BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER SYSTEMS
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

EMOTION RECOGNITION FROM ANALYSIS OF A PERSON'S SPEECH

ROZPOZNÁNÍ EMOCÍ Z ANALÝZY ŘEČI ČLOVĚKA

MASTER'S THESIS
DIPLOMOVÁ PRÁCE

AUTHOR **Bc. MARTIN KNUTELSKÝ**
AUTOR PRÁCE

SUPERVISOR **doc. AAMIR SAEED MALIK, Ph.D.**
VEDOUCÍ PRÁCE

BRNO 2023

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

Master's Thesis Assignment

141159

Institut:	Department of Computer Systems (UPSY)
Student:	**Knutelský Martin, Bc.**
Programme:	Information Technology and Artificial Intelligence
Specialization:	Machine Learning
Title:	**Emotion Recognition from Analysis of a Person's Speech**
Category:	Biocomputing
Academic year:	2022/23

Assignment:

1. Study and learn about the various emotions and moods and how they affect the various features of the speech of a person.
2. Get acquainted with audio and speech processing methods as well as machine learning techniques and their application to the recognition of emotions and moods.
3. Find out the challenges for emotion and mood interpretation from a person's speech as well as the limitations of the existing methods.
4. Design an algorithm for interpretation of emotion and mood from the audio of a person's speech.
5. Implement the designed algorithm.
6. Create a set of benchmark tasks to evaluate the quality of emotion and mood recognition from a person's speech (audio) as well as the corresponding computational performance and memory usage.
7. Conduct critical analysis and discuss the achieved results and their contribution.

Literature:

- According to supervisor's advice.

Requirements for the semestral defence:

- Items 1 to 4 of the assignment.

Detailed formal requirements can be found at https://www.fit.vut.cz/study/theses/

Supervisor:	**Malik Aamir Saeed, doc., Ph.D.**
Head of Department:	Sekanina Lukáš, prof. Ing., Ph.D.
Beginning of work:	1.11.2022
Submission deadline:	17.5.2023
Approval date:	31.10.2022

Abstract

This thesis deals with the analysis of emotion recognition from human speech. It aims to design and implement a system that can automatically infer emotional states from speech recordings. The solution is based on the Audio Spectrogram Transformer (AST), a derivative of the Vision Transformer neural network, which accepts mel spectrogram as input. The implementation comprehends the pipeline with two stages. In the first stage, a mel spectrogram is obtained from the input speech recording and in the second stage, the pretrained AST model computes output in the form of probabilities of considered emotional classes. The AST implementation was trained and evaluated on three datasets: RAVDESS, Emo-DB and EMOVO. The obtained results in the form of unweighted accuracy are 84.5 % for RAVDESS, 91.6 % for Emo-DB and 73.8 % for EMOVO. During training, the consumed energy of the graphical processing unit was recorded for the calculation of the carbon footprint in terms of emitted CO_2. The main contribution of this work is the utilization of neural network based on Transformer architecture, originally used for vision tasks, to classify emotions from speech. Another contribution is carbon footprint tracking of neural network training. The carbon footprint, expressed in emitted CO_2 mass is 1058.37 grams.

Abstrakt

Táto práca sa zaoberá analýzou rozpoznávania emócií z ľudskej reči. Jej cieľom je navrhnúť a implementovať systém, ktorý je schopný automaticky klasifikovať emočný stav z rečových nahrávok. Riešenie je založené na neurónovej sieti typu Audio Spectrogram Transformer (AST), odvodenej z neurónovej siete Vision Transformer, ktorej vstupom je mel spektrogram. Implementácia riešenia pozostáva z dvoch častí. Prvá časť sa zaoberá extrakciou mel spektrogramu zo vstupnej nahrávky reči, zatiaľ čo v druhej časti predtrénovaný AST model počíta odozvu, ktorej výstupom sú pravdepodobnosti pre uvažované emočné triedy. Tréning a vyhodnotenie implementácie bolo uskutočnené na troch dátových sadách: RAVDESS, Emo-DB a EMOVO. Získané výsledky vo forme neváženej presnosti sú 84.5 % pre RAVDESS, 91.6 % pre Emo-DB a 73.8 % pre EMOVO. Počas tréningu modelu bolo zaznamenávané emitované množstvo CO_2 na základe spotrebovanej energie grafickým procesorom. Hlavným výstupom tejto práce je využitie neurónovej siete vychádzajúcej z architektúry typu Transformer, určenej pôvodone pre obrazové úlohy, na rozpoznávanie emócií z ľudskej reči. Ďalším výstupom je hodnota uhlíkovej stopy tréningu neurónovej siete, vyjadrená ako hmotnosť vylúčeného CO_2, ktorá dosiahla hodnotu 1058.37 gramov.

Keywords

speech emotion recognition, speech signal processing, classification of emotions, machine learning, deep learning, Vision Transformer, Audio Spectrogram Transformer, carbon footprint

Kľúčové slová

rozpoznávanie emócií z reči človeka, spracovanie rečového signálu, klasifikácia emócií, strojové účenie, hlboké učenie, Vision Transformer, Audio Spectrogram Transformer, uhlíková stopa

Reference

KNUTELSKÝ, Martin. *Emotion Recognition from Analysis of a Person's Speech.* Brno, 2023. Master's thesis. Brno University of Technology, Faculty of Information Technology. Supervisor doc. Aamir Saeed Malik, Ph.D.

Rozšírený abstrakt

Témou tejto diplomovej práce je rozpoznávanie emócií z reči človeka. Toto rozpoznávanie je založené na zmenách rečových vlastností (napr. výška, rýchlosť, chvenie), ktoré sú prejavom buď vedomých alebo nevedomých emočných impulzoch. Tieto zmeny je možné extrahovať vo forme rečových príznakov, na základe ktorých je vyvodený záver akou emóciou bol rečový signál podfarbený.

Súčasné riešenia využívajú pre zostavenie systému rozpoznávajúceho emócie z reči techniky jej predspracovania, ktoré sú nasledované extrakciou príznakov. Tieto príznaky sú použité ako vstupy pre algoritmy strojového učenia, ktorých výstup sa interpretuje vzhľadom na zvolený emočný model. V súčasnosti systémy, ktoré dosahujú najlepšie výsledky využívajú pre svoju činnosť extrakciu kepstrálnych koeficientov (napr. mel-frekvenčné kepstrálne koeficienty alebo lineárne prediktívne kepstrálne koeficienty), ktoré sú použité ako vstupy pre hlboké neurónové siete. Najčastejšie využívanými typmi neurónových sietí pre rozpoznávanie emócií z reči sú konvolučné a rekurentné neurónové siete, prípadne ich kombinácia. Z relevantných zdrojov a prác zaoberajúcich sa