CNNs structure compared with CNNs and LSTM network structure, distributed CNNs structure shows better results. They explain this situation as the speech signal is sequentially related. EMODB database is used in the research.

Trigeorgis et al. (2016) propose an end-to-end recurrent model within their investigation that combines Convolutional Neural Networks (CNNs) with LSTM network signals. The aim of the study is to learn feature extraction and regression steps in a common trained model in order to predict emotion. The model uses a direct audio time signal to determine probable spontaneous emotion. First, the signal's background noise is removed and specific parts of this signal are enhanced. Then the temporal structure is established by using a recurrent network with LSTM cells. In the research they discover that interpretable cells are highly correlated with both prosodic and acoustic features. Remote Collaborative and Affective Interactions (RECOLA) database is used in the study.

Nwe et al. (2003) studied a text-independent method for emotion recognition from speech. They explain that speech consists of a semantic part that involves linguistic information which is related to the pronunciation of the language as well as a paralinguistic part that involves implicit emotions which is the initial step for the emotional state. In the study, log frequency power coefficients (LFPC) are used for energy distribution across the frequency spectrum and the Hidden Markov Model (HMM) is used for classification. A text-independent but speaker-dependent emotion database is created for the research and six emotions take part in the database namely, Anger, Disgust, Fear, Joy, Sadness and Surprise. The speech signal instances are

segmented into frames, for each frame feature vectors are obtained and a codebook is generated using feature vectors. Energy distribution, overall intensity, speaking rate and the variation of tone are determined as parameters of speech emotion. Normalized short-time LFPC is used as the energy distribution of the signal. HMM models are built and the experiment results pointed out that a 4-state discrete ergodic HMM provides the best performance, due to more successful results for emotion recognition achieved by human assessment.

Hertel et al. (2016) compare time and frequency domain features of audio signals by using multiple deep neural networks and a convolutional network for recognition tasks. They tried to identify more distinctive features on the basis of the time and frequency domain. In the pre-processing phase, stereo files are converted into mono audio files, they are resampled, their amplitudes are scaled and a rectangular sliding window is applied. They use Freiburg-106 and ESC-10 datasets and f-score for measurement. According to their study, convolutional networks achieved superior results than the standard deep neural networks. In addition to this, the networks which are trained in the frequency domain achieve a higher f-score than networks that are trained in the time domain for both datasets.

## 3. Overview

The paper is organized as follows. Chapter 2 begins with background information about psychological views on emotion. Speech and Auditory systems are explained. Chapter 3 presents details about machine learning algorithms and audio features. Figures for audio features are presented. Chapter 4 describes the utilized emotional speech datasets. Implementation details and test results of made work are presented in detail. In chapter 5, developed work is discussed and in chapter 6 conclusion is made and some future work is stated.

# 1. BACKGROUND

## 1.1 Psychological View on Emotion

What is an emotion? There is not an exact and agreed answer for this question because emotion is a multidisciplinary concept and all about psychology, neurology, behavioural and diverse other disciplines. Emotions consist of cognitive, physiological and behavioural changes and all these phenomena are interrelated. While physiological responses may be congenital, behavioural and cognitive responses are personal and vary from culture to culture. The way in which emotions are expressed and experienced is subjective. For these reasons, it is difficult to explain emotions. There are various theories for explaining the relationship between event, physiological action, interpretation and emotion. James - Lange, Cannon-Bard, Schachter - Singer, Lazarus Theories are some main theories. On the other hand, in order to classify emotions, we need a systematic approach. The concept of emotion can be handled with two common models: Discrete Emotional Model and Dimensional Emotion Model.

The Discrete Emotional Model is based on basic emotions. Ekman explains ‘basic’ as distinguishing emotions from each other and the other meaning for 'basic' is that it points out emotions evolved for handling fundamental life tasks. He adds his own perspective for the main function of emotion: emotions make someone deal with important interpersonal encounters by using what has been adaptive both in the history of our species and our own individual past. He also thinks emotions provide information for conspecifics and prepare the organism for different emotional states if emotions were developed for fundamental life tasks. Distinct patterns occur on the Autonomic Nervous System for different emotions and they form motor behaviours that arrange different actions. Unique neural pathways in the Central Nervous System provide the changes in expression and cognitive activities such as memories and expectations. He mentions the emotion family concept for clarifying how many emotions exist. An emotion has got a family for affective state, the family comprises unique characteristic of theme which comes from evolution and variations that

represent individual learnings. There are six basic emotions namely, Sadness, Happiness, Fear, Anger, Disgust and Surprise (Ekman, 1999).

The Dimensional Emotion Model uses dimensions such as valence, arousal and power. One approach - known as the Circumplex Model of Affect - considers valence and arousal. The status of valence ranges from pleasantness to unpleasantness and arousal is about physiological activation. That being said, emotion consists of a combination of these two independent neurophysiological systems. A representation for the circumplex model of affect takes place in the following figure (Posner et al., 2005).

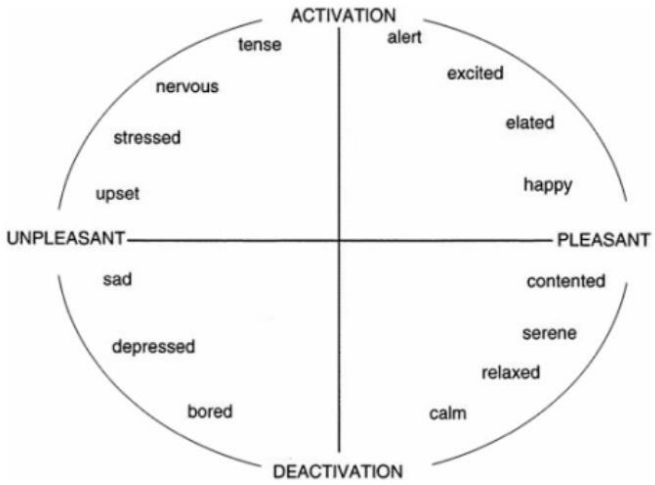



Figure 1.1.1: Circumplex model of affect

Emotional state changes over time and Cowie et al. (2001) compared the relation between time and emotional state as Figure 1.2 The figure includes categories for emotional states. Cowie et al. (2014) indicate that an emotional word can refer to more than one category. For instance, ‘happy’ can correspond to the short-term or long-term state.

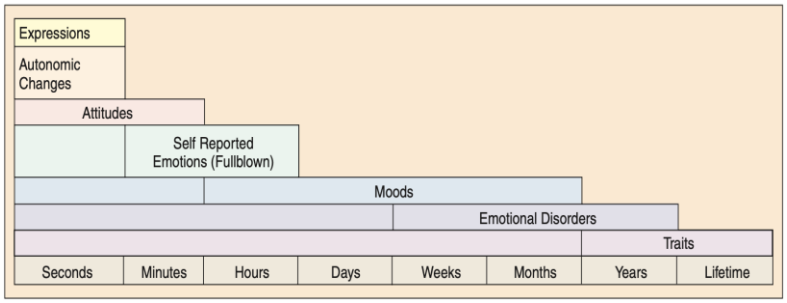



Figure 1.1.2: Emotion categories according to time

## 1.2 Speech and Auditory Systems

### 1.2.1 Speech system

Many parts of the human body are used for producing speech. They can be broadly divided into three mechanisms, the power of source (sound source), the vibrator (sound filter) and the resonator. Diaphragm takes place in the stomach and helps to push air from the lungs into the larynx. This airstream provides the source of energy for human speech. The airstream passes the vocal cords which are part of the larynx and makes them oscillate through the glottis. Vocal cord oscillation creates a sound and this sound is transformed to our unique personalized sound by a resonator system that includes throat, nose and mouth (Chen, 2016). The vocal folds are controlled by muscles and nerves in the larynx which determine the pitch. Vocal tract determines timbre and reinforces certain frequencies of the source, called formants (Latinus & Belin, 2011). Speech is an intricate process nevertheless the brain controls and manages the whole process momentarily. Below that, there is a figure shows speech system parts (Music 2º E.S.O).

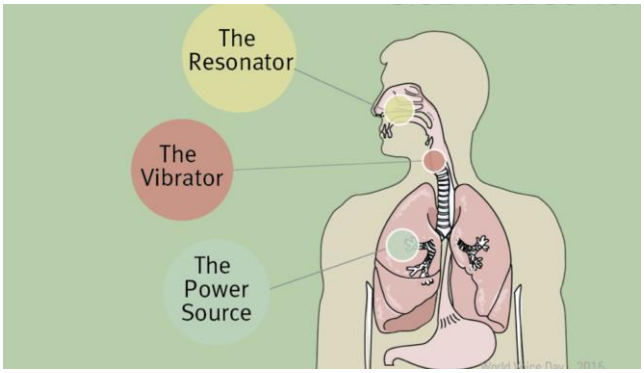



Figure 1.1.3: Speech system parts

### 1.2.2 Auditory system

The auditory system can be broadly divided into two mechanisms: the Peripheral and the Central Systems. The Peripheral System consists of three parts of the ear: outer, middle and inner ear and its function is collecting, filtering, amplifying and converting the sounds into electrical signals called neural impulses. These impulses are transferred to the Central System. The Central System is responsible for interpreting impulses and provides frequency and volume information about sound. Below that, there is a figure shows auditory system parts (Mariemont Hearing Center).

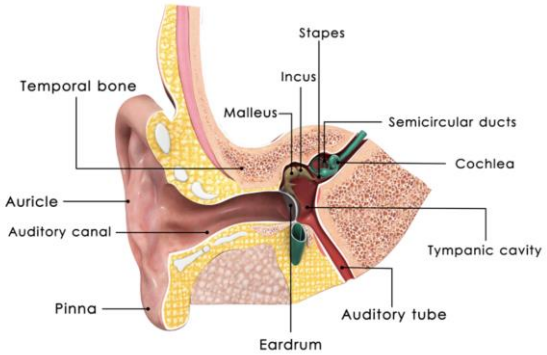



Figure 1.1.4: Auditory system parts

#### 1.2.2.1 Outer ear

The outer ear is composed of a pinna and auditory canal. Pinna is the visible part of the ear and the auditory canal is the tube that connects the pinna to the eardrum. The outer ear collects sound from the outside world and also reflects and absorbs sound waves, whereas changing waves are directed to the eardrum.

#### 1.2.2.2 Middle ear

The middle ear is composed of ossicles (malleus, incus and stapes), eardrum and Eustachian Tube. Sound waves create pressure and then cause the eardrum to vibrate. The vibrations are transmitted through ossicles to the fluid of the inner ear. The Eustachian tube balances the air pressure on both sides of the eardrum. As air is less dense than the fluid most of the sound wave energy would lose in the ear canal. The middle ear provides impedance transformation between the ear canal and cochlear fluid (Kurokawa & Goode, 1995).

#### 1.2.2.3 Inner ear

The inner ear is composed of the cochlea and semi-circular canals. The cochlea is responsible for hearing and it is divided into two chambers by a membrane (UTHealth McGovern Medical School). Different points of the basilar membrane resonate at different frequencies. Therefore, it can be accepted as a spectrum analyser. The frequency resonances of the basilar membrane show a log scale distribution. It means that the pitch depends on a logarithmic perception. Semi-circular canals are responsible for the balance of the body movement (Robinson, 1999). The organ of Corti which is placed in the cochlea contains the sound sensing cells called hair cells. The hair cells send electrical impulses to the brain via the auditory nerve and get signals from the brain (Jacobson, 1999).

# 2. MACHINE LEARNING BASED EMOTION RECOGNITION

Arthur Samuel, who was a computer scientist and pioneer of artificial intelligence research defined ‘Machine Learning’ as a computer's ability to learn without being explicitly programmed (1959). Machine learning methods can be characterized based on learning types, supervised and unsupervised learning.

## 2.1 Supervised Learning

Supervised learning utilizes labelled instances and designed to learn by examples. In this type of learning both the input and output variables are presented. The objective of the supervised learning method is to predict the correct label for an unlabelled data. A supervised learning algorithm can be shown simply as:

$$Y = f(x) \tag{2.1}$$

x: input value
Y: predicted output

Classification and Regression are subcategories of Supervised Learning. While classification methods are used for the categorical output, the regression methods are used for the continuous output.

### 2.1.1 Classification

Classification techniques categorize or classify data from the prior information. A model learns from the input data by weighting input features, then uses this learning to classify new instances. A dataset may consist of simply bi-class (for e.g., mail is spam or non-spam) or include multi-classes.

Kotsiantis (2007) studied classification problems in which the output of instances admits only discrete and unordered values. He pointed out that supervised classification is one of the most frequently applied tasks. Therefore, various techniques like Logic or Perceptron based techniques have been developed for Artificial Intelligence. On the other hand, Bayesian Networks, Instance based techniques have been developed for Statistics.

In this study, Perceptron based-techniques Convolutional Neural Network, Multi-Layer Perceptron and Statistical Based-Support Vector Machine algorithms are used for classification.

### 2.1.2 Regression

In a regression task, the goal of the model is to estimate and understand the important relationship between dependent and independent variables. It is a predictive statistical process and uses a continuous function to evaluate out how outputs change for given inputs. Most of the time, there are existing relationships between input and output variables. The most common regression algorithms are listed below:

1. Linear Regression
2. Non-Linear Regression
3. Logistic Regression
4. Polynomial Regression

Different regression studies have been made for understanding the relation among voice features.

### 2.2 Unsupervised Learning

In unsupervised learning, unlike supervised learning, labelled instances are not available. Algorithms aim to either discover similar example groups in the data or determine the distribution in the space of data by identifying hidden patterns in them. According to probability distribution of data, there are two classes for Unsupervised Learning, Parametric and Non-Parametric Unsupervised Learning. Clustering and Association are two popular methods for unsupervised learning problems.

### 2.3 Audio Features

Audio consists of nonstationary signals in nature. For analysing and examining them, a stationary state is needed. It is assumed that for a short time basis in time and frequency domains audio signals are stationary.

#### 2.3.1 Time (Temporal) domain features

Time domain features of a signal provide instantaneous information about audio and they are computed directly on the waveform. To interpret speech signals characteristics, short term and long-term levels are used. Global features describe and provide information about the entire signal, whereas local features provide information for only a segment of the signal. The short term (frame) level lasts 10 to 30 ms and the long-term level (clip, window) lasts longer than 0.5 seconds. Zero Crossing Rate, Amplitude Envelope, Energy of Signal, Minimum Energy and Root Mean Square Energy are time-based feature examples.

#### 2.3.2 Frequency (Spectral) domain features

Spectral domain features are about frequency components of the speech signal. They are obtained by converting the time domain to the frequency domain using the Fourier Transform or auto-regression analysis. The Fourier transform applies on a

continuous signal and gives the frequencies and magnitude of each frequency present within the signal. Chroma-related features and Tonality based features are frequency-based feature examples.

### 2.3.3 Spectrum shape domain features

Spectrum - shape related feature is a subcategory of frequency-based feature. Spectral Centroid, Spectral Contrast and MFCC are spectrum shape-based feature examples.

## 2.4 Feature Selection and Extraction

Classification is a type of pattern recognition problem. According to Wolf (1972), pattern recognition comprises measurement and classification phases. The output of the measurement phase contains several parameters that define the pattern. Features are the optimal combination of these parameters.

Wolf (1972) pointed out the following characteristics which an ideal feature should have for speech recognition. They are acceptable for speech emotion recognition also:

i. Occur naturally and frequently in normal speech

ii. Easily measurable

iii. Not be affected by reasonable background noise nor depend on specific transmission characteristics

Most research indicate that feature selection is the vital and prominent part of the emotion recognition system. Utilized features directly affect models and results. Although a considerable number of the studies have been conducted for selecting and extracting an optimal set of features, appropriate attributes for automatic emotion recognition from audio are still under research. Moreover, speech signals include traits that have no relation with emotional statements such as recording conditions, linguistic

content, gender and age. Speaker normalization (Schuller et al., 2010) and variability compensation via i-Vector modelling techniques are applied to eliminate these factors (Kua et al., 2014).

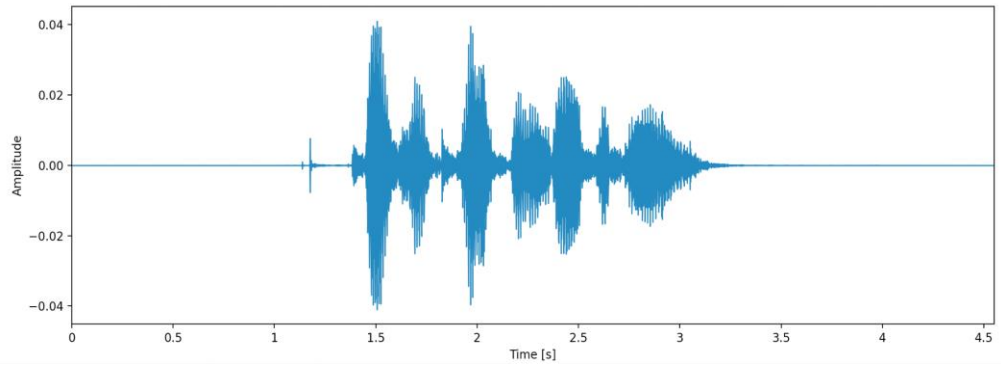
Input dimensions and the number of data samples affect the complexity of learning algorithms. Decreasing dimensions provide reduced memory, less variance, an easier understanding and facilitate computation and testing inference. Feature selection and feature extraction approaches are widely used for dimension reduction. Feature selection can be described as finding particular dimensions beyond all dimensions and discarding remaining features. Feature extraction can be described as generating a new combination of dimensions (Alpaydin, 2010).

There are various types of features for an audio signal which can be grouped as acoustic, context information, linguistic and hybrid features. Acoustic features can be analysed mainly in time and frequency dimensions. In this paper, Mel-Frequency Cepstral Coefficients (MFCC) and Chroma features are selected on frequency domain, Root Mean Square Energy (RMSE) and Zero Crossing Rate (ZCR) features are selected on time domain for comparing time and frequency-based features for speech emotion recognition and wav2vec Large pre-trained model is used for feature extraction.

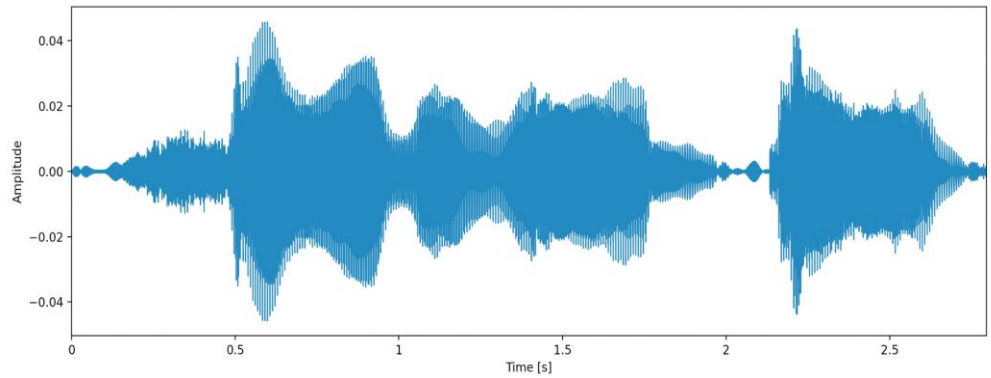
### 2.4.1 Amplitude

Amplitude is about the strength of the voice and shows the level of sound pressure. Amplitude envelope of waveforms that are used in the study plotted as following:

(a)




(b)




(c)

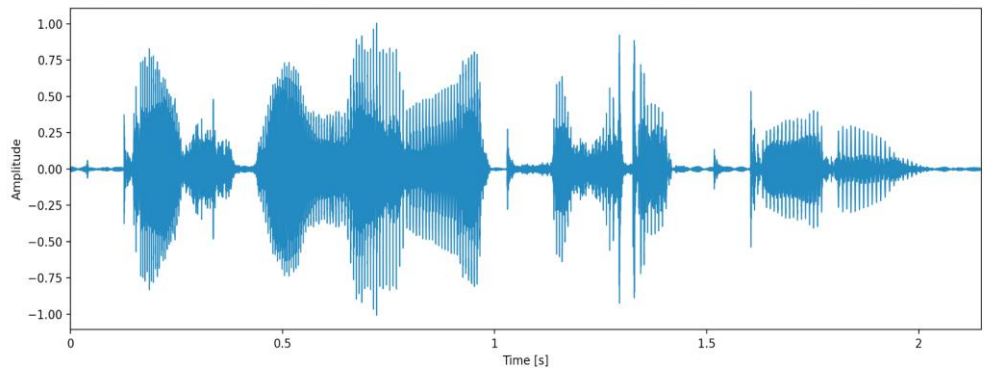
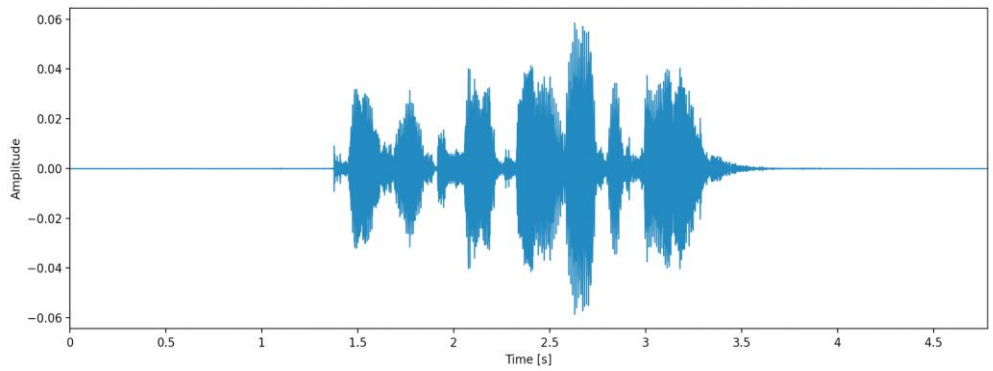
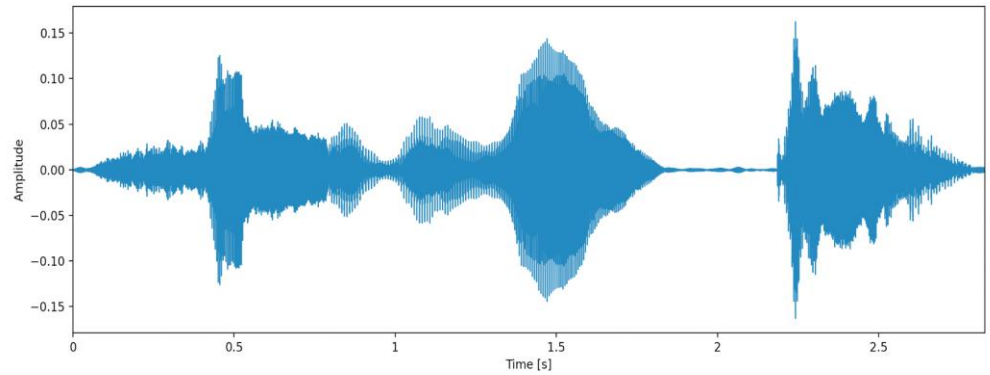



Figure 2.2.1: Waveforms are given in figures and the X-axis shows Time (sec) and the Y-axis shows amplitude. Figures in ‘a’ to ‘c’ belong to the emotion of neutral. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterances. In (b) record is taken from OAF actress from the TESS which represents “Say the word dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

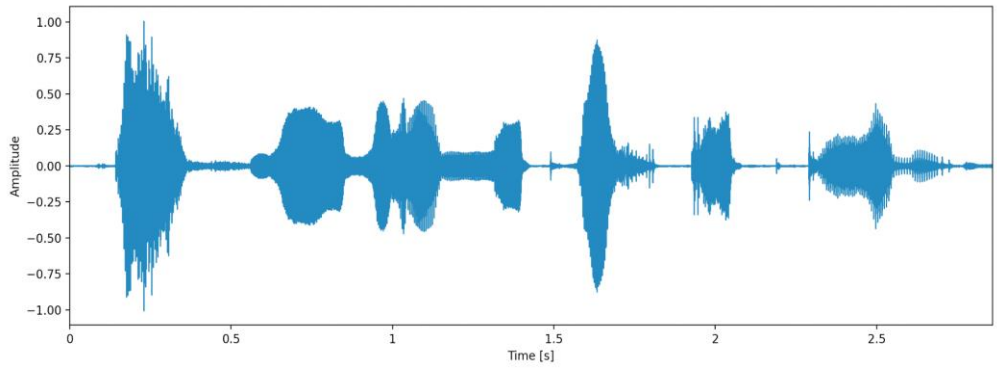
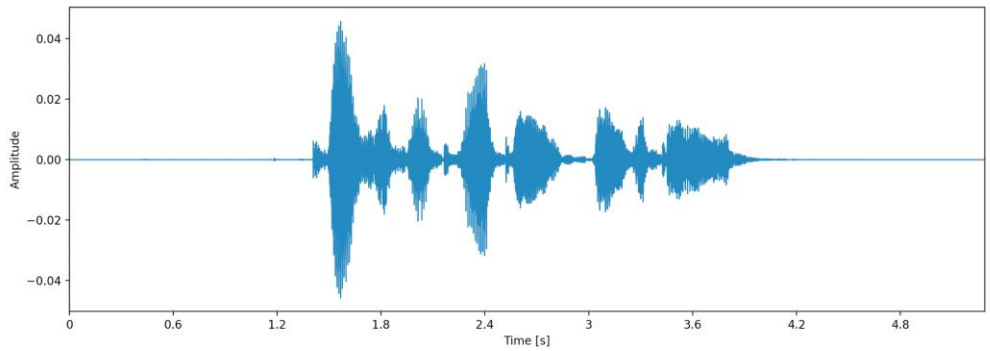
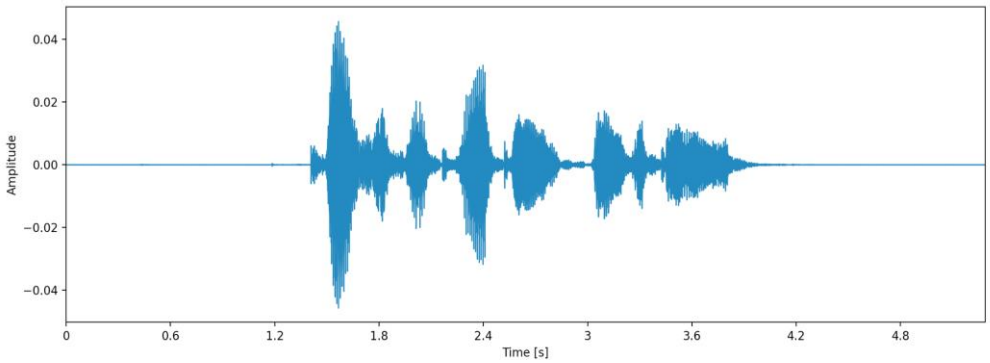



Figure 2.2.2: Waveforms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in 'a' to 'c' belong to the emotion of happiness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

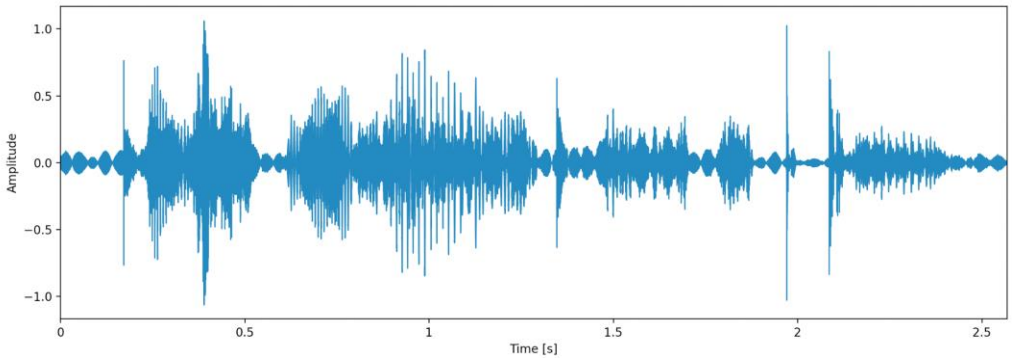
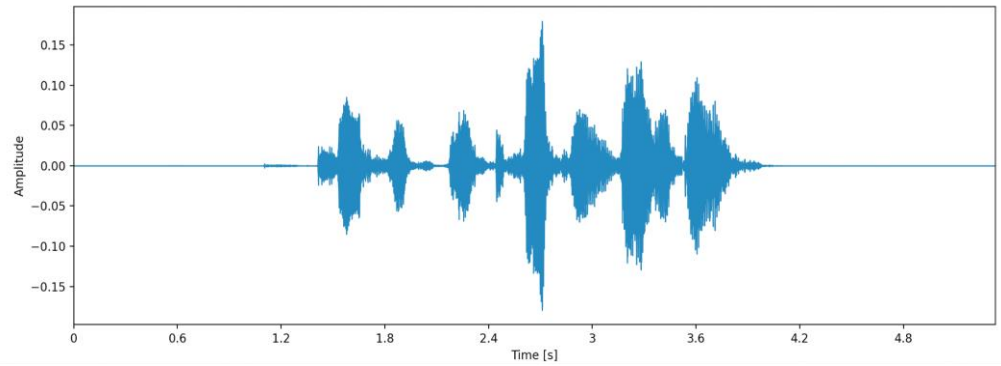
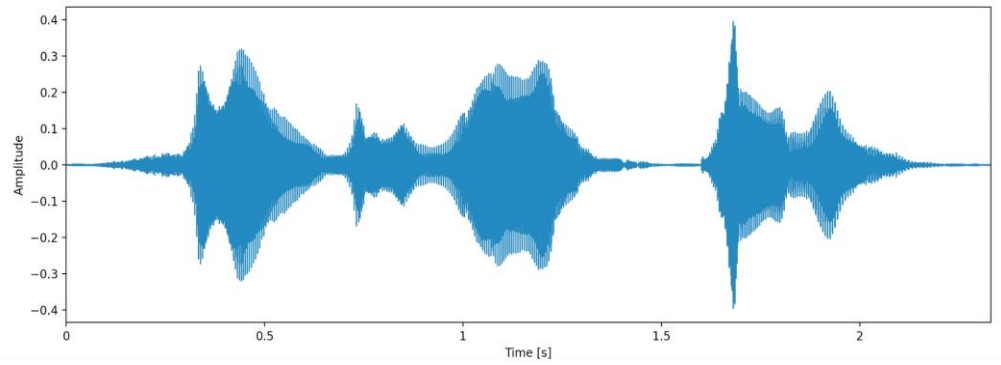



Figure 2.2.3: Waveforms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Amplitude. Figures in 'a' to 'c' belong to the emotion of sadness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday"

(a)




(b)




(c)

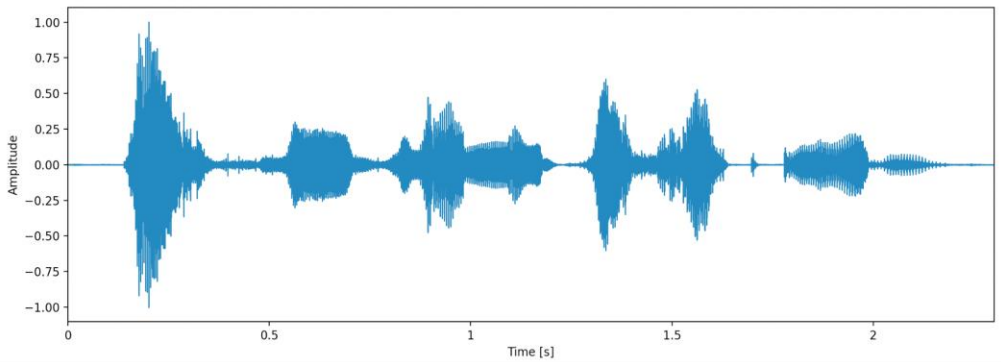
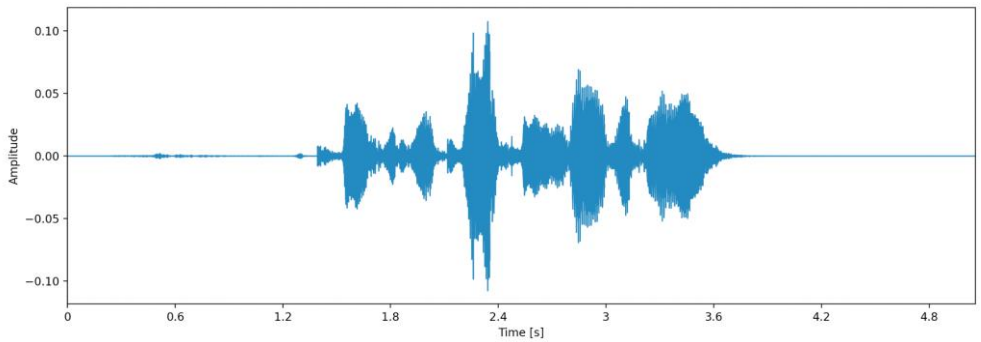
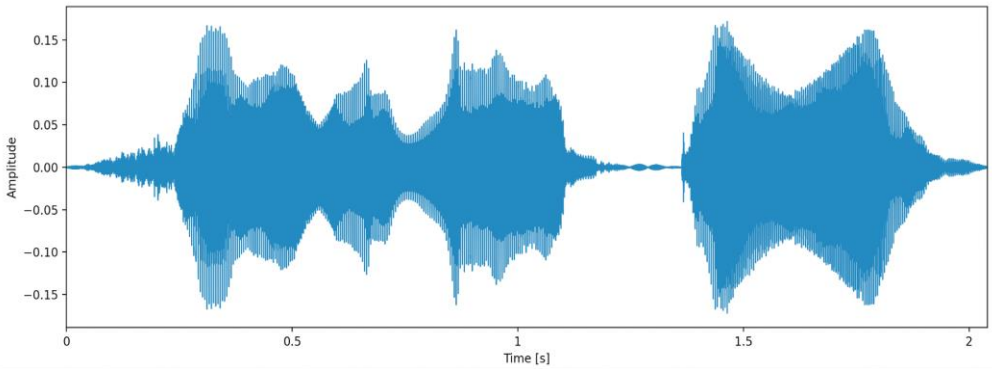



Figure 2.2.4: Waveforms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in ‘a’ to ‘c’ belong to the emotion of anger. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”

(a)




(b)




(c)

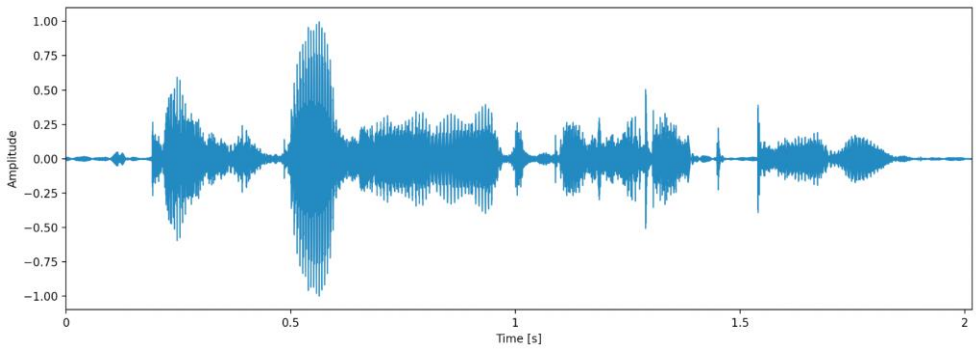
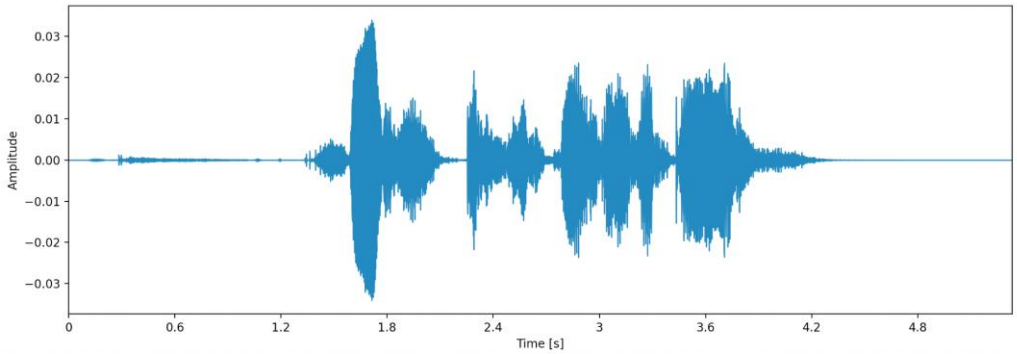
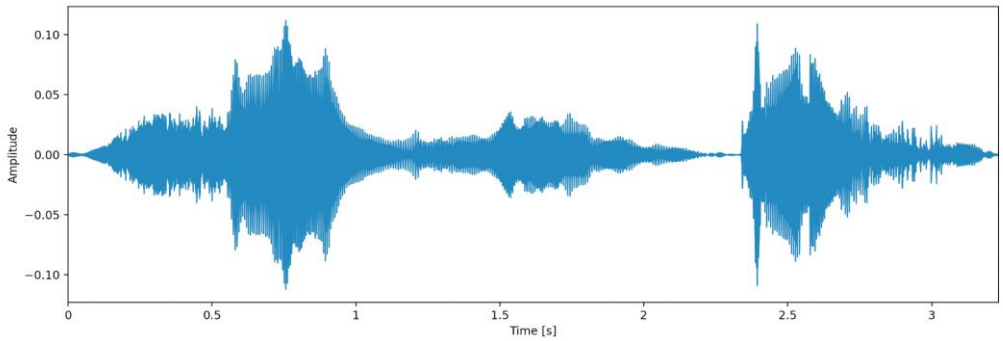



Figure 2.2.5: Waveforms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in ‘a’ to ‘c’ belong to the emotion of fear. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”

(a)




(b)




(c)

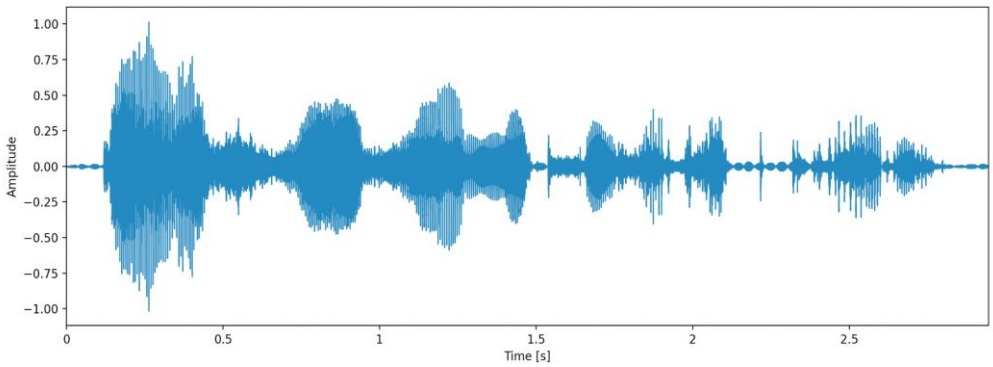
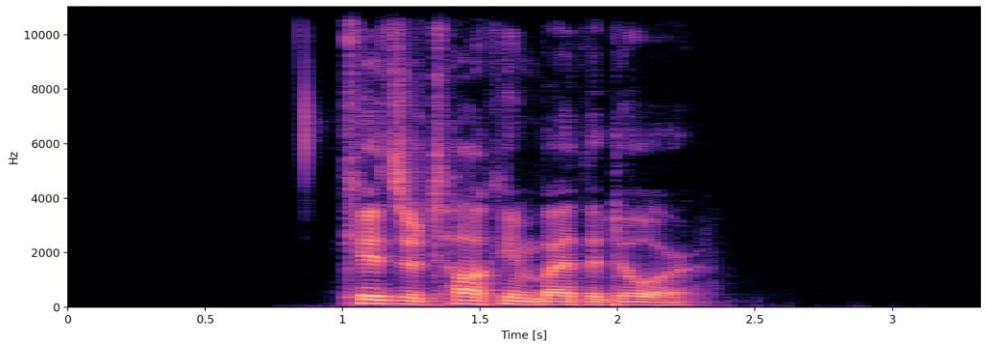



Figure 2.2.6: Waveforms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in 'a' to 'c' belong to the emotion of disgust. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

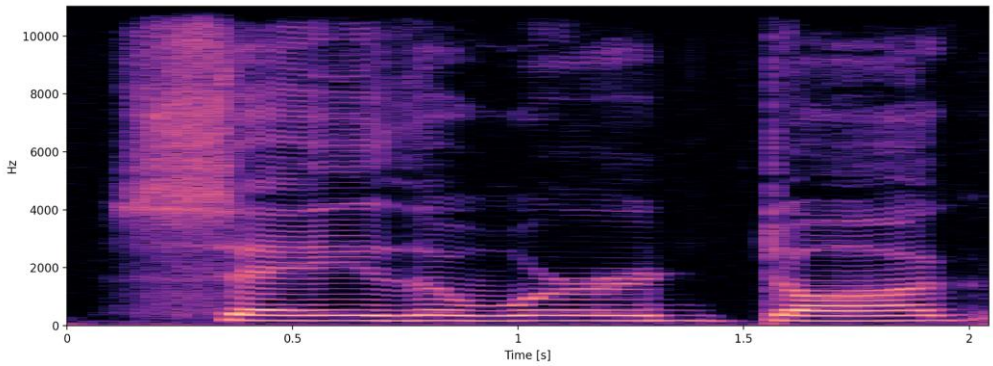
### 2.4.2 Spectrogram

Spectrogram is a representation of sound and displays frequency or other signals according to their changes to time. Example spectrograms that are used in the study plotted as following:

(a)




(b)




(c)

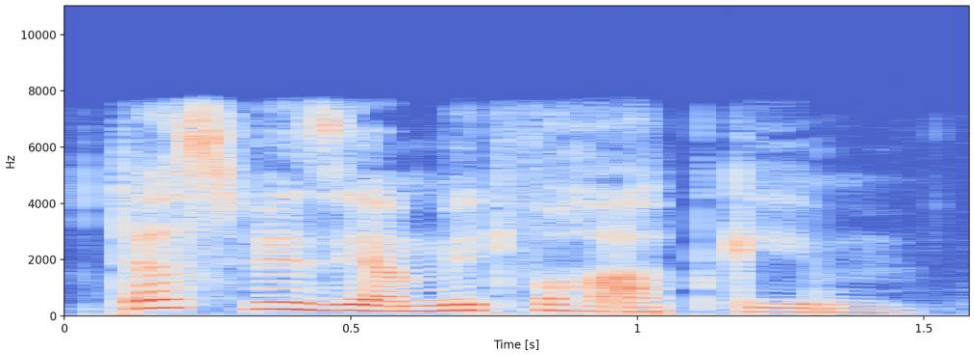
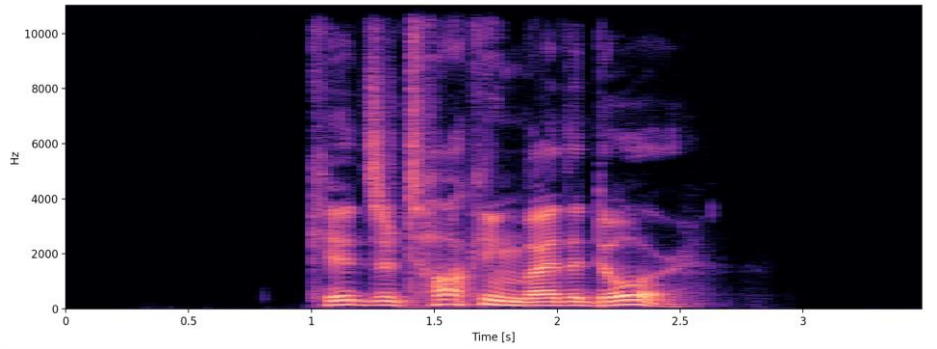
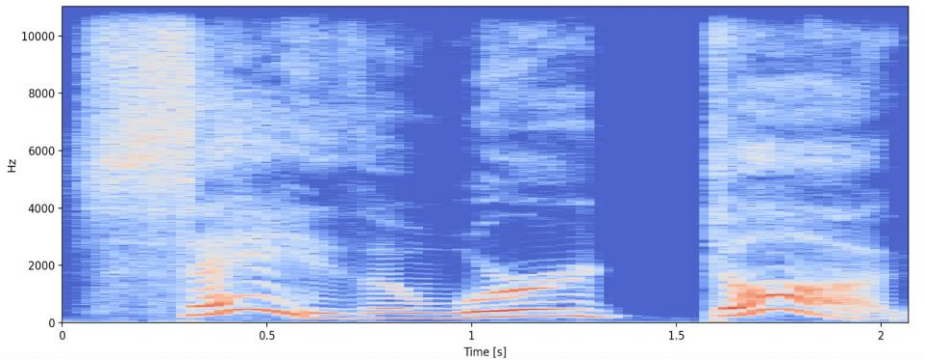



Figure 2.2.7: Spectrograms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in 'a' to 'c' belong to the emotion of neutral. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

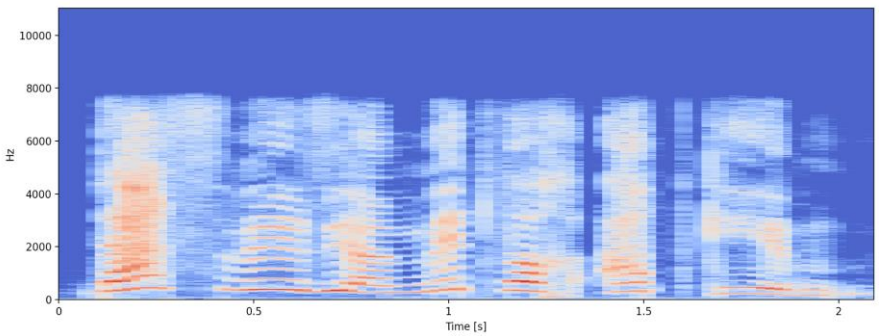
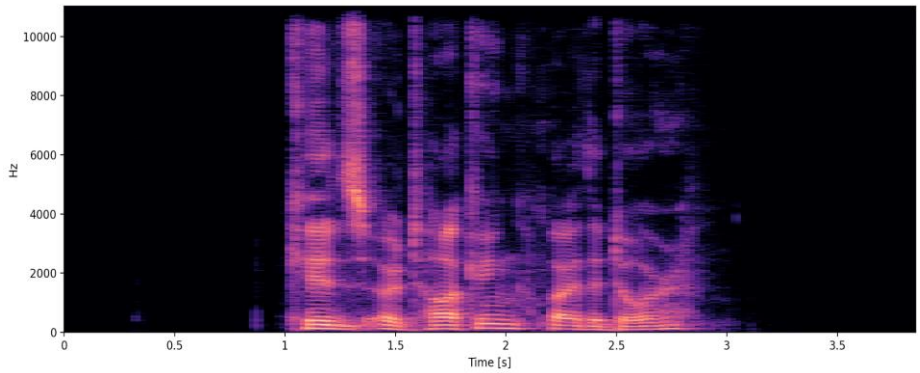
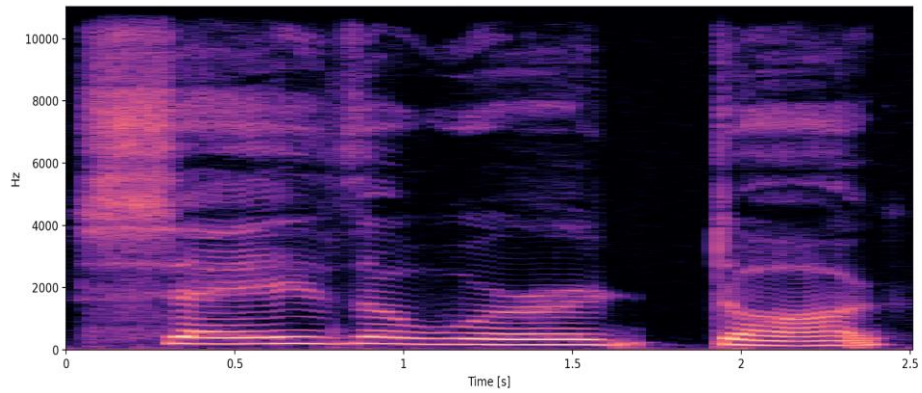



Figure 2.2.8: Spectrograms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in 'a' to 'c' belong to the emotion of happiness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

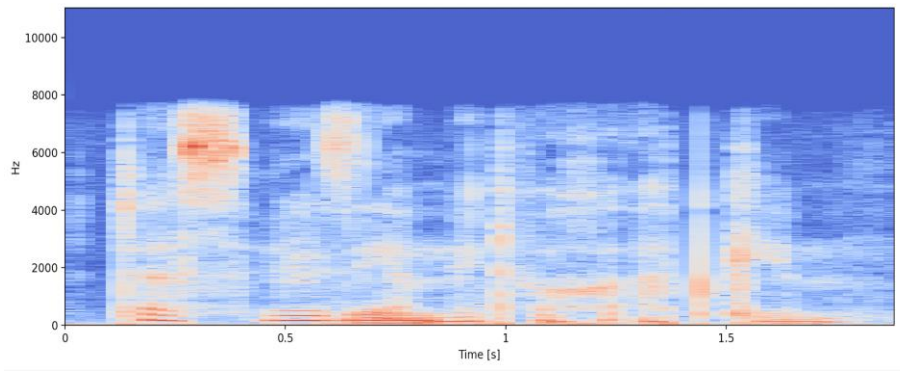
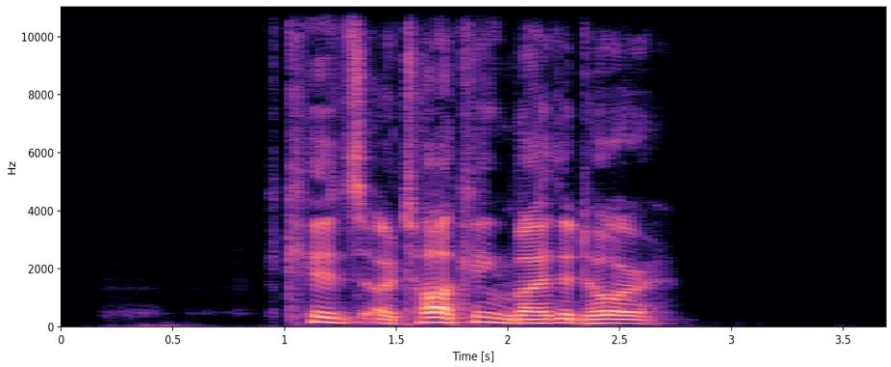
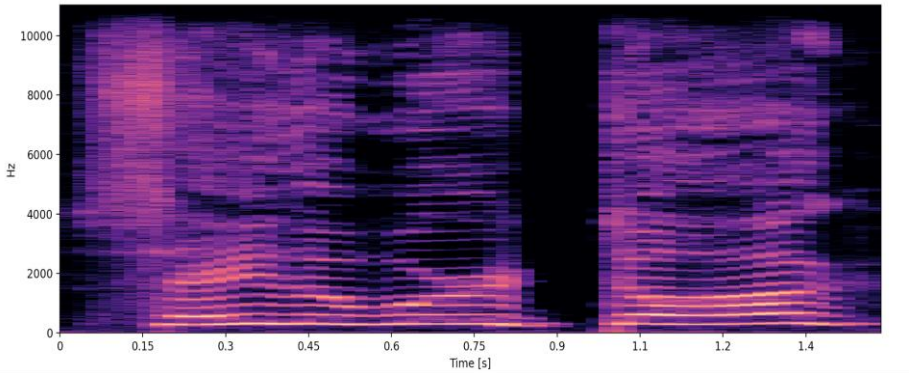



Figure 2.2.9: Spectrograms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in ‘a’ to ‘c’ belong to the emotion of sadness. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”

(a)




(b)




(c)

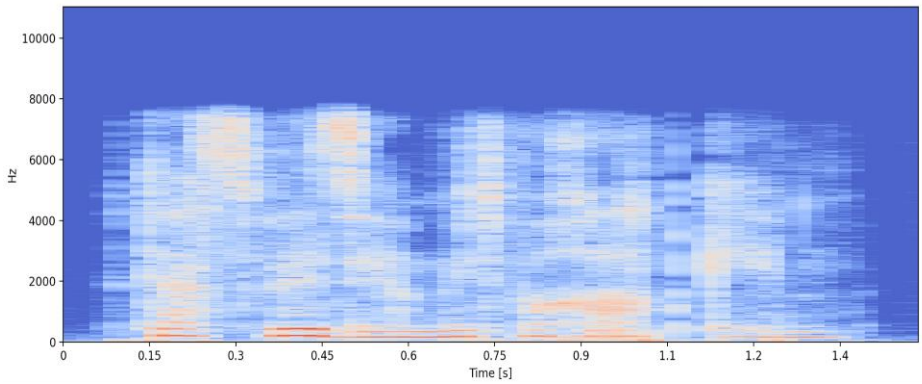
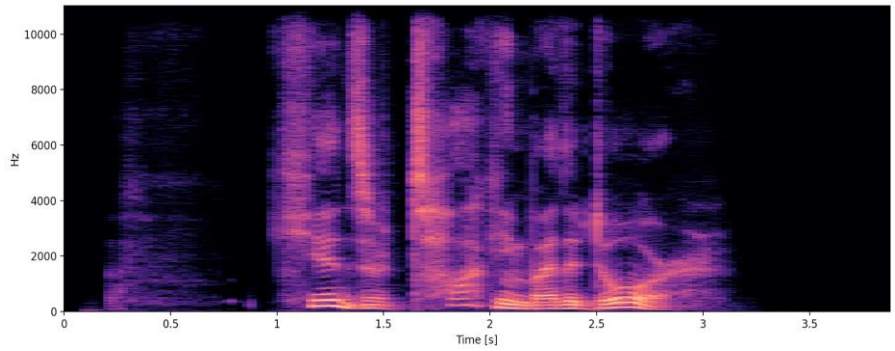
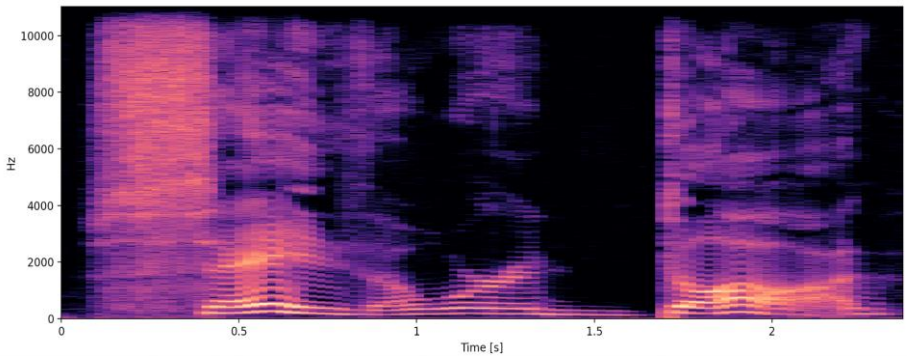



Figure 2.2.10: Spectrograms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in ‘a’ to ‘c’ belong to the emotion of fear. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

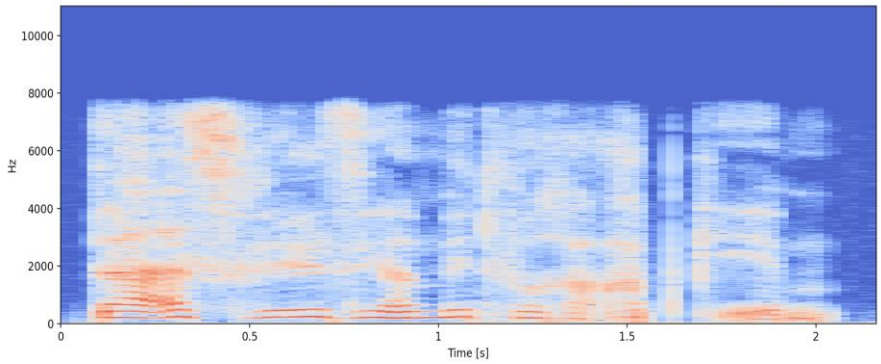
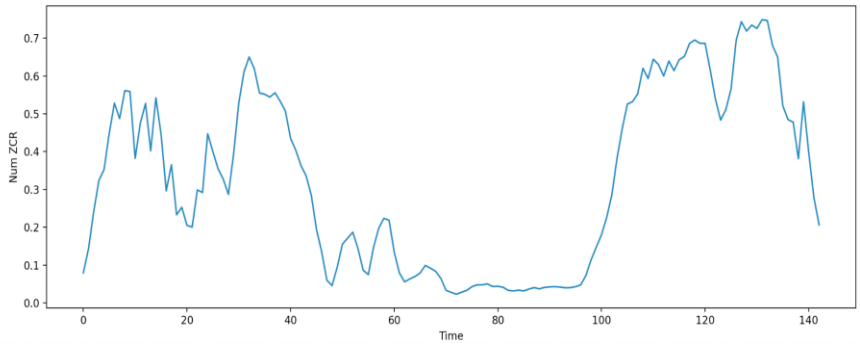



Figure 2.2.11: Spectrograms are given in figures and the X-axis shows Time (sec) and the Y-axis shows Frequency (Hz). Figures in ‘a’ to ‘c’ belong to the emotion of disgust. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

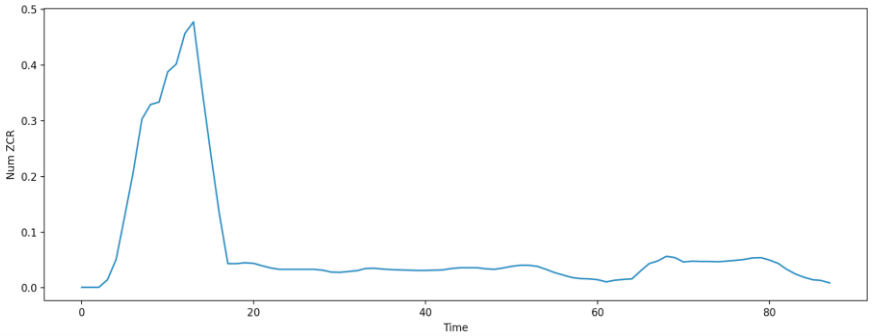
### 2.4.3 Zero Crossing Rate

Zero Crossing Rate (ZCR) is the number of times the signal changes value from positive to negative and vice versa in a specific time interval. It is used generally for voice activity detection. Example ZCR of audio records that are used in the study plotted as following:

(a)




(b)




(c)

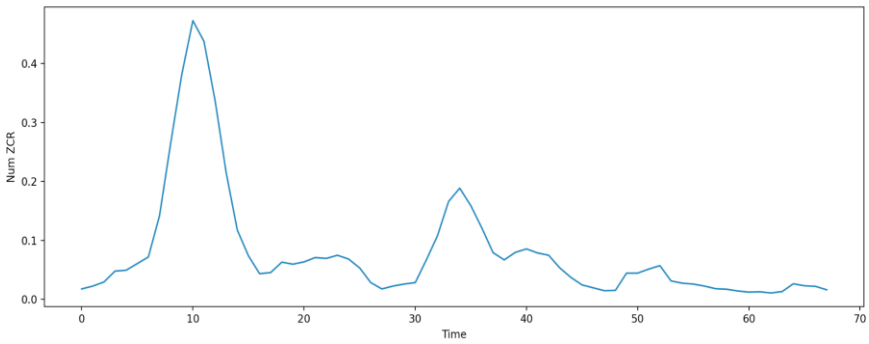
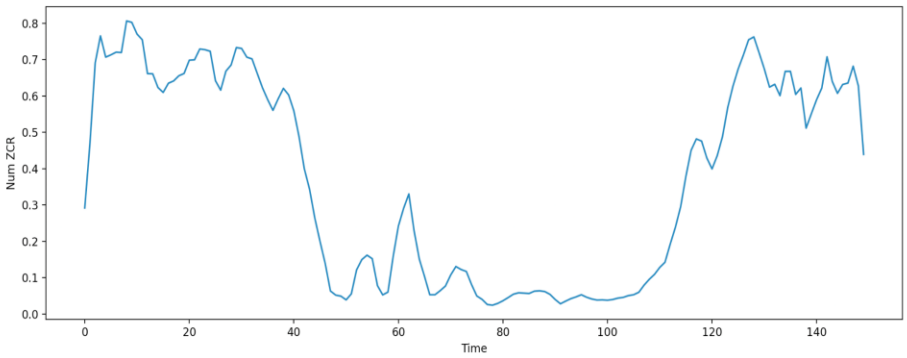
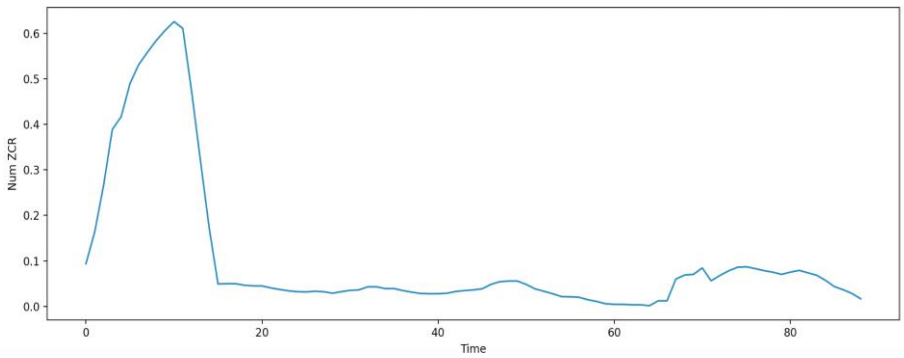



Figure 2.2.12: Number of zero cross rates are given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of neutral. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

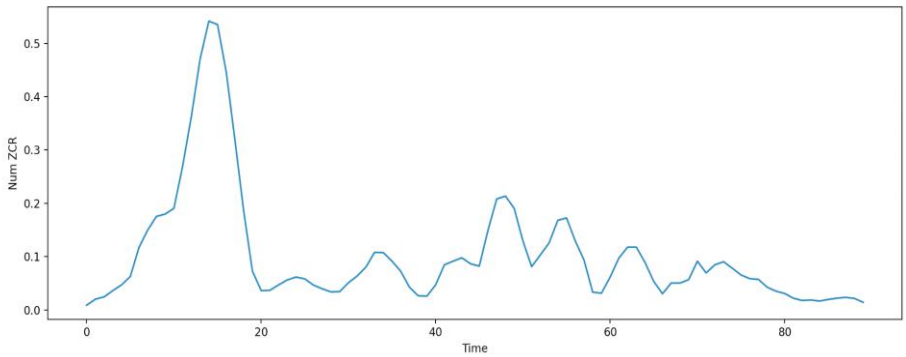
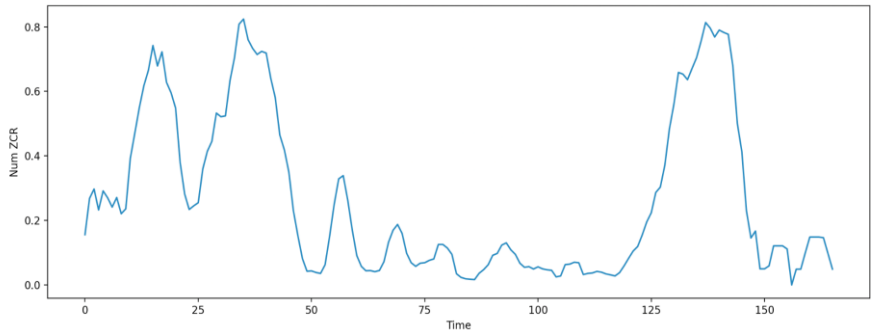
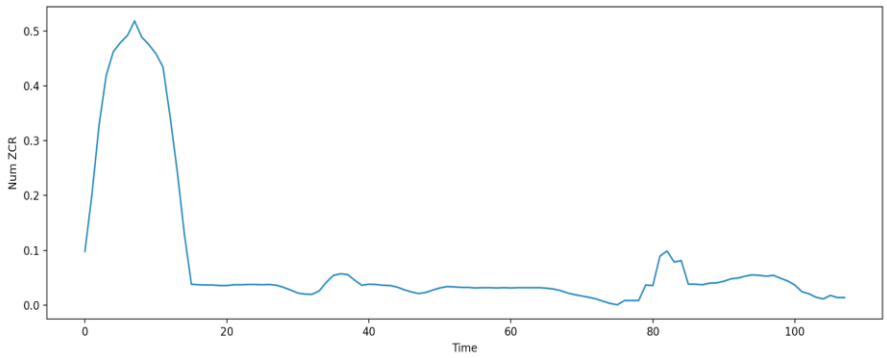



Figure 2.2.13: Number of zero cross rates are given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of happiness. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”

(a)




(b)




(c)

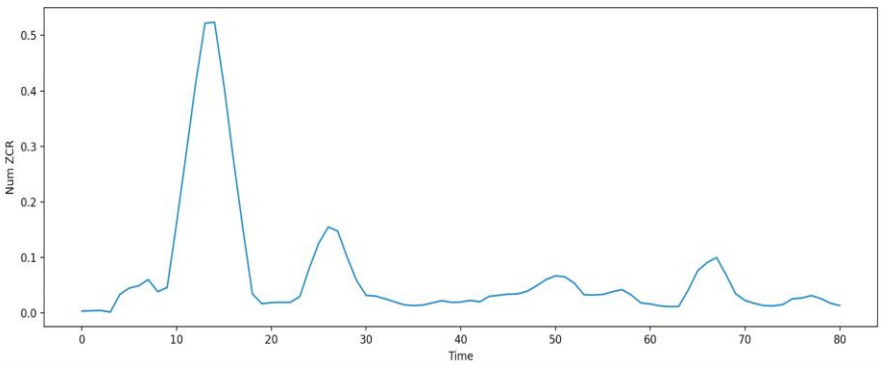
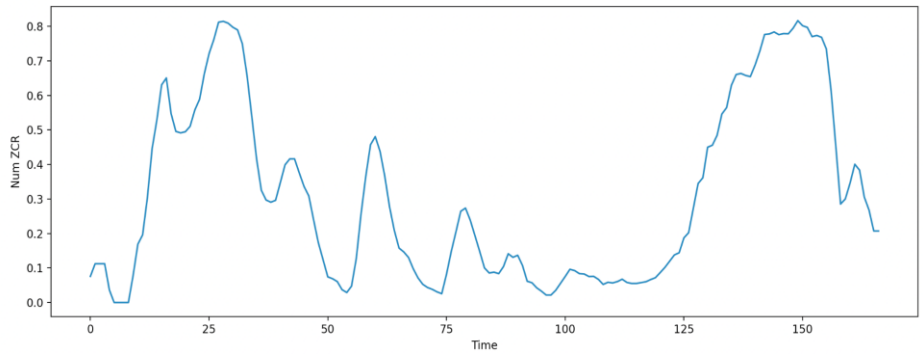
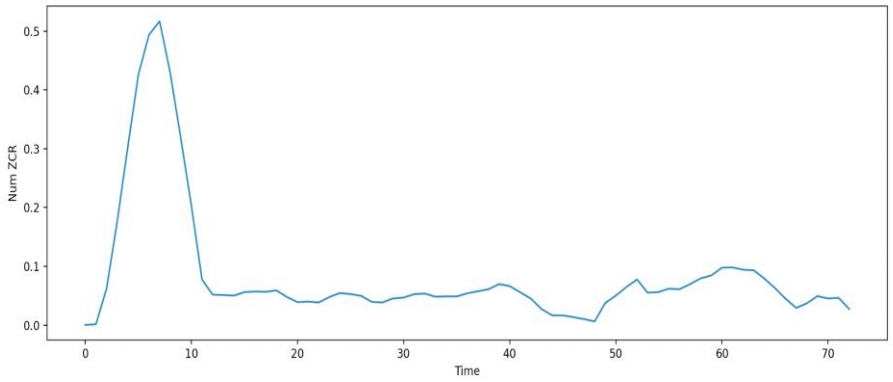



Figure 2.2.14: Number of zero cross rates are given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of sadness. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

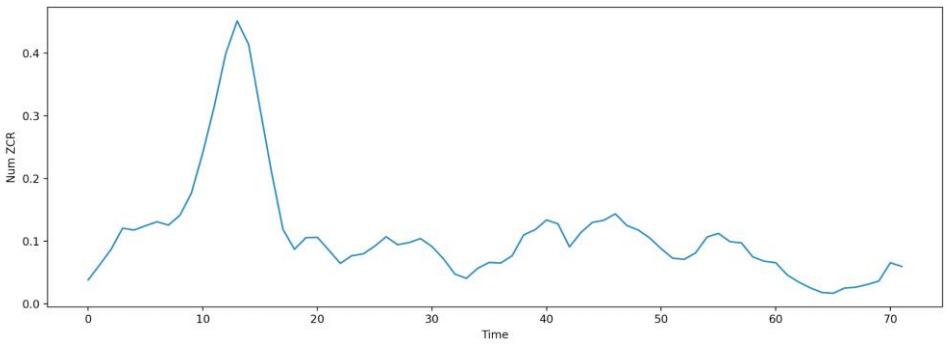
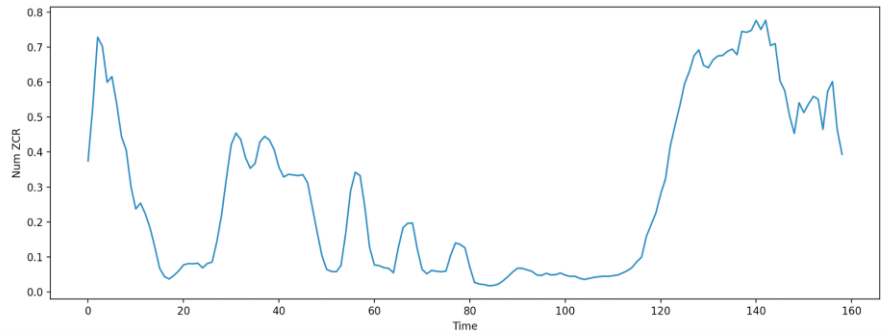
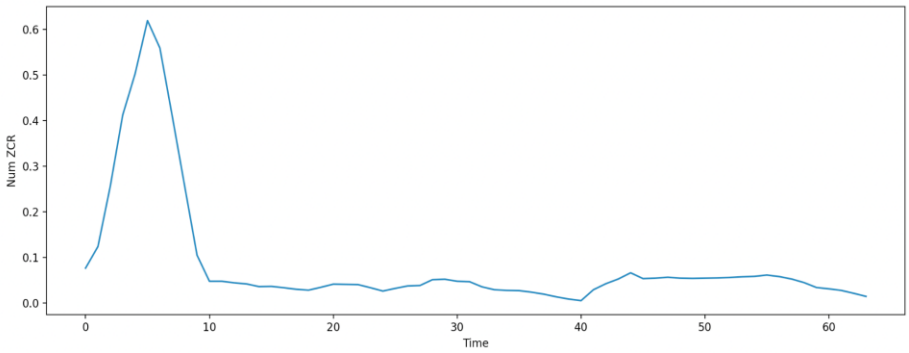



Figure 2.2.15: Number of zero cross rates are given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of anger. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

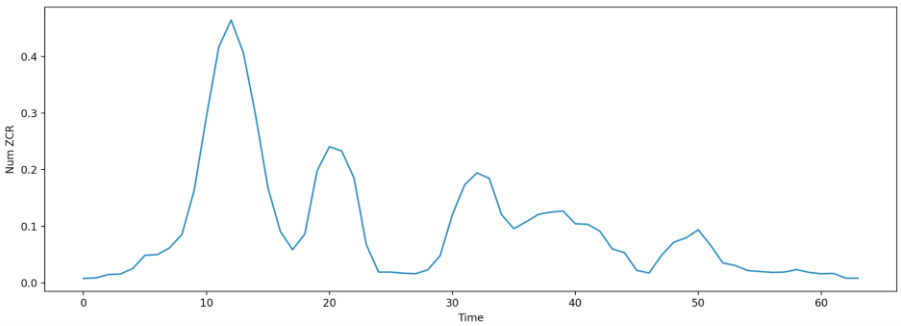
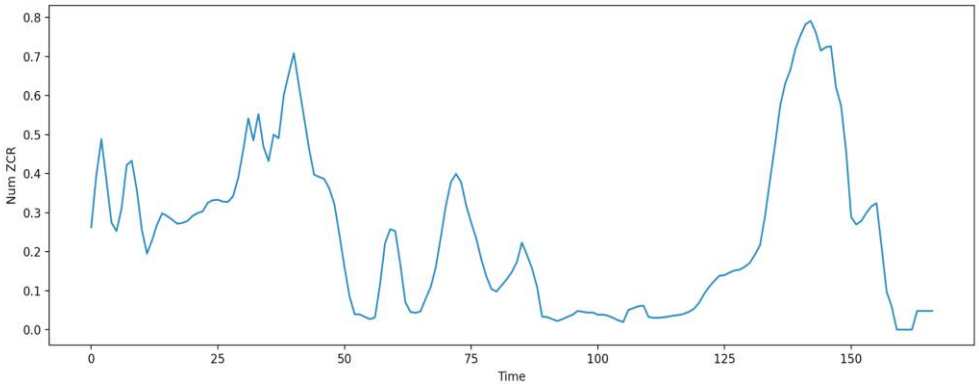
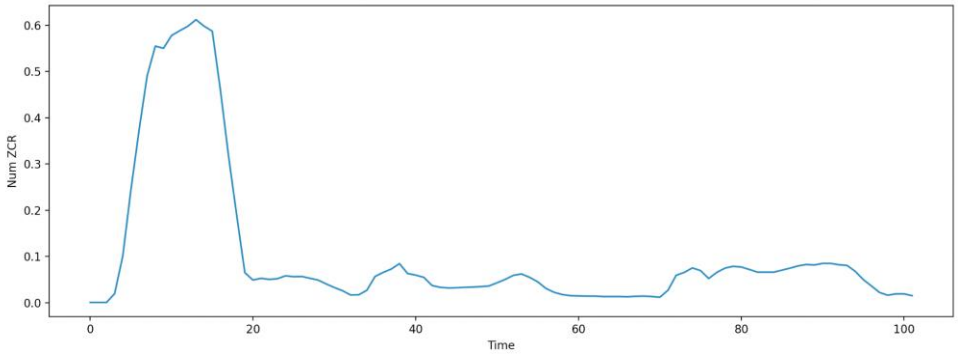



Figure 2.2.16: Number of zero cross rates are given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of fear. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

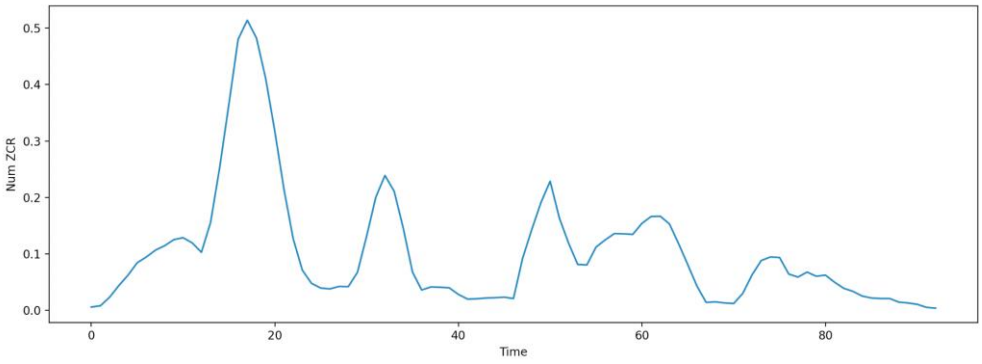
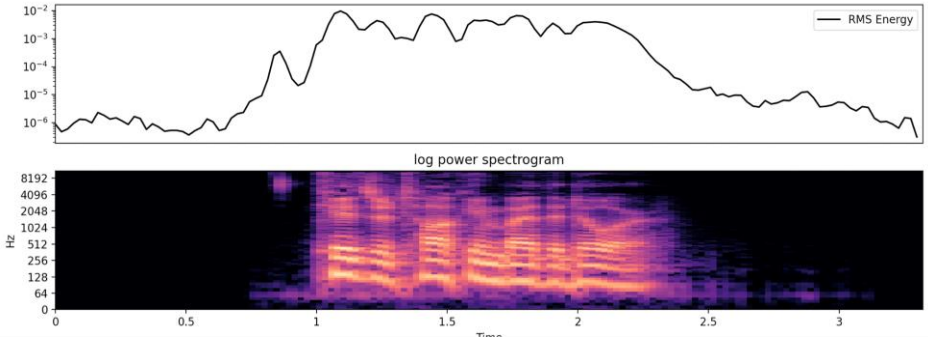



Figure 2.2.17: Number of zero cross rates are given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of disgust. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

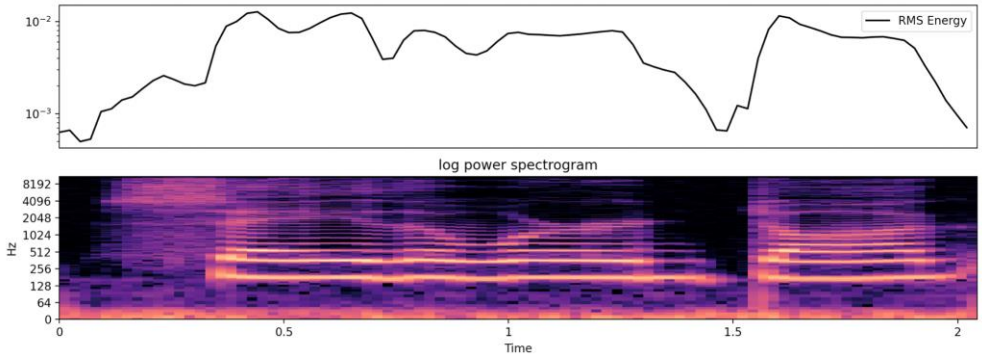
### 2.4.4 Root mean square energy

The Root Mean Square Energy (RMSE) is related to the energy that is carried by the wave. It is the square root of the magnitude of the audio frames mean square over a given duration. In this study audio raws are converted to spectrograms with STFT and then RMSE values are calculated. Example RMSE of audio records that are used in the study plotted as following:

(a)




(b)




(c)

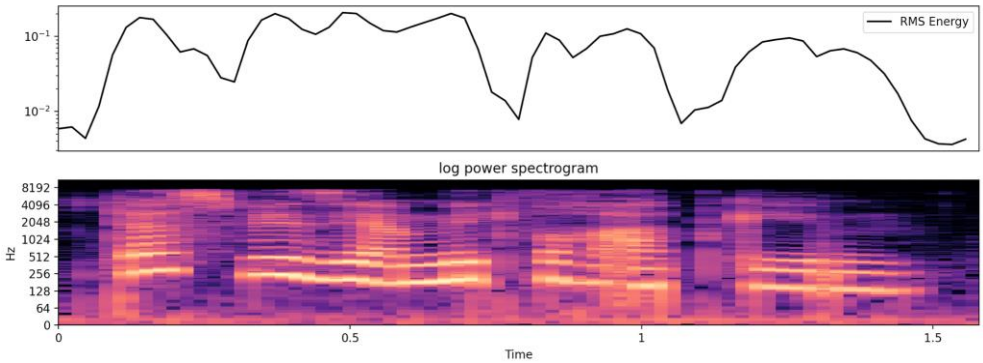
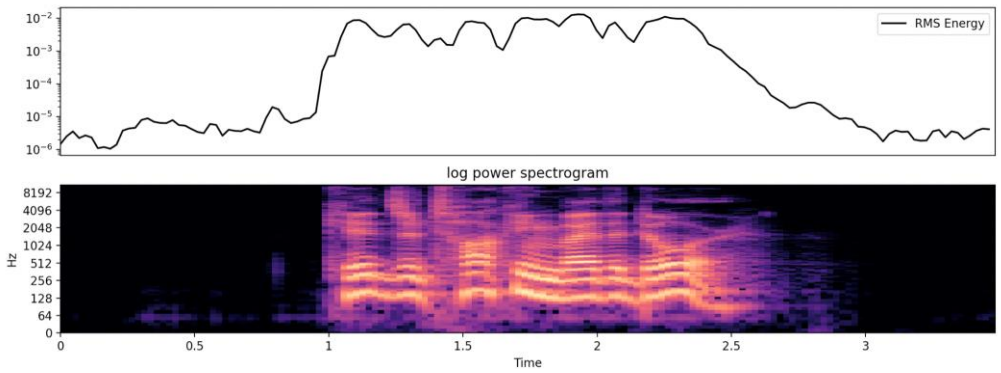
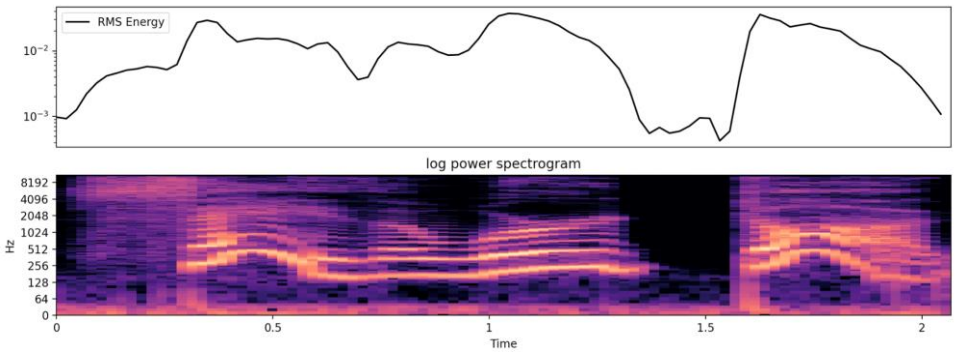



Figure 2.2.18: RMS Energy is given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of neutral. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

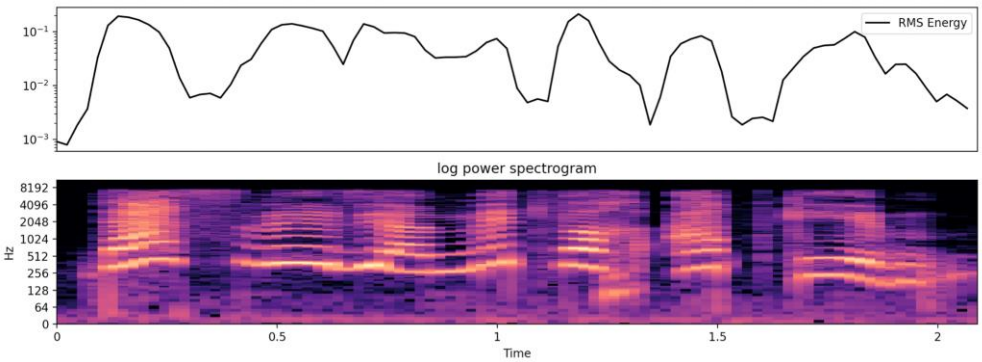
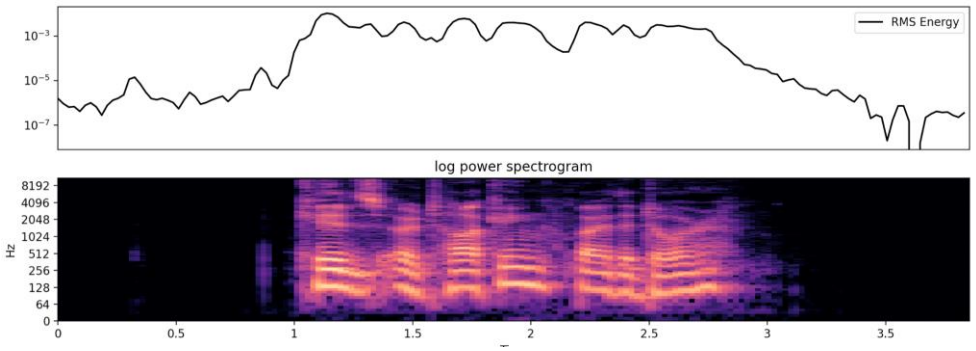
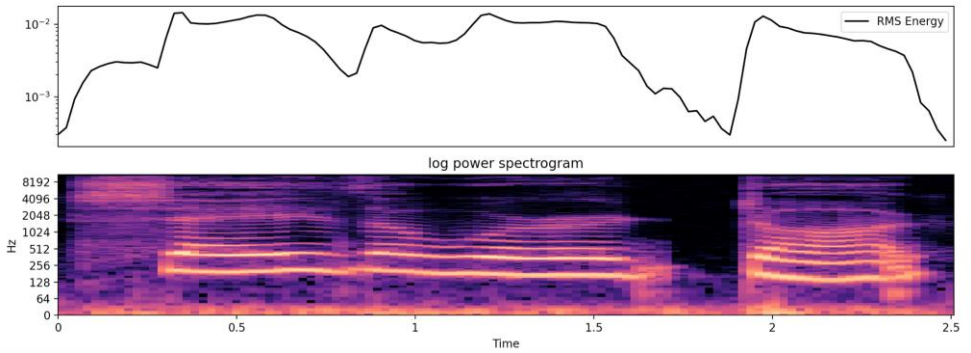



Figure 2.2.19: RMS Energy is given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of happiness. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

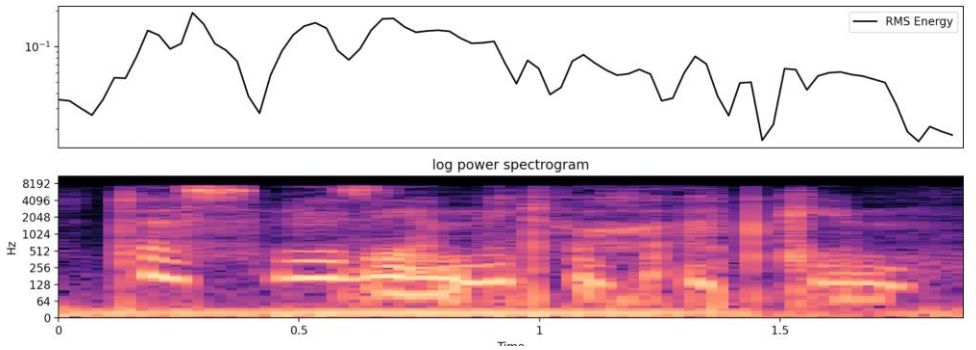
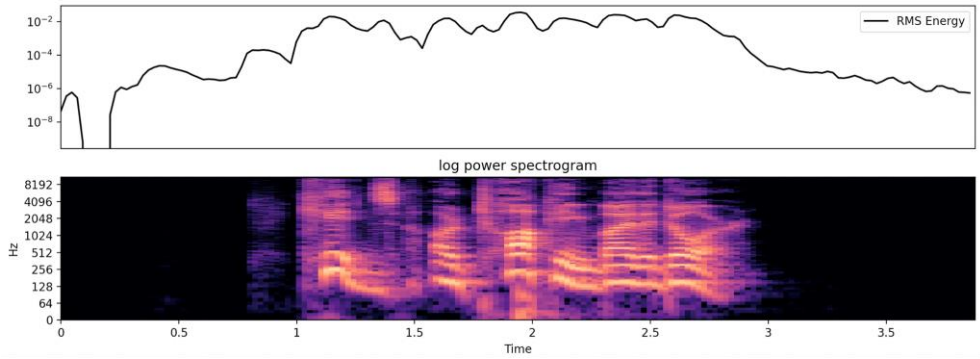
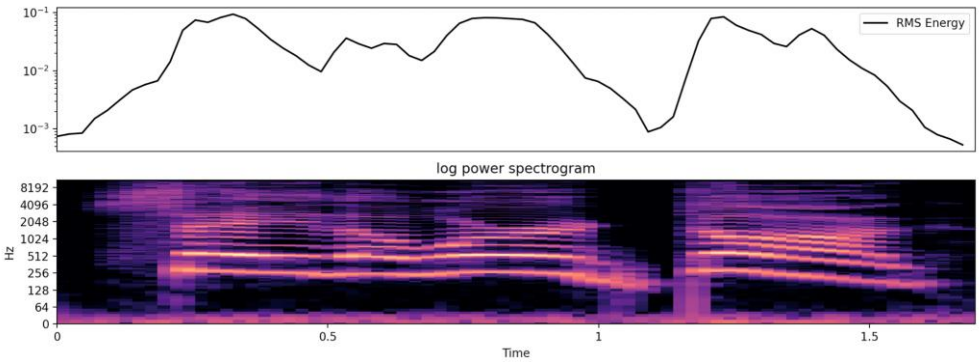



Figure 2.2.20: RMS Energy is given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of sadness. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

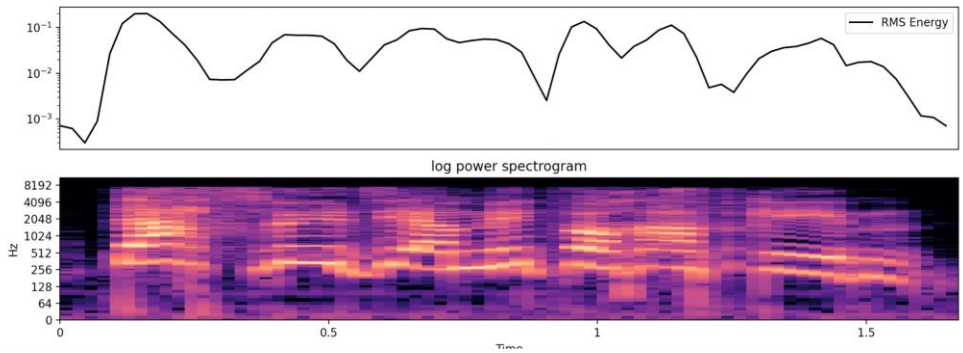
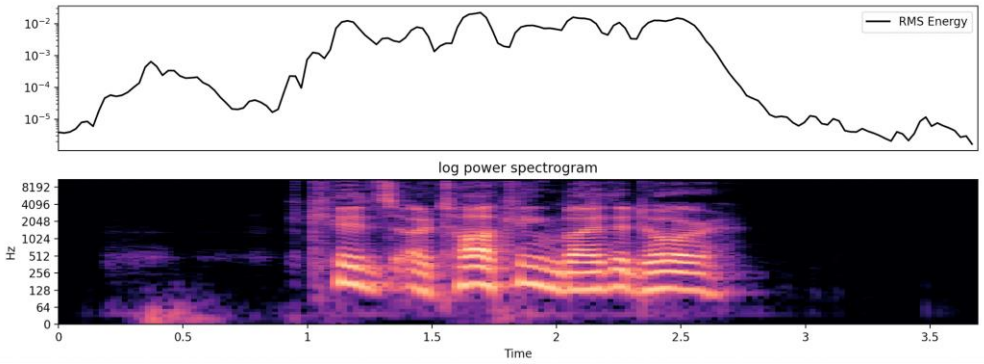
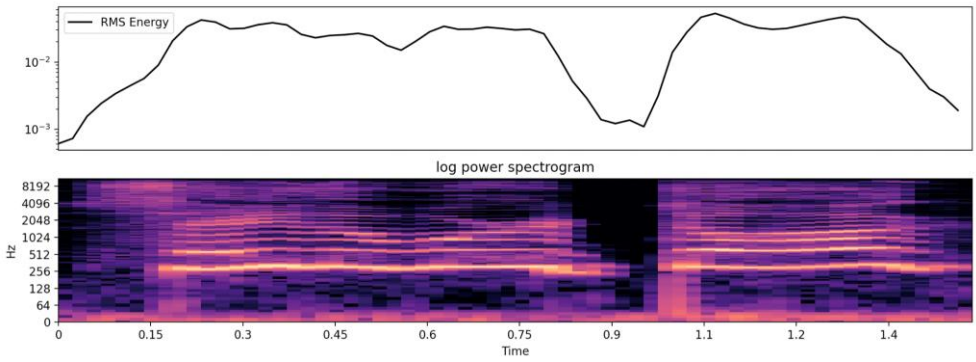



Figure 2.2.21: RMS Energy is given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of anger. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

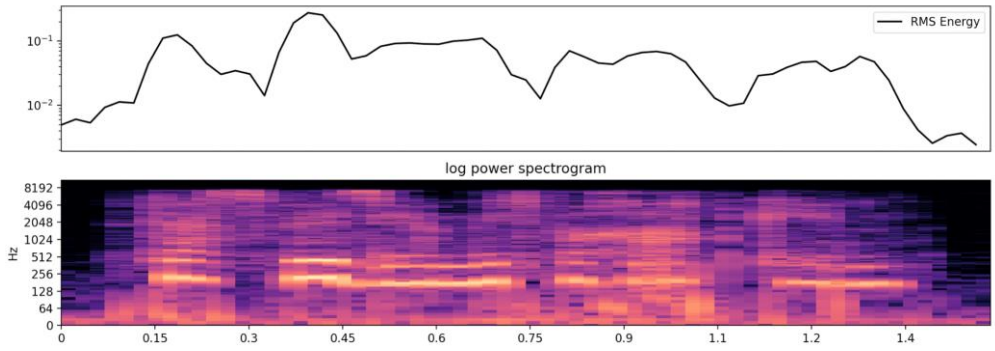
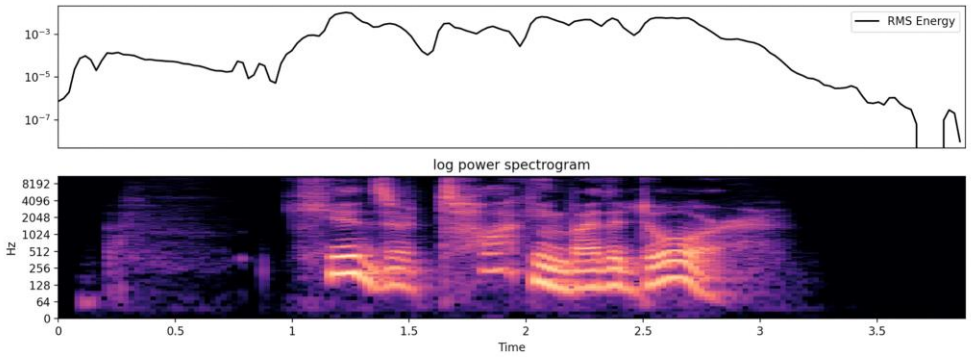
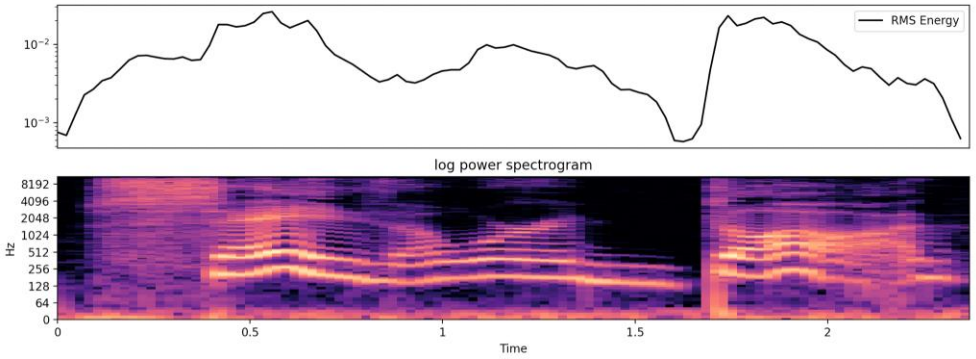



Figure 2.2.22: RMS Energy is given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of fear. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

(a)




(b)




(c)

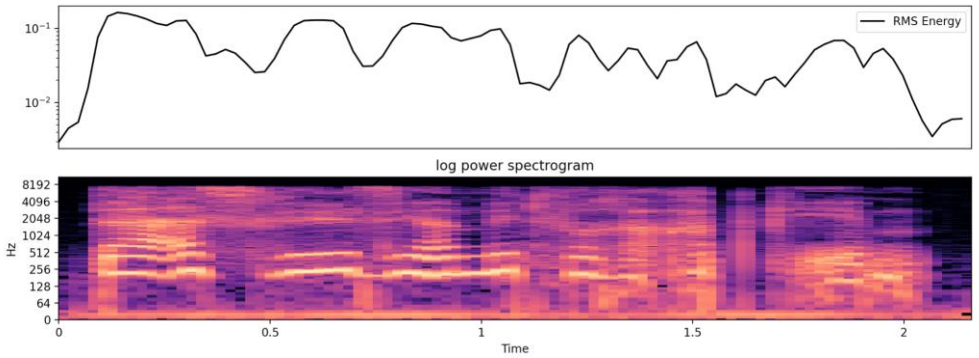



Figure 2.2.23: RMS Energy is given in figures. Figures in ‘a’ to ‘c’ belong to the emotion of disgust. In (a) record is taken from speaker 1 from RAVDESS which represents “Kids are talking by the door” utterance. In (b) record is taken from OAF actress from the TESS which represents “Say the world dog” utterance. In (c) record is taken from speaker 13 from EMODB for “Das will sie am Mittwoch abgeben” utterance which means “She will hand it in on Wednesday”.

### 2.4.5 Mel-Frequency cepstral coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCC) is a common feature extraction technique for audio signals. MFCC is based on the variation of the human ear's critical bandwidths with frequency filters and represents distinctive speech features. It is a computation technique that uses cosine transform of the real logarithm of the short-term energy spectrum and warping the frequencies on a Mel scale (Zheng et al., 2001). Mel scale has linear frequency spacing below 1000 Hz and logarithmic spacing above 1000 Hz. Pitch of 1 kHz tone and 40 dB above the perceptual audible threshold is defined as 1000 mels and used as a reference point (De Lara, 2005). The process of MFCC is figured below and can be explain as following:

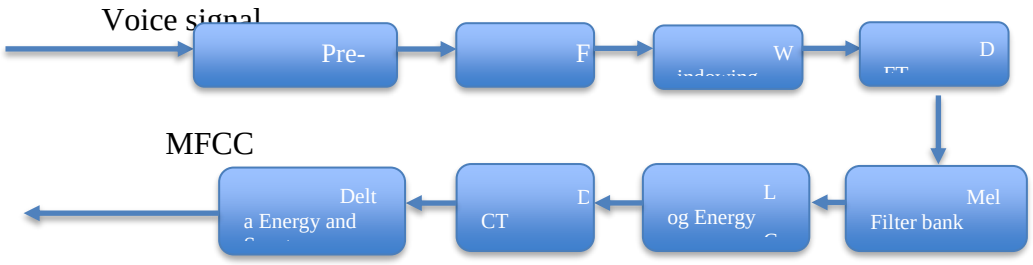
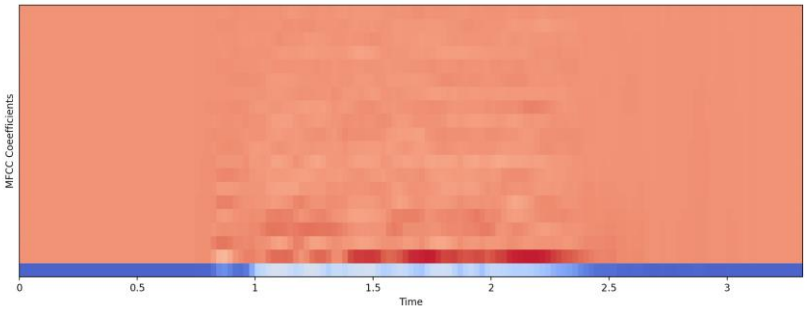



Figure 2.2.24: MFCC block diagram

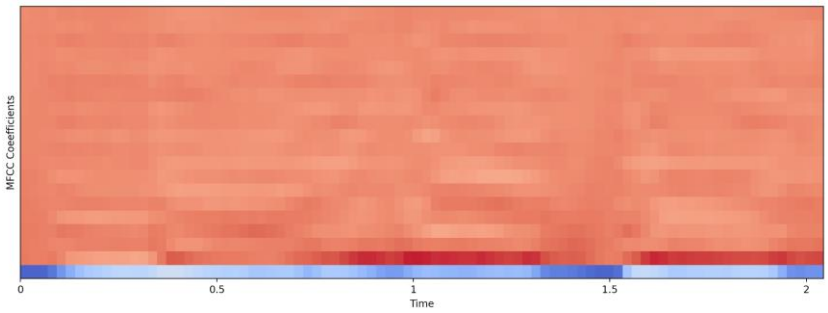
1. Emphasize speech signal with a filter to amplify high frequency regions.
2. Divide pre-emphasized speech signal into frames of generally 20 ms with 10 ms overlap.
3. Convert discrete-time signals obtained by windowing frames to discrete frequency samples and get magnitude spectrums
4. Convert Fourier transformed signal to Mel Scale passed through a filter bank of frequency response that is known as Mel Filter Bank.
5. Compute logarithm of the square magnitude of the output of the mel-filter bank and obtain a log-spectral-energy vector for each frame
6. Apply Discrete Cosine Transform on log energy output obtained from bandpass filters to generate mel-scale cepstral coefficients

MFCC of audio records that are used in the study plotted as following:

(a)




(b)




(c)

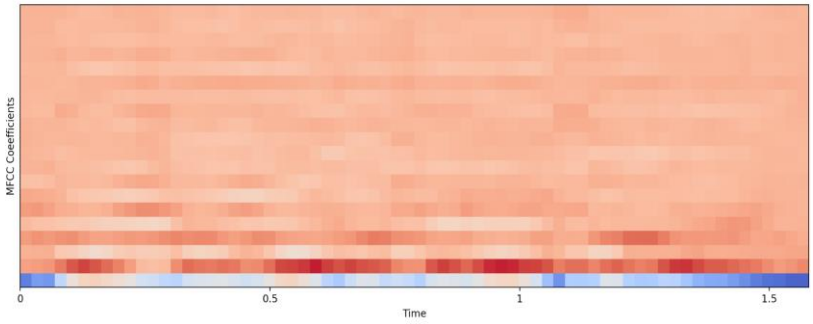
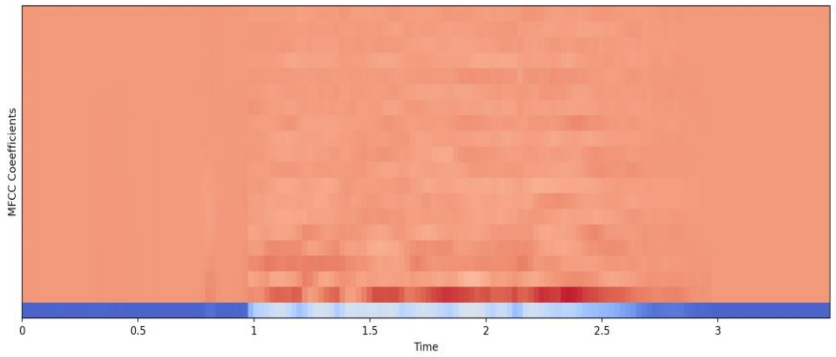
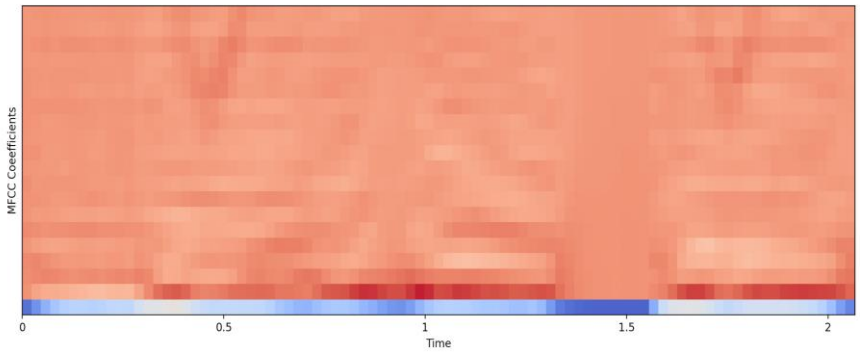



Figure 2.2.25: MFCC's are given in figures. Figures in 'a' to 'c' belong to the emotion of neutral. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday"

(a)




(b)




(c)

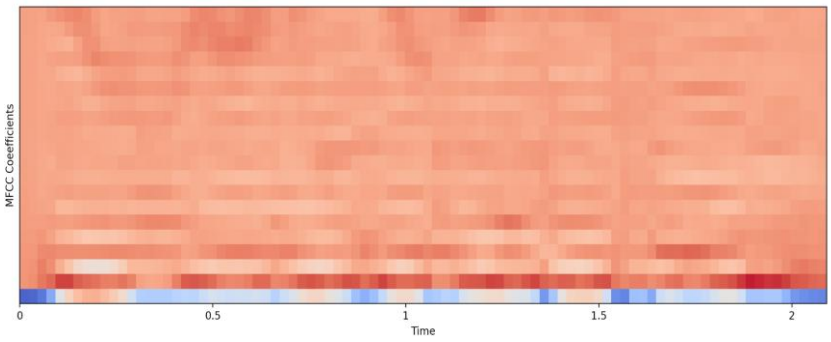
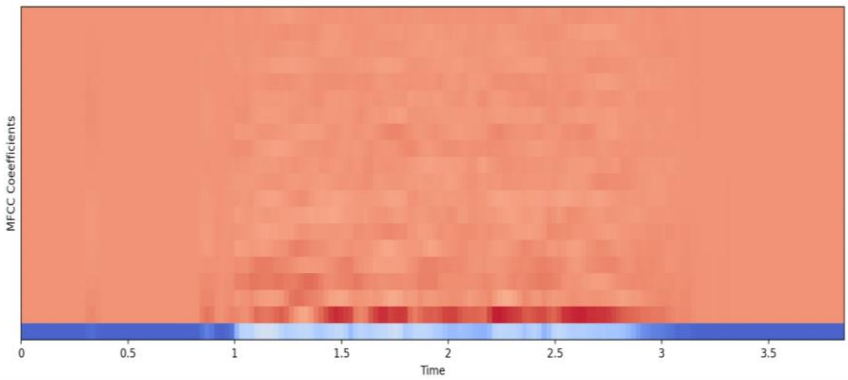
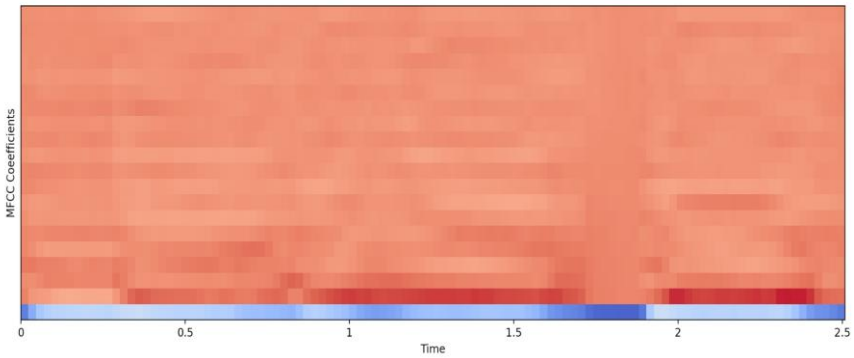



Figure 2.2.26: MFCC's are given in figures. Figures in 'a' to 'c' belong to the emotion of happiness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

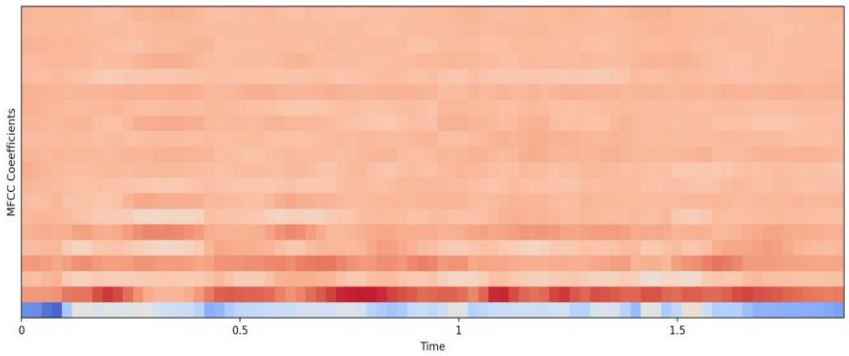
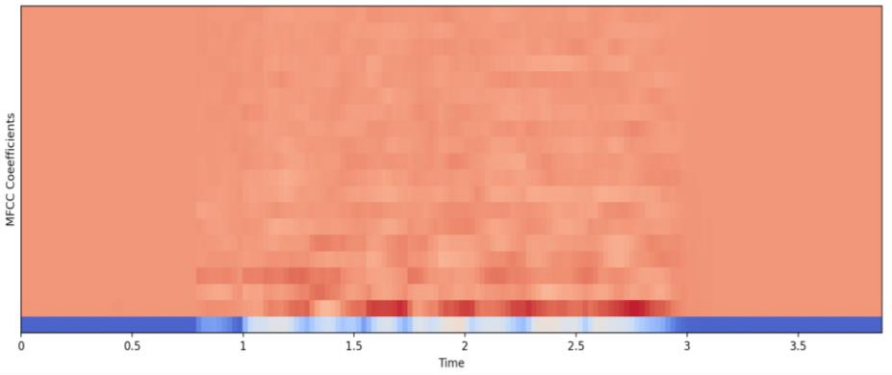
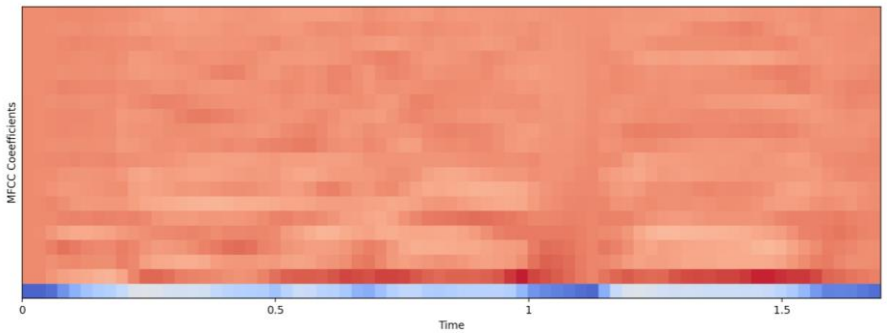



Figure 2.2.27: MFCC's are given in figures. Figures in 'a' to 'c' belong to the emotion of sadness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

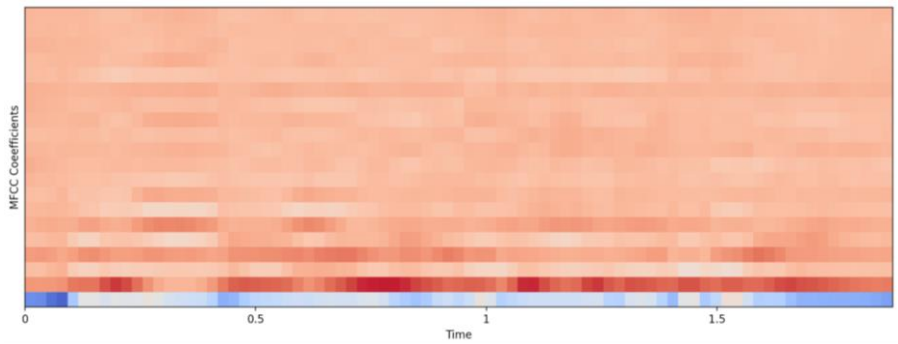
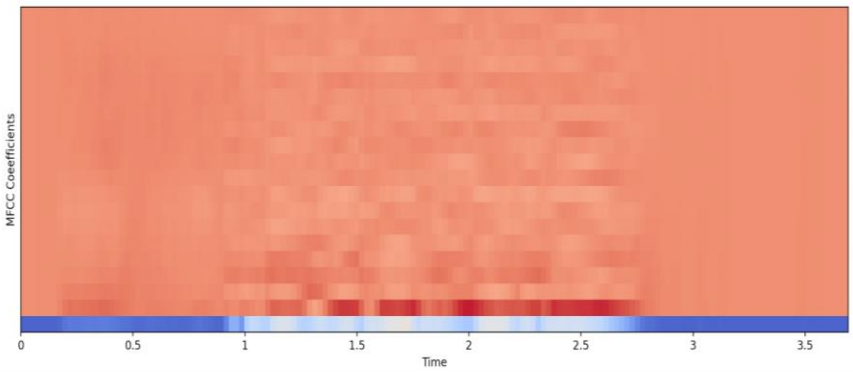
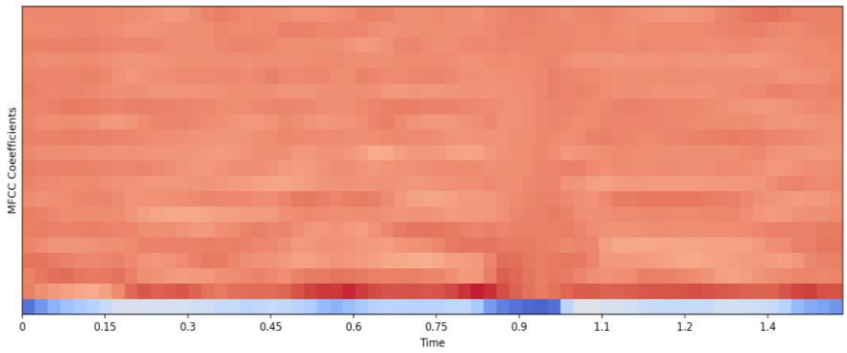



Figure 2.2.28: MFCC's are given in figures. Figures in 'a' to 'c' belong to the emotion of anger In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

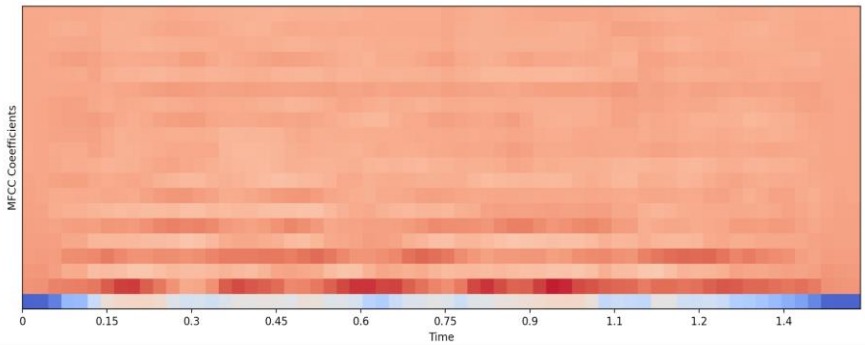
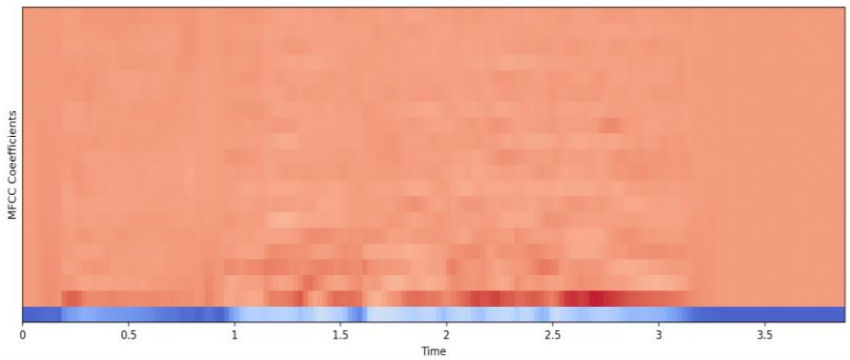
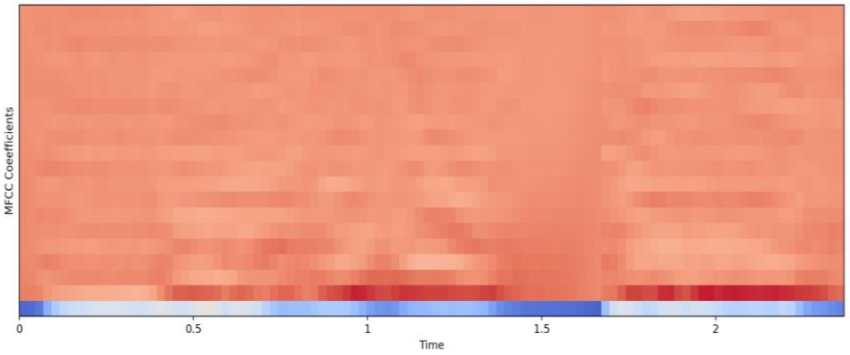



Figure 2.2.29: MFCC's are given in figures. Figures in 'a' to 'c' belong to the emotion of fear. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

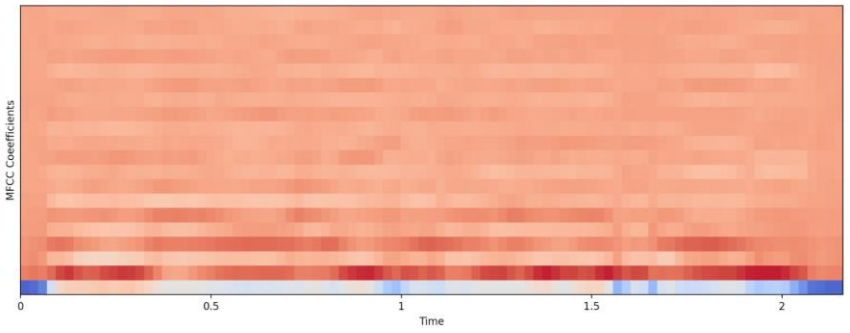



Figure 2.2.30: MFCC's are given in figures. Figures in 'a' to 'c' belong to the emotion of disgust. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

### 2.4.6 Chroma

Chroma represents the inherent circularity of pitch. Chroma values are represented by the following set and in the following page example Chroma for audio records that are used in the study plotted.

{C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, B}

In the chroma circle, pitches which are two octaves related are at the same angle. 12 pitch classes formed from these angles (Bello, 2018).

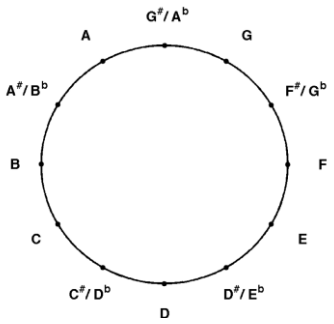


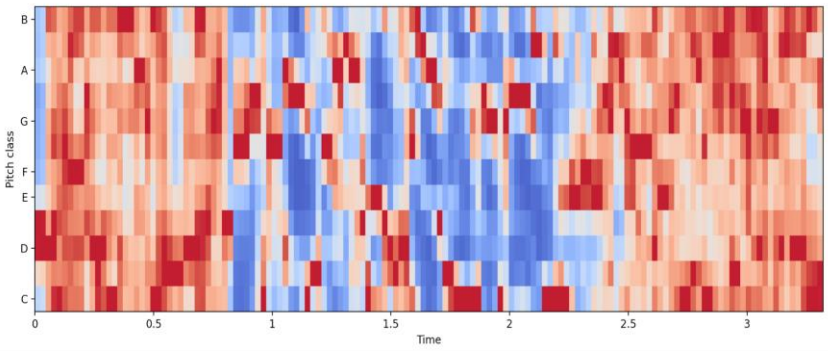

Figure 2.2.31: Pitch classes

The chroma is computed by applying a filter bank to a time-frequency representation of the audio. It shows how much energy of each pitch class is present in the signal. Chromagram comprises a time-series of chroma vectors (Korzeniowski & Widmer, 2016). It is calculated by a Short Time Frequency Transform or a Constant-q Transform. The signal transforms to the frequency domain from the time domain with these methods (Khadkevich & Omologo, 2011). In this study, STFT is used for calculation. Chromatograms indicate the harmonic structure of a short time window of the signal. Musical pitch perception and automatic chord recognition utilizes Chroma widely (Kattel et al., 2019).

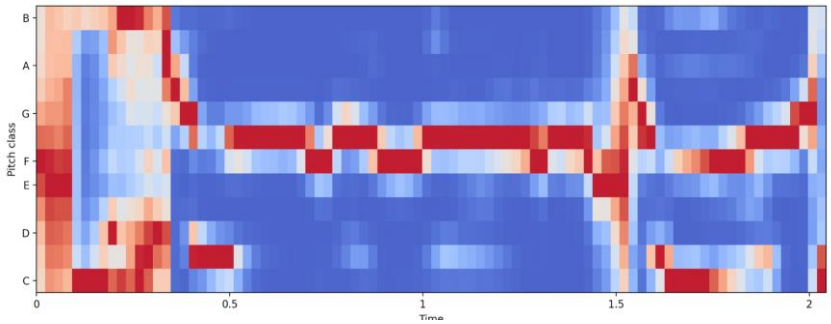
### 2.4.7 Short time fourier transform

Short Time Fourier Transform is a technique that applies on non-stationary signals and provides information in frequency components that varies over time. Fourier transform indicates frequency and hides time information. Frequencies visualized as spectrograms showing magnitude of the short-time Fourier transform for a signal. Selecting the time value for windowing changes the resolution of the frequency domain. For better resolution of frequency, a wide window should be used (Kehtarnavaz, 2008).

(a)




(b)




(c)

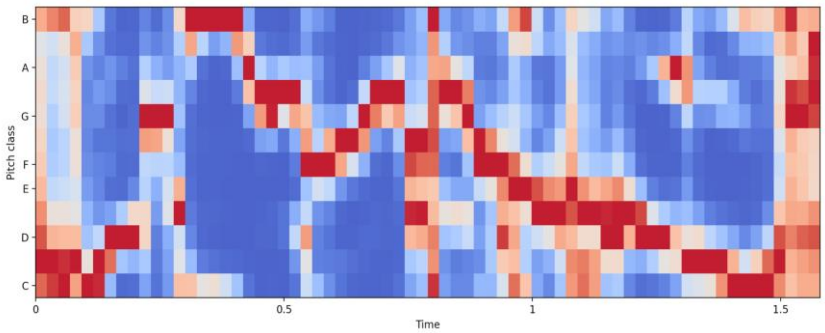
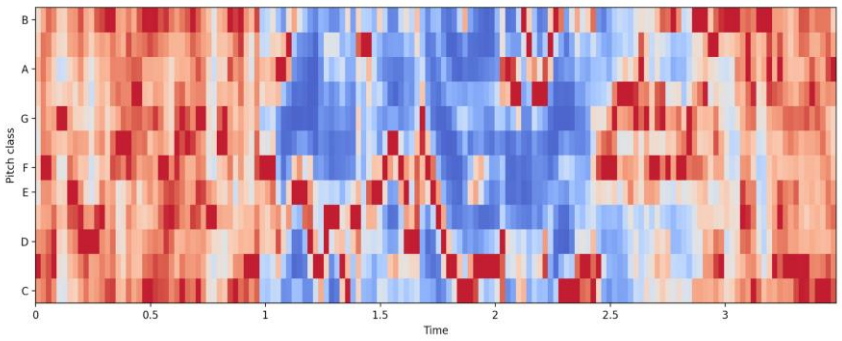
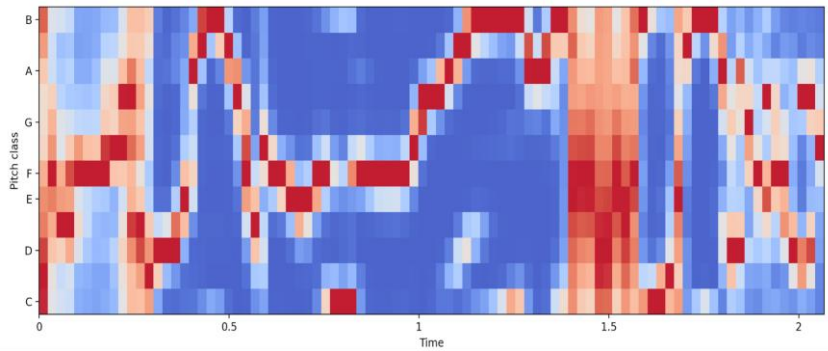



Figure 2.2.32: Chroma's are given in figures and the X-axis shows Time (sec) and the Y-axis shows Pitch Class. Figures in 'a' to 'c' belong to the emotion of neutral. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

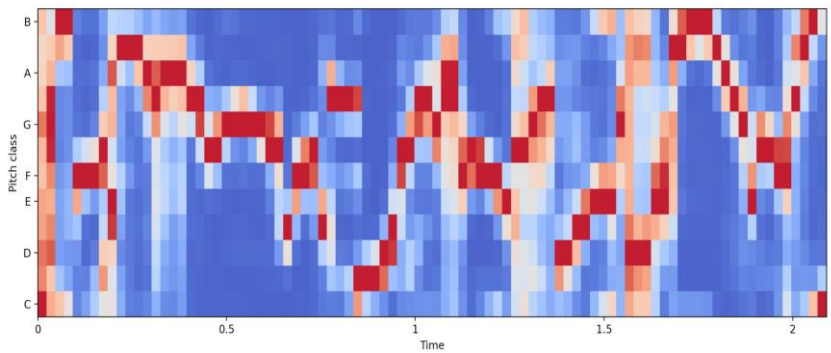
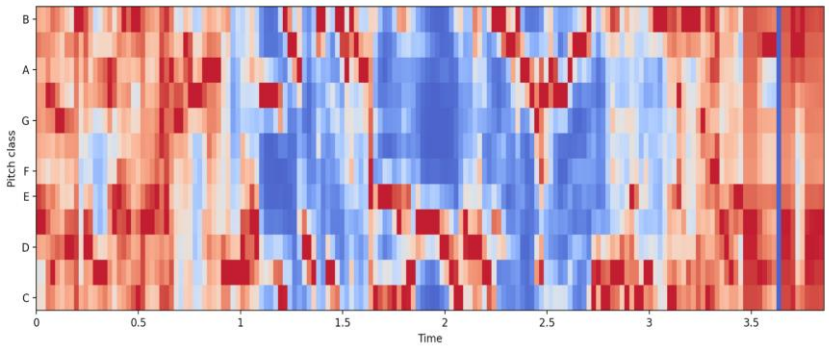
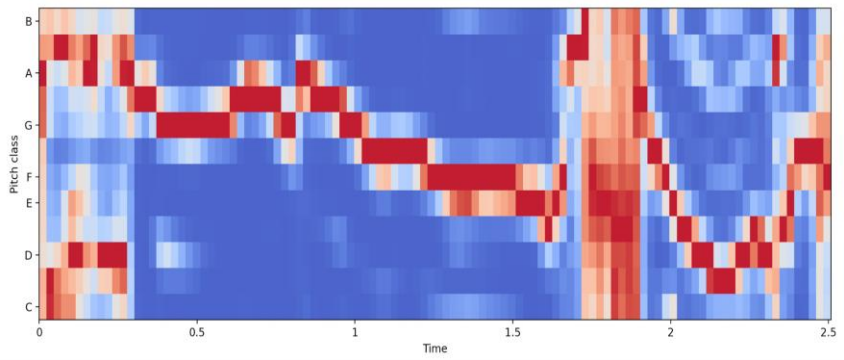



Figure 2.2.33: Chroma's are given in figures and the X-axis shows Time (sec) and the Y-axis shows Pitch Class. Figures in 'a' to 'c' belong to the emotion of happiness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

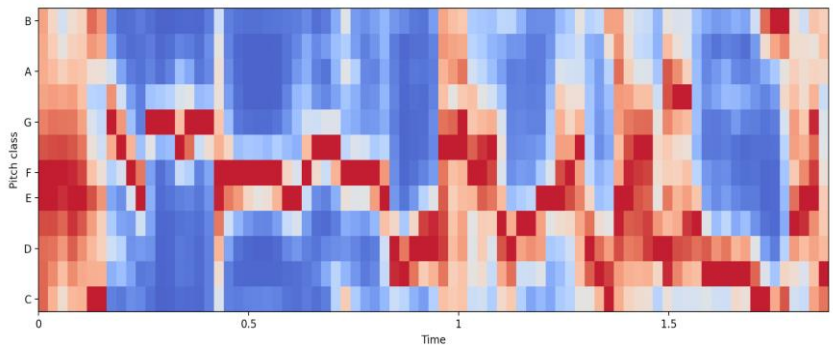
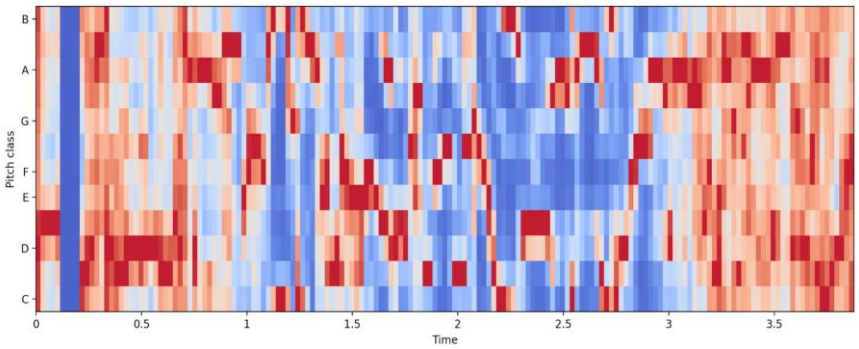
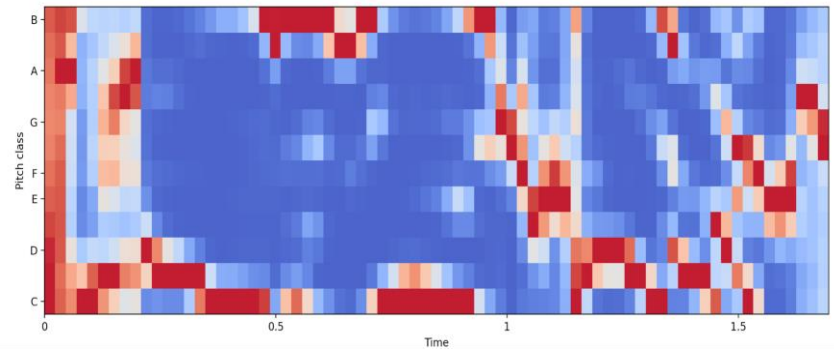



Figure 2.2.34: Chroma's are given in figures and the X-axis shows Time (sec) and the Y-axis shows Pitch Class. Figures in 'a' to 'c' belong to the emotion of sadness. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

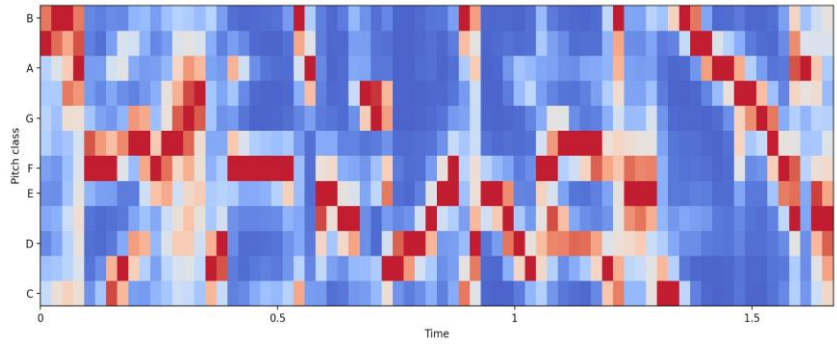
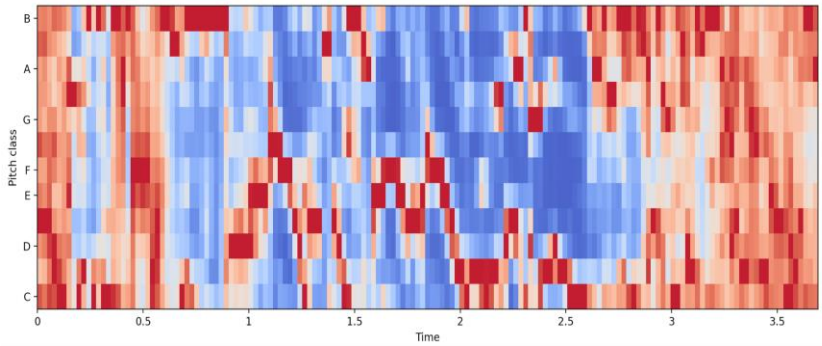
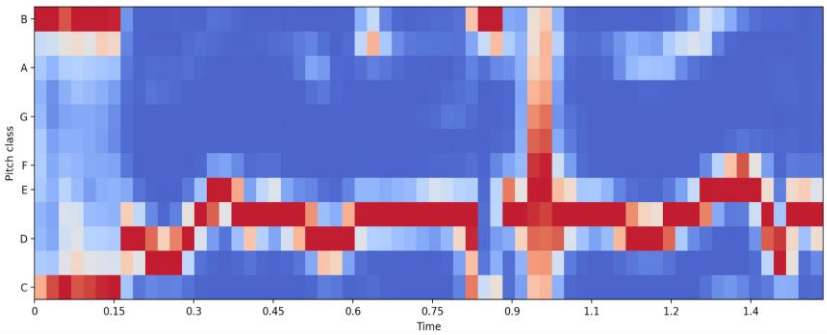



Figure 2.2.35: Chroma's are given in figures and the X-axis shows Time (sec) and the Y-axis shows Pitch Class. Figures in 'a' to 'c' belong to the emotion of anger. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

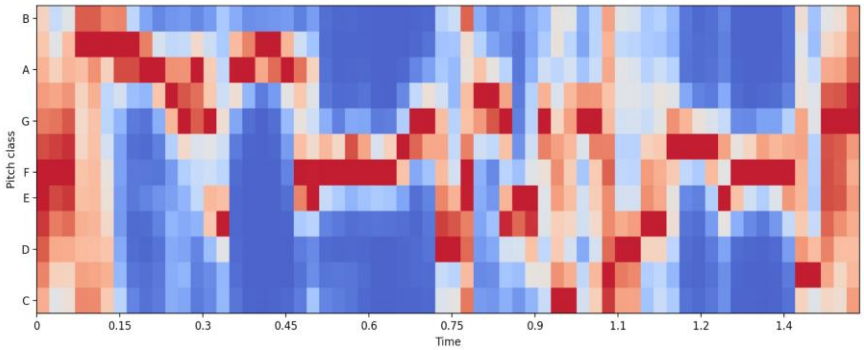
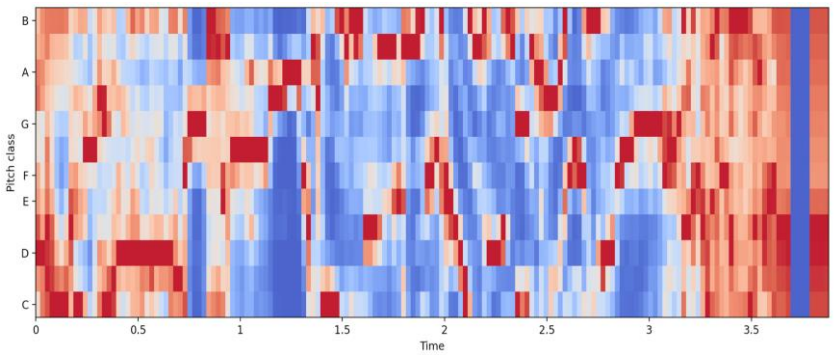
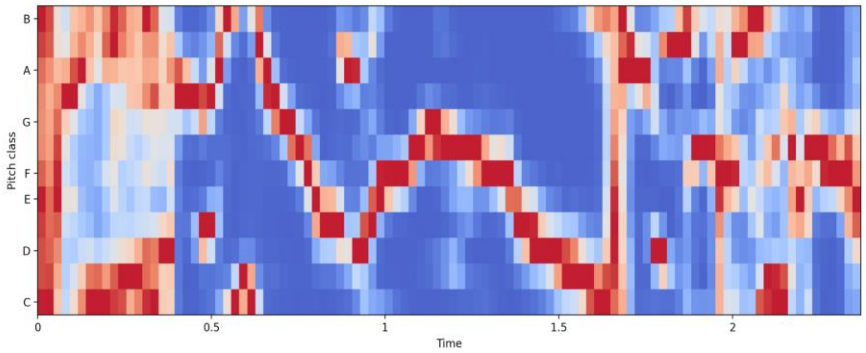



Figure 2.2.36: Chroma's are given in figures and the X-axis shows Time (sec) and the Y-axis shows Pitch Class. Figures in 'a' to 'c' belong to the emotion of fear. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

(a)




(b)




(c)

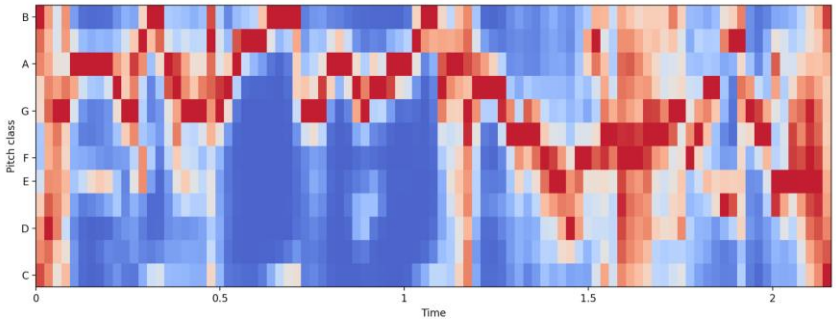



Figure 2.2.37: Chroma's are given in figures and the X-axis shows Time (sec) and the Y-axis shows Pitch Class. Figures in 'a' to 'c' belong to the emotion of disgust. In (a) record is taken from speaker 1 from RAVDESS which represents "Kids are talking by the door" utterance. In (b) record is taken from OAF actress from the TESS which represents "Say the world dog" utterance. In (c) record is taken from speaker 13 from EMODB for "Das will sie am Mittwoch abgeben" utterance which means "She will hand it in on Wednesday".

## 2.5 Unsupervised Pre-Trained Model

Preparing a labelled speech dataset is both time consuming and also requires considerable resources. An accomplished speech model requires a lot of labelled data whereas there are limited datasets for speech related research. That's why, using unlabelled speech data gives various advantages when creating models.

An unlabelled pre-training model named wav2vec is developed. The model is based on raw audio representations for speech recognition. It uses a convolutional neural network to compute representation of raw audios and aims to improve supervised speech recognition. In this study, wav2vec Large pre-trained model is performed on emotional speech datasets for obtaining model inputs. Below, there is an illustration for the feature extraction process with a pre-trained model:

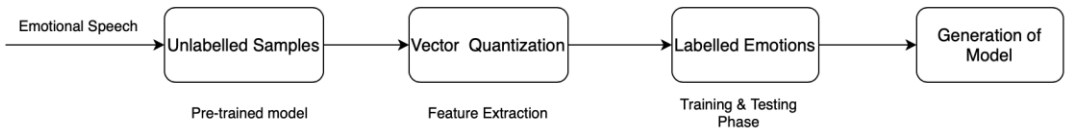



Figure 2.2.38: Block diagram of the proposed system for pre-trained model

### 2.5.1 Architecture of wav2vec Model

In the model there are two networks namely encoder and context networks. There is a figure represents architecture of pre-training model bellowing (Schneider et al., 2019).

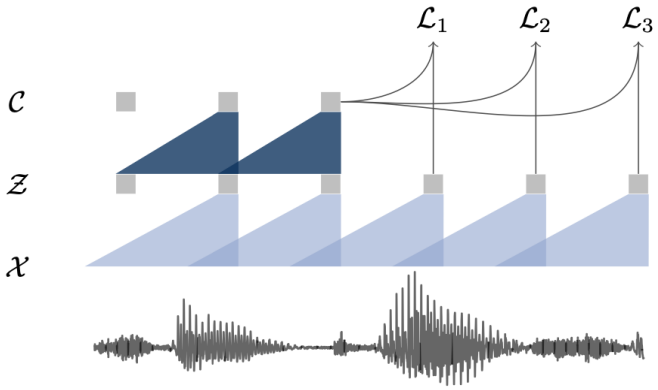



Figure 2.2.39: Architecture of wav2vec pre-trained model

$x_i$, which is an element of X, represents raw audio samples. Encoder network applies on X and low frequency feature representation $z_i$ which is an element of Z is obtained. Encoder layer involves a five-layer convolutional network, its kernel sizes are (10, 8, 4, 4, 4) and strides are (5, 4, 2, 2, 2). Z consists of about 30 ms of 16 kHz encoded audio and $z_i$ strides every 10ms. After this step, the context network applied on Z and C a single contextualized tensor with v field size is obtained

$$ci = g(zi \ldots zi - v) \tag{2.2}$$

g: Z → C

g network has nine layers with kernel size three. The encoder and context networks consist of 512 channels and a ReLU nonlinearity. Both feature and temporal dimension for each sample is normalized with a single normalization group. wav2vec Large model which is used in this study includes two additional linear transformations in the encoder and a larger context network created for larger datasets (Schneider et al., 2019).

Lian et al. (2019) utilize an unsupervised learning approach, namely Future Observation Prediction (FOP) in their study. They combined FOP with Fine-tuning and Hypercolumns and focused on long-term dynamic dependency in order to improve the performance of speech emotion recognition. The FOP Model relies on a masked multi-head self-attention mechanism to capture the long-term dynamic dependencies.

Center for Speech Technology Voice Cloning Toolkit and IEMOCAP datasets are used in their research. In the fine-tuning approach, there are three settings, for the first and second settings the model is applied separately to two datasest and for the third setting the model used without pre-training. Results show for fine tuning that the IEMOCAP set is better than the Center for Speech Technology Voice Cloning Toolkit set whereas the Center for Speech Technology Voice Cloning Toolkit set is better than the no pre-training set. In the Hypercolumns approach, mel-spectrograms which are extracted from waveforms are accepted as baseline because the FOP Model is not used for this step. The FOP Model has 2 layers of which the outputs of each layer add and concatenate their representation and together they get different combinations of features. In the research three classification algorithms are used: Support Vector Machine (SVM), Random Forest (RF) and Attention-Based LSTM. Results show that the mel-spectrogram has the lowest performance with 62.57% weighted accuracy. Random Forest and concatting the output representation have the highest performance for prediction with 65.54 % weighted accuracy A-LSTM algorithm.

To cope with scarce database problems for speech emotion recognition, Deng et al. (2014) used different corporas for train and test sets. Utilizing different corporas for test and train sets is providing decreasing performance. They use the Shared Hidden Layer Autoencoder (SHLA) method to solve scarce resource problems. This method learns common feature representations between the train and test sets in an unsupervised way. After this, extracted features are fed into supervised algorithms. SHLA uses the same parameters for both input and hidden layers but uses distanced parameters for the reconstruction train and test sets. Overall, SHLA tries to minimize this reconstruction error. In the research Airplane Behaviour Corpus as well as Speech Under Simulated and Acted Stress sets are used for training. For the testing phase, Deng et al used the FAU Aibo Emotion Corpus set. The following approaches are used for representation learning: Matched Training, Cross Training, Kernel Mean Matching, Denoising Autoencoder and SHLA and SVM is used as a classifier. In experiments, they used the FAU AEC set for training. Results indicate that the Cross Training has lowest performance for both Airplane Behaviour Corpus dataset with a 55.28% unweighted average recall and the Speech Under Simulated and Actual Stress dataset with 57.32 % unweighted average recall datasets, whereas SHLA has the best

performance for Airplane Behaviour Corpus 63.36 % with an unweighted average and Speech Under Simulated and Actual Stress 62.72%

Xia et al, used denoising autoencoders which they customized. In their approach input is mapped to two hidden representations. One of them is for emotional information and the other is for redundant information. In the first stage, weight and bias parameters that are used for neutral hidden representation are trained. In this step for avoiding overfitting, they use the Wall Street journal corpus which is different from the training dataset. In the second stage, two projections are created. Parameters that are obtained from the previous stage are used for corrupted input data and expecting to gain neutral information for one of the representations. Random parameters initialized and updated during pre-training in the other projection. Then, neural and emotional hidden representations are decoded for creating reconstructed representations. Two reconstructed representations are combined for the final reconstructed representation. The IEMOCAP database and SVM classifier are used in the research. Results show that the worst performance belongs to traditional Denoising Auto-encoder with 57.2 % average accuracy using 200 nodes and the best performance is obtained with proposed Denoising Auto-encoder with 61.5 % average accuracy with 800 nodes (Xia & Liu, 2013). In the following table, there are results for trials:

Table 2.2.1: Unsupervised learning results

| Approach | Accuracy (%) | Database | Accuracy Type | Classifier |
|---|---|---|---|---|
| Future Observation Prediction | 65.45 | - | Weighted Accuracy | Multi-head Self-Attention |
| Future Observation Pred+Fine-Tuning | 65.03 | IEMOCAP | Weighted Accuracy | Multi-head Self-Attention |
| Future Observation Pred+Hypercolumns | 63.56 | IEMOCAP | Weighted Accuracy | A-LSTM |

Table 2.1 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Matched Training | 62.41 | Speech Under Simulated and Actual Stress | Unweighted Average Recall | SVM |
| Cross Training | 57.32 | Speech Under Simulated and Actual Stress | Unweighted Average Recall | SVM |
| Kernel mean matching | 62.52 | Airplane Behavior Corpus | Unweighted Average Recall | SVM |
| Denoising Auto-encoder | 62.08 | Speech Under Simulated and Actual Stress | Unweighted Average Recall | SVM |
| Shared Hidden Layer Autoencoder | 63.36 | Airplane Behavior Corpus | Unweighted Average Recall | SVM |
| Denoising Auto-encoder + Traditional | 60.1 | IEMOCAP | Accuracy | SVM |
| Denoising Auto-encoder + Proposed | 61.5 | IEMOCAP | Accuracy | SVM |

### 2.6 Autoencoder

An autoencoder is a neural network that compresses input and reconstructs it as output. It uses a representation learning approach and tries to give output with the same as the input. There are three parts of an autoencoder, an encoder, a compressed representation also called code and a decoder. Encoder function converts inputs to code and a decoder function generates reconstruction of the input using this code. Main goal is learning informative representation of the raw data in an unsupervised way. If input features have a relation between them more successful outputs are obtained.

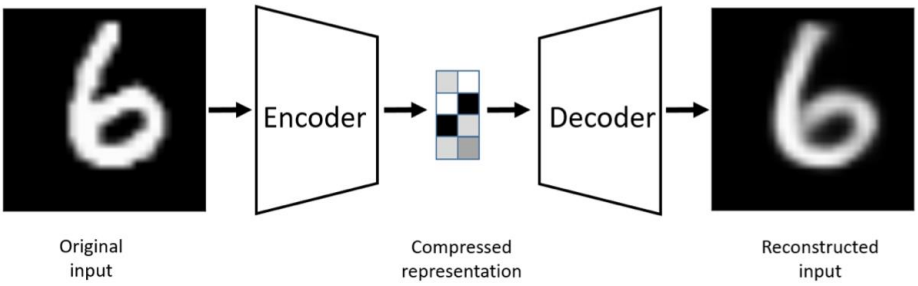



Figure 2.2.40: Representation for an autoencoder (Bank et al., 2020)

The approach tries to satisfy following formula:

$$argmin\ A, B\ E[\Delta(x, B \circ A(x)] \tag{2,3}$$

$A$: R$n$ → R$p$, A stands for encoder

$B$: R$p$ → R$n$, B stands for decoder,

E: Expectations over the distribution of $x$

Δ: Reconstruction loss function

There are various types of autoencoders e.g., Variational Autoencoders, Denoising Autoencoders. They can be used for compression, feature extraction, dimensionality reduction, classification or clustering (Bank et al., 2020).

### 2.7 Classification

Classification types of machine learning are explained in the following.

#### 2.7.1 Support vector machine

Support Vector Machines (Vapnik, 1982) is a supervised learning algorithm based on statistical learning theory. The method suggested for classification problems where patterns between variables are unknown. At first approximation, it was designed

to classify linear data of two classes later developed for nonlinear data of multi classes. The purpose of SVM is maximizing distance between support vectors belonging to different classes to obtain the optimal separation line. SVM is basically divided into two subcategories whether the dataset can be separated linearly or not.

#### 2.7.1.1 Linear svm

Suppose that a dataset of two classes is given and try a new data to group for one class. The problem can be formalized as (Scholkopf & Smola, 2018):

$$(x_1, y1), \ldots . (xm, ym) \subseteq X \; x \left\{ \text{-1,1} \right\} \tag{2,4}$$

$x_i$: instances
$y_i$: labels
X: nonempty patterns $x_i$

$y_i \in X \; x \left\{ \text{-1,1} \right\}$

This form is the simplest form and refers to binary classification. This situation can be figured below:

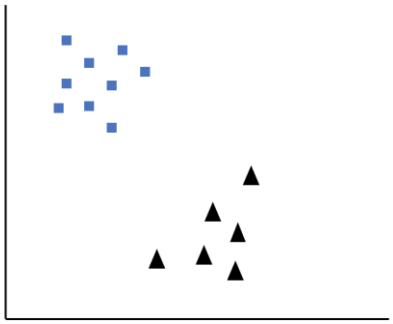


Figure 2.2.41: Linear data example

In the figure, there are two classes; first one is blue rectangles (positives) and second is black triangles (negatives). There are infinite lines to separate the classes, SVM tries to find more generalized line that provides maximum distance between data points of both classes, an example line can be shown as below:

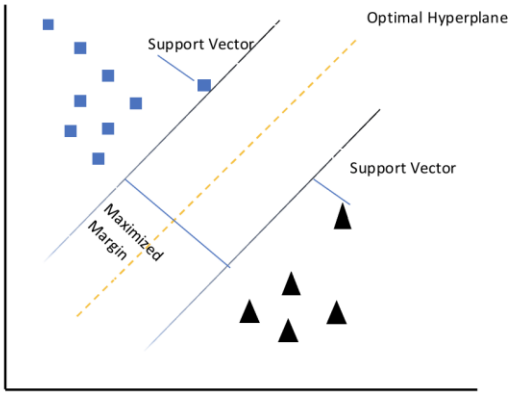



Figure 2.2.42: Separated data with Support Vectors

Margin: The distance between line and support vectors

Hyperplane: Decision boundary for classify the data points and it maximizes the margin

Support vector: Data points which are closest to the hyperplane

SVM's goal is maximizing margin and it tries to make a decision boundary which makes the margin as wide as possible. The hyperplane that has the maximum margin can be called an optimal hyperplane. Model's success has a strong relation with margin.

#### 2.7.1.2 Nonlinear (Kernel) svm

Real life problems cannot be solved with binary classification. That's why, SVM is developed for nonlinear datasets. When a dataset cannot separate with a line, it can be transformed to separable data in higher dimension by adding more axes where the training set is separable. SVM and some other linear classifiers use Kernel Trick for mapping higher dimensional space.

### 2.7.1.3 Kernel trick

Nonlinear datasets can be classified by adding extra dimensions to become linearly separable. Mapping a higher dimensional is known as “Kernel Trick”. Kernels are similarity functions that return dot products for data points. The data does not map feature space actually, instead a kernel function is defined and the dot produce <x.z> is replaced with a kernel function K (x.z) for both training and testing data

K: X x X → R

$$K(x \rightarrow, z \rightarrow) = <\Phi(x \rightarrow), \Phi(z \rightarrow)> \quad (2,5)$$

Kernels can be computed easily and efficiently. Polynomial, Radial Basis Function (RBF) and Sigmoid functions are the most common Kernel functions.

## 2.8 Artificial Neural Network

Artificial Neural Network (ANN) is designed to simulate functioning of the human brain. Basically, neurons which are individual cells in the brain are fundamental parts that are sampled in Artificial Intelligence. A biological neuron can be shown as Figure 2.43 (Hamdi et al., 2016).

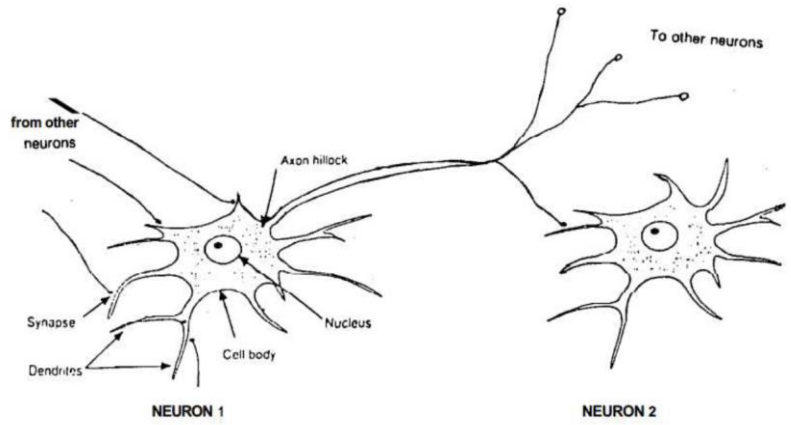



Figure 2.2.43: A biological neuron

The human brain consists of billions of neurons. Information is processed by the connections that exist between neurons. Dendrites collect inputs (signals) from other neurons and if the sum of inputs is great enough, an electrical impulse occurs and it is sent from axon to boutons. Boutons network with other neurons via connections called synapses. Our brains have the ability to process information thanks to synaptic connections.

The Neural Network's foundation is 'perceptron'. A perceptron includes four different parts:

1. Input layer

2. Weight and Bias

3. Net Sum

4. Activation function

An artificial neuron can be shown as Figure 2.44 (Hamdi et al., 2016).

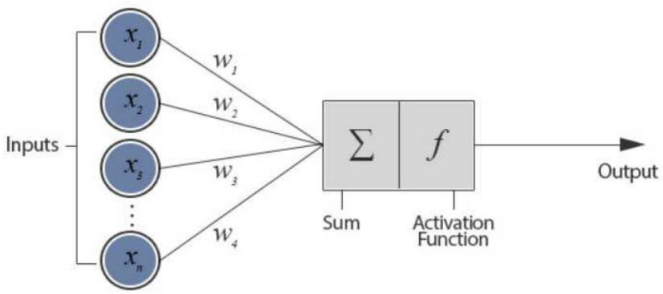



Figure 2.2.44: A perceptron example

A standard perceptron has many inputs and these inputs are all individually weighted. Each input is multiplied by its weight, by this way weighted inputs are created. Then the weighted inputs are summed up. The sum result changes from zero to infinite, if sum equals to zero, bias is added to change the result. On the other hand, a threshold value is set up to limit the sum so as not to get an infinite result. Activation function is used for limiting the response.

### 2.8.1 Single Perceptron Layer

Simple Perceptron consist of three parts:

1. Layer of input neuron

2. Layer of output neuron

3. Single layer of weights between input and output neurons

A schematic representation for simple perceptron is given in Figure 2.45 (Noriega, 2005).

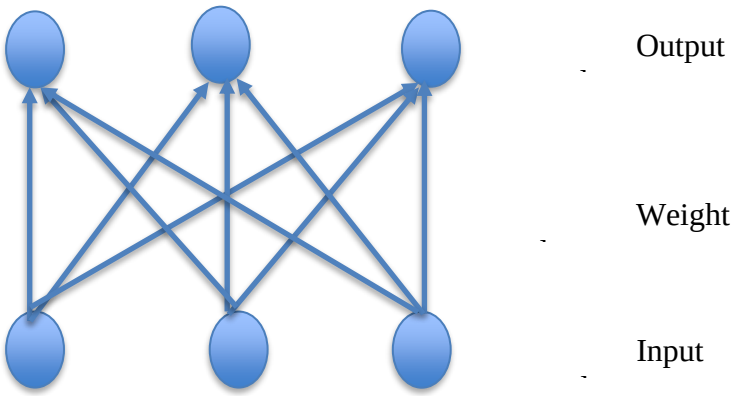



Figure 2.2.45: A simple perceptron architecture

Mathematically, input and output layers can be represented as vectors of values and the weight (w) can be represented as a matrix. The matrix is a grid that has the number of input (i) and output (o) nodes. Network output is a function that has a relation with input and matrix:

$$O = f(IWio) \tag{2,6}$$

### 2.8.2 Multi-Layer Perceptron

In addition to the single layer perceptron, the multi-layer perceptron contains a hidden layer. Hidden layer locates between input and output layers and its output is connected to the inputs of other neurons. Multi-layer perceptrons are capable of approximating nonlinear functions. There is not a certain count for hidden layers, it is adaptable according to the problem. An example multi-layer perceptron is given in Figure 2.46 (Hamdi et al., 2016).

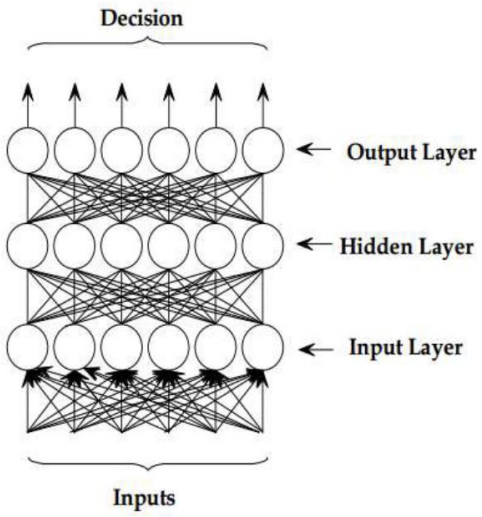



Figure 2.2.46: A multi-layer perceptron architecture

As seen in Figure 2.46, MLP classifier based on feedforward artificial neural network. Each node is fully connected in a feed-forward way. There can be any number of nodes per layer and there are usually multiple hidden layers to solve complex problems.

Updating weights and biases for finding the correct values is called a learning rule. At the beginning of the process weights are set to random numbers between -1 and +1. The amount of changes during each step size is called the learning rate.

Determining the output of a neural network, activation functions are used. The function takes place in each neuron and decides whether it should be activated or not, related to whether input is relevant 71ort he model's prediction. As activation function

is calculated for each data sample, so as not to cause performance problems, it must be efficient.

### 2.9 Convolutional Neural Networks

Convolutional Neural Network (CNN) is a type of artificial neural network that applies mostly image recognition and processing. CNN consists of two main parts:

1. Feature extractors
2. Classifier

Alom et al figured overall architecture of CNN as beloved Figure 2.47 (Alom et al., 2019).

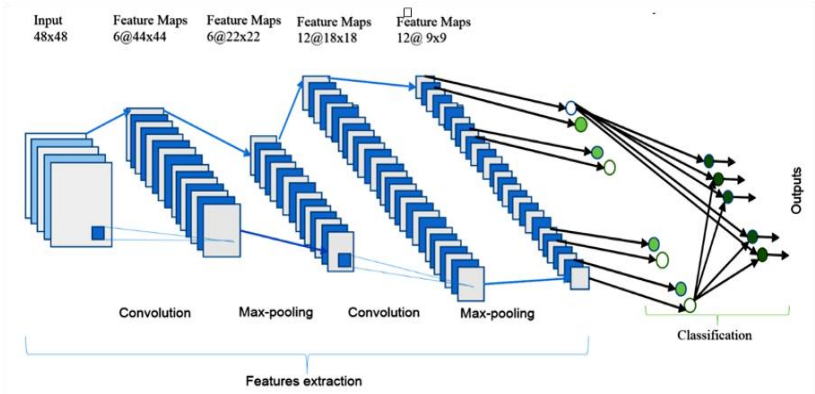



Figure 2.2.47: A CNN layers example

Feature extraction phase includes Convolution and Max Pooling layers. Convolution Layer is the core building block layer of this algorithm.

#### 2.9.1 Convolutional layer

Simply, a convolution uses a filter, which is smaller than input data, to an input and applies convolution operations to extracting features. Convolution operation involves multiplication of weights with the inputs. Filter also called as kernel, forms a set of weights and the same filter implements the entire input by sliding. At each location element based

matrix multiplication is done and the result is summed up. At the end of filtering a map of activations called a feature map is obtained. The following figures show obtaining a feature map:

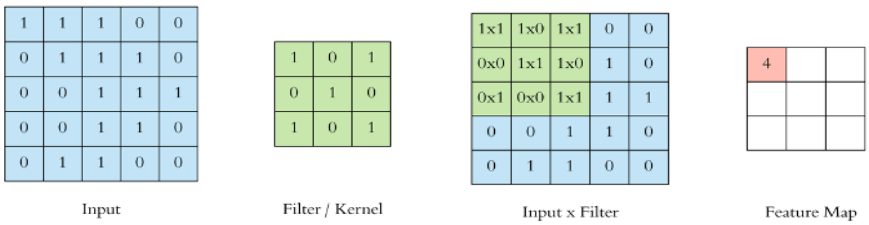



Figure 2.2.48: A convolutional layer and feature map example (Dertat, 2017).

Due to the shape of the filter, this is a 3X3 convolution and the receptive field is also 3x3 size. Filter is applied to overlapping parts of the input from left to right and from top to bottom. This gives discovering feature opportunities that take place anywhere in the input. Multiple convolutions can be applied on an input and distinct feature maps can be obtained by using different filters.

In Convolutional Layer the result of the convolution operation goes through generally non-linear activation functions such as Sigmoid, Hyperbolic, Tangent functions to form feature maps. In addition, bias is used for each output map. Backpropagation technique can be used to increase computational strain on the activation function, and its derivative function. Backpropagation performs a backward pass for tuning model's weights and biases after each forward pass through a network. The goal is optimizing the weights to correctly mapping inputs and outputs.

### 2.9.2 Sub-Sampling layer

Pooling is the other basic operation in CNN. It is used to reduce the dimensionality of feature maps while keeping the important information after a convolution operation. Pooling enables extracting dominant features and shortens the training time.

Max Pooling and Average Pooling are two types of operations that are mostly performed. Max Pooling returns the maximum value from the portion of the feature map. Average Pooling selects the average value from the portion of the feature map.

### 2.9.3 Classification- Fully connected layer (FC Layer)

After the convolution and pooling layers, a feed-forward neural layer called a fully connected layer is used for classification. It computes the score of each class from the extracted features by using Soft-Max classification technique. Feature maps are flattened into a column vector with scalar values. Backpropagation with gradient descent applies to every iteration of training and Rectified Linear Unit (Relu) is the common choice for the activation function.

# 3. RESULTS

## 3.1 Dataset

Recordings in the emotional speech datasets can be created from natural or simulated emotions. Within the study, acted databases in English and German languages are preferred namely, RAVDESS, TESS and EMO-DB which are most common and publicly available. These three datasets are combined into one dataset to obtain a larger set and the combined set is used for implementation.

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Database contains 7356 files, 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent. Both speech and song files exist in the dataset. Speech includes eight emotions which are Calm, Happy, Sad, Angry, Fearful, Surprise and Disgust expressions, and the song contains Calm, Happy, Sad, Angry, and Fearful emotions. In the study only speech files are used (Livingstone & Russo, 2018).

Toronto Emotional Speech Set (TESS) contains a set of 200 target utterances in English and voiced by two age groups 26 for Young and 64 for Old. Recordings include seven emotions which are Anger, Disgust, Fear, Happiness, Pleasant, Surprise, Sadness and Neutral and there are 2800 files in total (Pichora-Fuller & Dupuis, 2020).

Berlin Database of Emotional Speech (EMO-DB) contains 5 female and 5 male actors whose ages are between 21 and 35 simulated emotions for 5 short and 5 longer sentences. These ten German utterances are similar with daily communication and interpretable in all selected emotions. There are also some second versions in the database and it consists of 800 sentences including seven emotions which are Anger, Boredom, Disgust, Fear, Happiness and Sadness (Burkhardt et al., 2005).

There are more details about databases in the following tables.

Table 3.3.1: Durations and total counts of the used datasets

| Dataset | Total Duration (min) | Average Duration (sec) | Total Samples |
|---|---|---|---|
| Ravdess | 89 | 3.7 | 1440 |
| Tess | 96 | 2.05 | 2800 |
| EmoDB | 25 | 2.8 | 535 |

Table 3.3.2: RAVDESS database record counts and durations based on emotions

| Emotion | Count of Files | Total Duration (min) |
|---|---|---|
| Neutral | 96 | 5,6 |
| Calm | 192 | 12,15 |
| Happy | 192 | 11,6 |
| Sad | 192 | 11,8 |
| Angry | 192 | 12,39 |
| Fearful | 192 | 11,4 |
| Disgust | 192 | 12,6 |
| Surprised | 192 | 11,16 |

Table 3.3.3: TESS database record counts and durations based on emotions

| Emotion | Count of Files | Total Duration (min) |
|---|---|---|
| Neutral | 400 | 13,7 |
| Happy | 400 | 13,15 |
| Sad | 400 | 16,01 |
| Angry | 400 | 12,28 |
| Fearful | 400 | 11,05 |
| Disgust | 400 | 16,32 |
| Surprised | 400 | 13,39 |

Table 3.3.4: EMO-DB database record counts and durations based on emotions

| Emotion | Count of Files | Total Duration (min) |
|---|---|---|
| Angry | 127 | 5,59 |
| Boredom | 81 | 3,8 |
| Disgust | 46 | 2,57 |
| Fearful | 69 | 2,57 |
| Happy | 71 | 3,01 |
| Sad | 62 | 4,19 |
| Neutral | 79 | 3,1 |

Table 3.3.5: Total RAVDESS, TESS and EMO-DB databases record counts and durations based on emotions

| Emotion | Count of Files | Total Duration (min) | Average Duration (sec) |
|---|---|---|---|
| Happy | 663 | 27,9 | 2,5 |
| Calm | 192 | 12,2 | 3,8 |
| Surprise | 592 | 24,5 | 2,5 |
| Neutral | 575 | 22,5 | 2,34 |
| Sad | 654 | 31 | 2,9 |
| Fear | 661 | 25,1 | 2,3 |
| Disgust | 638 | 31,5 | 3 |
| Boredom | 81 | 3,8 | 2,7 |
| Anger | 719 | 31,5 | 2,7 |

### 3.2 Implementation Details

In the study, nine emotions namely, Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, Boredom and Surprise are classified. The aforementioned databases TESS, EMO-DB, RAVDESS are combined to obtain a one larger dataset. Some of the emotions are common at three datasets whereas some of them are not. That's why, the number of audio files for each emotion differs in the final dataset. This situation causes disadvantages for emotions which have less number of files. However, using more than one dataset increases the sample count and provides robust training. Totally, 4775 audio files are handled; 3581 files implemented for training and 1194 files used for testing phases.

As mentioned before SVM, MLP and CNN techniques are performed for classification. MFCC, Chroma, ZCR and RMSE features used as traditional features

and alternatively an unsupervised pre-trained model wav2vec Large is used for obtaining representations which is relatively a new technique.

All audio files are resampled at the 16000 sampling rate. Before RMSE calculation, audio signals' Fourier transform computed by the STFT to get a more accurate representation of energy over time. Then the signals represented by the STFT with frame length 1024 were given as input for RMSE calculation and its mean value is taken. For Zero Crossing Rate calculation, audio signals with 1024 frame length are used and their mean values are taken. For MFCC calculation 40 coefficients are taken into consideration. Similar with RMSE, STFT form is used for Chroma calculation and its mean value is taken for classification

SVM algorithm finds the best hyperplane that maximizes margins for categorization. Support Vector Classification class from Sklearn library is used for SVM models. Kernels are set as Radial Basis Function (RBF), regulation parameter C chosen as 1, the size of Kernel cache is set as 700.

MLP is a relatively simple form of artificial neural networks. MLPClassifier class from Sklearn library is used. For Multi-Layer Perceptron models, 300 hidden units are performed, adaptive is chosen as learning rate parameter, the Relu function used as activation function and maximum iteration number is set as 500.

CNN utilizes layers to understand patterns and it is a quite popular deep learning architecture. In the study, the Sequential form from Keras is used for the CNN model. Model consists of two convolutional layers, two pooling layers, a fully connected layer and a Softmax layer. Relu acts as an activation function. To avoid especially overfitting problems dropout employed in each layer. Also, RMSprop optimizer is implemented for CNN models.

### 3.3 Test Results

#### 3.3.1 Speech emotion recognition using time domain features

Several experiments on the combined dataset are run for time-based features RMSE and ZCR. In order to test the algorithm, %25 of data separated for testing phase and %75 used for training phase. In all experiments, models with the time-based features get the lowest degrees. Table 3.6 shows the methods and their corresponding accuracy and weighted accuracy results on the test data. It seems from the results that the role of RMSE and Zero Crossing Rate in speech emotion recognition is not determinative regardless of the algorithm.

Table 3.3.6: Model classification results for time-based features

| Method | Features | Accuracy | Weighted Accuracy |
|---|---|---|---|
| SVM | RMSE & ZCR | 0.20 | 0.26 |
| MLP | RMSE & ZCR | 0.20 | 0.27 |
| CNN | RMSE & ZCR | 0.21 | 0.14 |

Table 3.3.7: SVM Classification Report for Time Based Features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.41 | 0.17 | 0.24 | 421 |
| Boredom | 0.00 | 0.00 | 0.00 | 0 |
| Calm | 0.00 | 0.00 | 0.00 | 0 |
| Disgust | 0.12 | 0.19 | 0.15 | 102 |
| Fearful | 0.04 | 0.22 | 0.07 | 32 |

Table 3.7 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Happy | 0.42 | 0.18 | 0.26 | 375 |
| Neutral | 0.00 | 0.00 | 0.00 | 144 |
| Sad | 0.45 | 0.28 | 0.35 | 264 |
| Surprised | 0.00 | 0.00 | 0.00 | 0 |
| Accuracy | | | 0.20 | 1194 |
| Macro Avg | 0.16 | 0.12 | 0.12 | 1194 |
| Weighted Avg | 0.38 | 0.20 | 0.26 | 1194 |

Table 3.3.8: MLP Classification Report for Time Based Features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.45 | 0.19 | 0.27 | 420 |
| Boredom | 0.00 | 0.00 | 0.00 | 0 |
| Calm | 0.00 | 0.00 | 0.00 | 0 |
| Disgust | 0.00 | 0.00 | 0.00 | 0 |
| Fearful | 0.00 | 0.00 | 0.00 | 0 |
| Happy | 0.50 | 0.19 | 0.27 | 447 |
| Neutral | 0.10 | 0.13 | 0.11 | 109 |
| Sad | 0.38 | 0.29 | 0.33 | 218 |
| Surprised | 0.00 | 0.00 | 0.00 | 0 |
| Accuracy | | | 0.20 | 1194 |
| Macro Avg | 0.16 | 0.09 | 0.11 | 1194 |

Table 3.8 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Weighted Avg | 0.42 | 0.20 | 0.27 | 1194 |

Table 3.3.9: CNN Classification Report for Time Based Features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.20 | 0.45 | 0.27 | 180 |
| Boredom | 0.00 | 0.00 | 0.00 | 20 |
| Calm | 0.00 | 0.00 | 0.00 | 48 |
| Disgust | 0.00 | 0.00 | 0.00 | 159 |
| Fearful | 0.00 | 0.00 | 0.00 | 165 |
| Happy | 0.18 | 0.52 | 0.27 | 166 |
| Neutral | 0.17 | 0.14 | 0.15 | 144 |
| Sad | 0.32 | 0.37 | 0.35 | 164 |
| Surprised | 0.00 | 0.00 | 0.00 | 148 |
| Accuracy | | | 0.21 | 1194 |
| Macro Avg | 0.10 | 0.16 | 0.12 | 1194 |
| Weighted Avg | 0.12 | 0.21 | 0.14 | 1194 |

The tables above show the classification reports of audio time-based features with SVM, CNN and MLP Classifiers. As can be seen

- from Table 3.7, Happy and Sad emotions have the higher f1 scores
- from Table 3.8 Happy, Sad and Angry emotions have the higher f1 scores
- from Table 3.9 Happy, Sad and Angry emotions have the higher f1 scores

Classification results demonstrate that RMSE and ZCR features are good at predicting Happy and Sad emotions. On the other hand, the weighted accuracy ratio of Happy and Sad emotions' f1 scores are quite low for an acceptable recognition rate. Consequently, it can be said that RMSE and ZCR time-based features are insufficient for speech emotion recognition.

### 3.3.2 Speech emotion recognition using frequency domain features

Several experiments on the combined dataset are run for frequency-based feature Chroma. In order to test the algorithm, %25 of data separated for testing phase and %75 used for training phase. Table 3.10 shows the methods and their corresponding accuracy and weighted accuracy results on the test data. It seems from the results that the recognition rates with Chroma feature are similar to each other for whole algorithms

Table 3.3.10: Model classification results for chroma feature

| Method | Features | Accuracy | Weighted Average |
|---|---|---|---|
| SVM | Chroma | 0.54 | 0.55 |
| MLP | Chroma | 0.54 | 0.54 |
| CNN | Chroma | 0.56 | 0.57 |

Table 3.3.11: SVM Classification Report for Frequency Based Features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.61 | 0.42 | 0.50 | 261 |
| Boredom | 0.00 | 0.00 | 0.00 | 0 |

Table 3.11 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Calm | 0.00 | 0.00 | 0.08 | 0 |
| Disgust | 0.50 | 0.34 | 0.41 | 234 |
| Fearful | 0.52 | 0.79 | 0.62 | 108 |
| Happy | 0.42 | 0.63 | 0.50 | 109 |
| Neutral | 0.73 | 1.00 | 0.84 | 105 |
| Sad | 0.73 | 0.60 | 0.66 | 199 |
| Surprised | 0.55 | 0.46 | 0.50 | 178 |
| Accuracy | | | 0.54 | 1194 |
| Macro Avg | 0.45 | 0.47 | 0.45 | 1194 |
| Weighted Avg | 0.58 | 0.54 | 0.55 | 1194 |

Table 3.3.12: MLP Classification Report for Frequency Based Features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.54 | 0.53 | 0.54 | 182 |
| Boredom | 0.10 | 0.20 | 0.13 | 10 |
| Calm | 0.00 | 0.00 | 0.00 | 1 |
| Disgust | 0.65 | 0.30 | 0.41 | 347 |
| Fearful | 0.53 | 0.76 | 0.63 | 116 |
| Happy | 0.40 | 0.75 | 0.53 | 89 |
| Neutral | 0.73 | 0.93 | 0.82 | 113 |
| Sad | 0.66 | 0.57 | 0.61 | 191 |

Table 3.12 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Surprised | 0.51 | 0.52 | 0.51 | 145 |
| Accuracy | | | 0.54 | 1194 |
| Macro Avg | 0.46 | 0.51 | 0.46 | 1194 |
| Weighted Avg | 0.59 | 0.54 | 0.54 | 1194 |

Table 3.3.13: CNN Classification Report for Frequency Based Features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.52 | 0.54 | 0.53 | 180 |
| Boredom | 0.23 | 0.15 | 0.18 | 20 |
| Calm | 0.16 | 0.10 | 0.13 | 48 |
| Disgust | 0.36 | 0.60 | 0.45 | 159 |
| Fearful | 0.70 | 0.58 | 0.64 | 165 |
| Happy | 0.66 | 0.45 | 0.53 | 166 |
| Neutral | 0.99 | 0.72 | 0.84 | 144 |
| Sad | 0.73 | 0.66 | 0.69 | 164 |
| Surprised | 0.45 | 0.59 | 0.51 | 148 |
| Accuracy | | | 0.56 | 1194 |
| Macro Avg | 0.53 | 0.49 | 0.50 | 1194 |
| Weighted Avg | 0.60 | 0.56 | 0.57 | 1194 |

The tables above show the classification reports of audio frequency-based features with SVM, CNN and MLP Classifiers. As can be seen

- from Table 3.11, Neutral and Sad emotions have the higher f1 scores
- from Table 3.12 Neutral and Fearful emotions have the higher f1 scores
- from Table 3.13 Neutral and Sad emotions have the higher f1 scores

Classification results demonstrate that Chroma feature is good at predicting Neutral and Sad emotions. Besides, the weighted accuracy ratios for whole algorithms are higher than the ratios belonging to the model with the time-based features.

### 3.3.3 Speech emotion recognition using spectral shape domain features

Several experiments on the combined dataset are run for spectral shape-based feature MFCC. In order to test the algorithm, %25 of data separated for testing phase and %75 used for training phase. Table 3.14 shows the methods and their corresponding accuracy and weighted accuracy results on the test data. It seems from the results that recognition differs from statistic and neural network techniques for MFCC feature

Table 3.3.14: Model classification results for MFCC feature

| Method | Features | Accuracy | Weighted Average |
|---|---|---|---|
| SVM | MFCC | 0.64 | 0.67 |
| MLP | MFCC | 0.81 | 0.81 |
| CNN | MFCC | 0.84 | 0.84 |

Table 3.3.15: SVM Classification Report for spectral shape-based feature

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.86 | 0.63 | 0.73 | 245 |
| Boredom | 0.0 | 0.0 | 0.0 | 1 |
| Calm | 0.85 | 0.31 | 0.46 | 132 |
| Disgust | 0.73 | 0.55 | 0.63 | 210 |
| Fearful | 0.66 | 0.83 | 0.73 | 132 |
| Happy | 0.10 | 1.00 | 0.18 | 16 |
| Neutral | 0.84 | 0.81 | 0.83 | 149 |
| Sad | 0.68 | 0.74 | 0.71 | 152 |
| Surprised | 0.62 | 0.59 | 0.60 | 157 |
| Accuracy | | | 0.64 | 1194 |
| Macro Avg | 0.59 | 0.61 | 0.54 | 1194 |
| Weighted Avg | 0.75 | 0.64 | 0.67 | 1194 |

Table 3.3.16: MLP Classification Report for spectral shape-based feature

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.92 | 0.82 | 0.87 | 203 |
| Boredom | 0.55 | 0.65 | 0.59 | 17 |
| Calm | 0.42 | 0.54 | 0.47 | 37 |
| Disgust | 0.72 | 0.96 | 0.82 | 119 |

Table 3.20 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Fearful | 0.84 | 0.85 | 0.84 | 164 |
| Happy | 0.84 | 0.67 | 0.74 | 210 |
| Neutral | 0.82 | 0.92 | 0.87 | 128 |
| Sad | 0.85 | 0.74 | 0.79 | 187 |
| Surprised | 0.79 | 0.91 | 0.84 | 129 |
| Accuracy | | | 0.81 | 1194 |
| Macro Avg | 0.75 | 0.78 | 0.76 | 1194 |
| Weighted Avg | 0.82 | 0.81 | 0.81 | 1194 |

Table 3.3.17: CNN Classification Report for spectral shape-based feature

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.84 | 0.90 | 0.87 | 180 |
| Boredom | 0.67 | 0.40 | 0.50 | 20 |
| Calm | 0.70 | 0.73 | 0.71 | 48 |
| Disgust | 0.85 | 0.88 | 0.86 | 159 |
| Fearful | 0.89 | 0.87 | 0.88 | 165 |
| Happy | 0.77 | 0.80 | 0.78 | 166 |
| Neutral | 0.87 | 0.86 | 0.86 | 144 |
| Sad | 0.90 | 0.84 | 0.87 | 164 |
| Surprised | 0.86 | 0.85 | 0.86 | 148 |
| Accuracy | | | 0.84 | 1194 |

Table 3.17 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Macro Avg | 0.82 | 0.79 | 0.80 | 1194 |
| Weighted Avg | 0.84 | 0.84 | 0.84 | 1194 |

The tables above show the classification reports of audio spectral shaped based features with SVM, CNN and MLP Classifiers. As can be seen

- from Table 3.15, Neutral and Fearful, Angry emotions have the higher f1 scores
- from Table 3.16, Neutral and Angry emotions have the higher f1 scores
- from Table 3.17, Sad, Fearful, Angry emotions have the higher f1 scores

Classification results demonstrate that MFCC feature is good at predicting Angry emotion. Besides that, the weighted accuracy ratios for whole algorithms are higher than the results belonging to the model with both time and frequency-based features.

### 3.3.4 Speech emotion recognition using pre-trained model-based features

Several experiments on the combined dataset are run for pre-trained model wav2vec large. In order to test the algorithm, %25 of data separated for testing phase and %75 used for training phase. Table 3.18 shows the methods and their corresponding accuracy and weighted accuracy results on the test data. According to the results, MLP and CNN have the same recognition rate whereas SVM has a lower rate for pre-trained model. However, there is no considerable difference between them.

Table 3.3.18: Model results for pre-trained vectors

| Method | Features | Accuracy | Weighted Average |
|---|---|---|---|
| SVM | Pre-trained vectors | 0.87 | 0.87 |
| MLP | Pre-trained vectors | 0.91 | 0.91 |
| CNN | Pre-trained vectors | 0.91 | 0.91 |

Table 3.3.19: SVM Classification Report for pre trained based features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.95 | 0.88 | 0.91 | 195 |
| Boredom | 0.70 | 0.88 | 0.78 | 16 |
| Calm | 0.92 | 0.57 | 0.70 | 77 |
| Disgust | 0.92 | 0.91 | 0.92 | 161 |
| Fearful | 0.93 | 0.88 | 0.90 | 176 |
| Happy | 0.73 | 0.88 | 0.80 | 138 |
| Neutral | 0.88 | 0.91 | 0.90 | 139 |
| Sad | 0.80 | 0.89 | 0.84 | 149 |
| Surprised | 0.89 | 0.92 | 0.90 | 143 |
| Accuracy | | | 0.87 | 1194 |
| Macro Avg | 0.86 | 0.86 | 0.85 | 1194 |
| Weighted Avg | 0.88 | 0.87 | 0.87 | 1194 |

Table 3.3.20: MLP Classification Report for pre trained based features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.94 | 0.94 | 0.94 | 181 |
| Boredom | 0.85 | 0.89 | 0.87 | 19 |
| Calm | 0.85 | 0.77 | 0.81 | 53 |
| Disgust | 0.92 | 0.92 | 0.92 | 159 |
| Fearful | 0.93 | 0.97 | 0.95 | 159 |
| Happy | 0.89 | 0.87 | 0.88 | 171 |
| Neutral | 0.94 | 0.92 | 0.93 | 148 |
| Sad | 0.88 | 0.88 | 0.88 | 163 |
| Surprised | 0.91 | 0.95 | 0.93 | 141 |
| Accuracy | | | 0.91 | 1194 |
| Macro Avg | 0.90 | 0.90 | 0.90 | 1194 |
| Weighted Avg | 0.91 | 0.91 | 0.91 | 1194 |

Table 3.3.21: CNN Classification Report for pre trained based features

| Emotion | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Angry | 0.91 | 0.96 | 0.93 | 180 |
| Boredom | 0.95 | 0.90 | 0.92 | 20 |
| Calm | 0.65 | 0.90 | 0.75 | 48 |

| | | | | |
|---|---|---|---|---|
| Disgust | 0.95 | 0.91 | 0.93 | 159 |

Table 3.21 (continued from previous page)

| | | | | |
|---|---|---|---|---|
| Fearful | 0.96 | 0.93 | 0.95 | 165 |
| Happy | 0.87 | 0.90 | 0.88 | 166 |
| Neutral | 0.94 | 0.92 | 0.93 | 144 |
| Sad | 0.90 | 0.84 | 0.86 | 164 |
| Surprised | 0.96 | 0.90 | 0.93 | 148 |
| Accuracy | | | 0.91 | 1194 |
| Macro Avg | 0.81 | 0.82 | 0.81 | 1194 |
| Weighted Avg | 0.91 | 0.91 | 0.91 | 1194 |

The tables above show the classification reports of audio pre- trained model-based features with SVM, CNN MLP Classifiers. As can be seen

- from Table 3.19, Angry and Disgust emotions have the higher f1 scores
- from Table 3.20 Angry and Fearful emotions have the higher f1 scores
- from Table 3.21 Angry, Fearful emotions have the higher f1 scores

Classification results demonstrate that the model with the pre-trained model gives superior results for whole emotions but especially for Angry and Fearful emotions. Besides that, the weighted accuracy ratios for whole algorithms are higher than all other models with traditional audio features.

# 4. DISCUSSION

In real life, an event occurs and affects people, meanwhile an emotional state appears and emotions change with other events. The interactions between one and his/her environment is complex, interactions have powerful impacts and shape the emotions. In literature there are different treatises on subjective emotion recognition from speech. Yüncü infers from his study that subjective emotion recognition accuracy rates are lower than automatic recognition accuracy results (Yüncü, 2013). On the other hand, uncontrolled environmental variables, from the point of psychological and cognitive appraisal and classifying natural emotions are difficult for machines. In order to avoid dealing with irrelevant circumstances, simulated datasets are chosen in this study. For obtaining an effective and competitive automatic emotion recognition system, different datasets are combined with different languages and emotions. In addition to this, speech emotion recognition is one of the under-resourced problems and this problem can be overcomed by bringing together different databases in this study.

Support Vector Machine models' recognition rates are 0.20, 0.55, 0.67 and 0.87 for time based, frequency based, spectral shape based and pre-trained model-based features respectively. For spectral shape based and pre-trained model-based features this algorithm has the lowest prediction rates. For time-based features' recognition rate there is no dramatic difference between MLP Classifiers and for frequency-based feature the results are almost the same. This indicates that conventional audio features and SVM algorithms have less ability to classify speech emotions.

Multi-Layer Perceptron models' recognition rates are 0.27, 0.54, 0.81 and 0.91 for time based, frequency based, spectral shape based and pre-trained model-based features respectively. For frequency-based feature this algorithm has the lowest prediction rate and for pre-trained model-based representations the highest rate is the same as CNN. This shows that MLPC and time and frequency-based features have less ability to classify emotions. Apart from these, determining hidden layer size is one of

the problems of MLP. In the study different numbers of layers are performed and hidden units size is determined as 300 as a result of these trails.

Convolutional Neural Network models' recognition rates are 0.14, 0.57, 0.84 and 0.91 for time based, frequency based, spectral shape based and pre-trained model-based features respectively. For time-based features this algorithm has the lowest prediction rate and for pre-trained model-based representations the highest rate is the same as MLPC. As a result, CNN and time and frequency-based features have less ability to classify speech emotions with conventional audio features. In addition to these, to avoid potential overfitting problem, networks regularized with applying dropout in each layer.

Neurons designed for information processing and their learning process interfere with parameters. Audio signals and time series gained from emotional utterances are not linear and they have complicated forms. Results of experiments in this study demonstrate that SVM has less ability to classify emotions among the other neural network techniques. It appears that certain patterns for emotions occur in the neural network's input space obviously than SVM's plane and neural networks outperform the classical classifier. Despite the close results between MLPC and CNN, MLPC performs remarkably faster than CNN. Computational efficiency and consuming energy are crucial points of a speech emotion recognition system. From this point of view, MLP and CNN models are more robust than SVM models whereas MLP Classifier is the most advantageous technique for speech emotion classification. On the other hand, the differences are not statistically significant between CNN-MLP and MLP-SVM models with pre-trained model for nine emotions. (Significance is established using paired t-test, adjusted p-value$< 0.05$)

Speech conveys an explicit message which is the linguistic part and implicit message which is the paralinguistic non-verbal part. Linguistic features not used in this work and the non-linguistic part seems to provide satisfactory results. Audio propagates by wave motions and at the course of wave motions energy is emitted. Furthermore, it changes over time and this change causes difficulties. Owing to these and many other anatomic reasons there are a myriad of derivable features. It is still

unknown which audio features are more relevant with emotions. In this study for examining which features represent the emotion better, time and frequency dimensions of audio are taken into consideration. Independent of these dimensions, a pre-trained model practiced instead of choosing features manually. Results point out that the input obtained from the pre-trained model provides superior predictions. Its main advantage over hand-crafted features is that it takes into consideration all of the audio features and attends prominent features. The pre-trained model is an unsupervised model and in this model the aim is finding the regularities in the input, irrelevant or little related emotional attributes are identified and removed. Both time and frequency features are suitable for emotion recognition. However, in the spectrum shape-based feature there are more discriminative features than time-based features. It seems that the time domain features are supplemental features for a SER system.

All in all, Neural Networks and a pre-trained model for feature extraction gives superior predictions for emotion recognition classification. To verify the effectiveness of the proposed method, the MLPC model is implemented on another independent emotional speech dataset named Acted Emotional Speech Dynamic Database (AESDD) which is not in the training phase. The AESDD dataset has 604 recordings for Anger, Fear, Sad, Happy, Disgust emotions in Greek language, there are almost equal number of recordings for each emotion. The accuracy of subjective emotion recognition is around %74 (Multidisciplinary Media & Mediated Communication Research Group). The MLP Classifier with wav2vec large pre-trained model gives 81% recognition rate which is higher than human listeners estimations. Moreover, this is a quite remarkable ratio for an emotion recognition system. This result also indicates that the developed model does not depend on speaker or language, the inference model can be implemented on any other emotional speech dataset. Moreover, pre-trained models may help skip pre-processing steps at least some of them.

When concentrating on the individual emotions classification results, it seems that some classes are extraordinarily evident to recognize. Fearful and Angry emotions have generally the highest recognition rates and they have %95 and %94 f1 scores respectively for the pre-trained model

Table 4.1: Speech Emotion Recognition Literature Survey

| Paper | Database | Features | Emotions | Classifier | Accuracy |
|---|---|---|---|---|---|
| Joy et al. (2021) | RAVDESS | MFCC, Mels, Chroma, Contrast, Tonnetz | Happy,Calm, Angry, Surprise, Fearful,Sad, Disgust | MLP | 70% |
| Xu et. al (2020) | IEMOCAP | MFCC | Happy,Sad, Excited,Neutral | Attention based CNN | 76% |
| Kumbhar et al (2019) | RAVDESS | MFCC | Happy,Calm, Angry, Surprise, Fearful,Sad, Disgust | LSTM | 80% |
| Liu et al. (2018) | CASIA | Time Domain Feature, Frequency Domain Feature | Normal, Happy, Sad,Fear, Surprise | CNN, DNN | 86% |
| Tawari et al. (2010) | EMODB | Speech Intensity, MFCC, Pitch, Speaking Rate | Surprise,Joy, Anger,Fear, Disgust, Sadness, Neutral | SVM | 84% |
| Iliou et al. (2010) | EMODB | Pitch, MFCC Energy, Formants | Joy,Anger, Fear,Disgust, Neutral, Surprise, Sadness | Probabilistic Neural Network | 94% |
| Casale et al. (2008) | EMODB | Log Energy Coefficient, Cepstral Coefficients, Pitch, Voice Class | Joy,Anger, Fear,Disgust, Neutral, Surprise, Sadness | SVM | 92% |

Table 4.1 provides a comparison for speech emotion recognition between different algorithms. The effect and importance of utilized algorithm and feature on classification results are obvious from the table. For instance, even both Tawari et al. (2010) and Casale et al. (2008) developed SVM models with EMODB, Casale et al. (2008) achieved higher results. Casale et al. (2008) developed models with EMODB similar with Iliou et al. (2010) Even their results are close, neural network achieved higher prediction rate. It is an interesting point that both Kumbhar et al (2019) and Joy et al. (2021) developed models with neural networks, RAVDESS set and they have MFCC feature as common and their results are quite different. In all these studies like many other studies seven emotions were classified. However, more emotions exist in real life. That's why, increasing the classified number of emotions is a notable advancement for emotion recognition from speech. In this study, nine emotions were classified. The enhancement of classes may cause some problems for example imbalance data in the models. Even though it can be said that the obtained classification results of this study is quite satisfactory. Liu et al. (2018) fused time domain and frequency domain features to classify five emotions. The developed model in this study provides higher results than Liu et al.'s study also.

In the study, CNN algorithm is applied on the combined dataset as well as RAVDESS, TESS and EMO-DB datasets separately. Table 4.2 provides individual datasets' recognition results. Time based audio features have lowest recognition rates and pre-trained model provides highest rates for individual datasets as well as the combined dataset. It is an interesting point that the results which are obtained by TESS dataset remarkably higher than the other datasets. For instance, TESS dataset's recognition rate is six times of RAVDESS dataset recognition rate and two times of EMO-DB dataset for Chroma feature. TESS dataset may consists of easily predictable audio records. For more realistic results, EMO-DB or RAVDESS datasets' results can be accepted as reference point. From this point of view, combined dataset provides highest recognition rate with CNN algorithm. Besides that combined dataset consists of nine emotions whereas the other datasets contains fewer emotions.

Table 4.2: Individual dataset recognition results

| Database | Feature | Accuracy | Weighted Accuracy |
|---|---|---|---|
| RAVDESS | ZCR-RMSE | 14% | 0.5% |
| RAVDESS | Chroma | 17% | 15% |
| RAVDESS | MFCC | 61% | 60% |
| RAVDESS | Pre-Trained | 79% | 79% |
| TESS | ZCR-RMSE | 22% | 19% |
| TESS | Chroma | 79% | 78% |
| TESS | MFCC | 100% | 100% |
| TESS | Pre-Trained | 100% | 100% |
| EMO-DB | ZCR-RMSE | 24% | 0.9% |
| EMO-DB | Chroma | 36% | 32% |
| EMO-DB | MFCC | 83% | 83% |
| EMO-DB | Pre-Trained | 89% | 89% |

## CONCLUSIONS AND FURTHER WORK

In this thesis, the theoretical background of machine learning approaches and psychological history of emotion are explained. The main purpose is determining discriminative speech features for emotion recognition and primary concern is the percentage of correctly estimated emotions.

As speech emotion recognition is a classifying task, the goal is detecting meaningful patterns in data. A statistical technique and two types of neural networks employed on the same combined dataset for classifying nine emotions. Totally, three ML algorithms are implemented for comparing discrimination of audio features. Time, Frequency and Spectral Shaped dimensions of audio and a pre-trained model compared for accomplishing most distinctive features. It is empirically proved that the models with the pre-trained representations reach superior outputs for all emotional states. In the learning algorithms case, according to the weighted accuracy rates neural networks offer more valid results compared with statistical modelling technique.

In general, it can be said that emotion recognition from speech has two significant parts. First one is feature selection and the other is classification method. For both of these, different strategies are proposed and it is obvious from the results that the main concern is determining features.

Real time emotion recognition has great challenges because of complex structure of voice, limited time for estimation, linguistic and paralinguistic factors of speech. To overcome these challenges various approaches have been developed. Numerous classifiers performed on many different emotional datasets to achieve high accuracy results. Recently CNN and RNN algorithms, sometimes separately and sometimes combined, applied on various datasets to create an advantageous model.

In future works, a hybrid model can be developed which contains both a pre-trained model and spectral shaped based features. Speech contains silent and noisy sections which increase the computational complexity. Time performance is the other major factor which should be a great deal of careful consideration. Although there are

many advancements on SER, custom architectures should be designed to fuse accuracy and time performance. Even further for a more realistic emotion estimation all physical gestures like voice, body parts of movement and facial expression can be obtained together as humans use them collectively to express themselves. In this study, the verbal part of the speech is ignored. Linguistic part can be combined also. Solely the effect of voice in emotion recognition among all other expressions is still not clearly defined and it seems it is not an easy task to do. Multimodal emotion recognition systems -facial expressions, bio signals and voice- that uses machine intelligence can be developed in later studies.

Duration effect of the utterances on emotional state is analysed in this research also. As a result, more accurate predictions are gained from shorter recordings for most of the emotions. However, current emotional speech datasets are not sufficient to decide the effects of duration on emotions. New emotional datasets can be developed with appropriate utterances for examining duration effect.

## REFERENCES


[Auditory system parts illustration], (n.d.), Mariemont Hearing Center, https://mariemonthearingcenter.com/how-the-ear-works/

[Speech system parts illustration], (n.d.), Music 2º E.S.O, Unit 3, http://music2eso.weebly.com/unit-3-human-voice.html

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K., (2019), A state-of-the-art survey on deep learning theory and architectures, Electronics, 8(3), p. 292.

Alpaydın, E., (2010), Introduction to machine learning, London: MIT Press.

Bank, D., Koenigstein, N., & Giryes, R., (2020), Autoencoders, ArXiv, abs/2003.05991.

Bello, J. P., (2018), Chroma and Tonality. Retrieval date: December 7, 2021, https://s18798.pcdn.co/jpbello/wp-content/uploads/sites/1691/2018/01/6-tonality.pdf

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B., (2005), A database of German emotional speech, Interspeech, (5), pp. 1517-1520.

Casale, S., Russo, A., Scebba, G., & Serrano, S., (2008), Speech emotion classification using machine learning algorithms, IEEE International Conference on Semantic Computing, pp. 158–165.

Chen, J. C., (2016), Elements of human voice, New Jersey: World Scientific Publishing Co.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G., (2001), Emotion recognition in human-computer interaction, IEEE Signal Processing Magazine, 18(1), pp. 32–80.

De Lara, J. R. C., (2005), A method of automatic speaker recognition using cepstral features and vectorial quantization [Paper presentation], Iberoamerican Congress on Pattern Recognition: Berlin, Heidelberg.

Deng, J., Xia, R., Zhang, Z., Liu, Y., & Schuller, B., (2014), Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition [Paper presentation], IEEE international conference on acoustics, speech and signal processing (ICASSP): Florence, Italy.

Dertat, A., (2017), Applied Deep Learning - Part 4: Convolutional Neural Networks, Retrieval date: December 7, 2021, https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2

Ear Anatomy – Inner Ear, (n.d.), UTHealth McGovern Medical School, https://med.uth.edu/orl/online-ear-disease-photo-book/chapter-3-ear-anatomy/ear-anatomy-inner-ear

Ekman, P., (1999), Basic Emotions. Handbook of Cognition and Emotion (pp. 45-60), Chichester: John Wiley & Sons Ltd.

Gunes, H., & Pantic, M., (2010), Automatic, dimensional and continuous emotion recognition, International Journal of Synthetic Emotions (IJSE), 1(1), pp. 68-99.

Hamdi, B., Limam, S., & Aguili, T., (2016), Uniform and concentric circular antenna arrays synthesis for smart antenna systems using artificial neural network algorithm, Progress In Electromagnetics Research B, pp. 67, 91-105.

Hertel, L., Phan, H., & Mertins, A., (2016), Comparing time and frequency domain for audio event recognition using deep learning [Paper presentation]. International Joint Conference on Neural Networks (IJCNN): Vancouver, BC, Canada.

Iliou, T., & Anagnostopoulos, C. N., (2010), Classification on speech emotion recognition - a comparative study, International Journal on Advances in Life Sciences, 2, pp. 18–28.

Jacobson, B. D., (1998 - 1999), Human Ear, Massachusetts Institute of Technology, https://web.mit.edu/2.972/www/reports/ear/ear.html

Jain, M., Narayan, S., Balaji, P., Bharath, K. P., Bhowmick, A., Karthik, R., & Muthu, R. K., (2020), Speech emotion recognition using support vector machine, ArXiv, abs/2002.07590.

Joy, J., Kannan, A., Ram, S., & Rama, S., (2020), Speech emotion recognition using neural network and MLP classifier, IJESC, 10(4), pp. 25170-25172.

Kattel, M., Nepal, A., Shah, A. K., & Shrestha, D., (2019), Chroma feature extraction [Paper presentation], Conference: Chroma Feature Extraction using Fourier Transform.

Kehtarnavaz, N., (2008), Digital Signal Processing System Design, Elsevier.

Khadkevich, M., & Omologo, M., (2011), Time-frequency reassigned features for automatic chord recognition [Paper presentation], IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP): Prague, Czech Republic

Korzeniowski, F., & Widmer, G., (2016), Feature learning for chord recognition: The deep chroma extractor, ArXiv, abs/1612.05065.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P., (2007), Supervised machine learning: A review of classification techniques, Emerging artificial intelligence applications in computer engineering, 160(1), pp. 3-24.

Kua, J. M. K., Sethu, V., Le, P., & Ambikairajah, E., (2014), The UNSW submission to INTERSPEECH 2014 ComParE cognitive load challenge. [Paper presentation] Fifteenth Annual Conference of the International Speech Communication Association: Singapore.

Kumbhar, H. S. & Bhandari, S. U., (2019), Speech Emotion Recognition using MFCC features and LSTM network [Paper presentatiton], 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA): Pune, India.

Kurokawa, H., & Goode, R. L., (1995), Sound pressure gain produced by the human middle ear. Otolaryngology-head and neck surgery: official journal of American Academy of Otolaryngology-Head and Neck Surgery, 113(4), 349–355.

Latinus, M., & Belin, P,. (2011), Human voice perception, Current Biology, 21(4), pp. 143-145.

Lian, Z., Tao, J., Liu, B., & Huang, J., (2019), Unsupervised representation learning with future observation prediction for speech emotion recognition. ArXiv, abs/1910.13806.

Lim, W., Jang, D., & Lee, T., (2016), Speech emotion recognition using convolutional and recurrent neural networks [Paper presentation], IEEE Asia-Pacific signal and information processing association annual summit and conference (APSIPA): Jeju, Korea.

Livingstone, S. R., & Russo, F. A., (2018), 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, 13(5). pp. 1-35.

Multidisciplinary Media and Mediated Communication Research Group, (n.d.). Acted Emotional Speech Dynamic Database (AESDD), http://m3c.web.auth.gr/research/aesdd-speech-emotion-recognition/

Neumann, M., & Vu, N. T., (2017), Attentive convolutional neural network-based speech emotion recognition: A study on the impact of input features, signal length, and acted speech, ArXiv, abs/1706.00612.

Noriega, L., (2005), Multilayer perceptron tutorial, Retrieval date: December 7, 2021, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.608.2530&rep=rep1&type=pdf

Nwe, T. L., Foo, S. W., & De Silva, L. C., (2003), Speech emotion recognition using hidden Markov models, Speech communication, 41(4), 603-623.

Pichora-Fuller, M. K. & Dupuis, K., (2020), Toronto emotional speech set (TESS), Scholars Portal Dataverse, V1, Retrieval date: December 7, 2021, https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=doi:10.5683/SP2/E8H2MF

Posner, J., Russell, J. A., & Peterson, B. S., (2005), The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Development and psychopathology, 17(3), pp. 715-734.

Robinson, D. J., (1999), The human auditory system [Paper presentation], 107th convention of the Audio Engineering Society (AES): New York, USA

Samuel, A. L., (1959), Some studies in machine learning using the game of checkers, IBM Journal of research and development, 3(3), pp. 210-229.

Satt, A., Rozenberg, S., & Hoory, R., (2017), Efficient emotion recognition from speech using deep learning on spectrograms, Interspeech, pp. 1089-1093.

Scherer, K. R., (2003), Vocal communication of emotion: A review of research paradigms, Speech communication, 40(1-2), pp. 227-256.

Schmitt, M., Ringeval, F., & Schuller, B. W. (2016). At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. Interspeech, pp. 495-499.

Schneider, S., Baevski, A., Collobert, R., & Auli, M., (2019), wav2vec: Unsupervised pre-training for speech recognition, ArXiv, abs/1904.05862.

Scholkopf, B., & Smola, A. J., (2018), Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT Press.

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G., (2010), Cross-corpus acoustic emotion recognition: Variances and strategies, IEEE Transactions on Affective Computing, 1(2), pp. 119-131.

Swain, M., Routray, A., & Kabisatpathy, P., (2018), Databases, features and classifiers for speech emotion recognition: a review, International Journal of Speech Technology, 21(1), pp. 93-120.

Tawari, A., & Trivedi, M. M., (2010), Speech emotion analysis: Exploring the role of the content, IEEE Transactions on Multimedia, 12, pp. 502–509.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S., (2016), Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network [Paper presentation], IEEE international conference on acoustics, speech and signal processing (ICASSP): Shanghai, China.

Wolf, J. J., (1972), Efficient acoustic parameters for speaker recognition, The Journal of the Acoustical Society of America, 51(6B), 2 pp. 044-2056.

Xia, R., & Liu, Y., (2013), Using denoising autoencoder for emotion recognition, Interspeech, pp. 2886-2889.

Xu, M., Zhang, F. & Khan, S. U., (2020), Improve accuracy of speech emotion recognition with attention head fusion [Paper presentation], 10th Annual Computing and Communication Workshop and Conference (CCWC): Las Vegas, USA.

Yüncü, E., (2013), Speech emotion recognition using auditory models (Master's thesis, Middle East Technical University).

Zheng, F., Zhang, G., & Song, Z., (2001), Comparison of different implementations of MFCC, Journal of Computer science and Technology, 16(6), pp. 582-589.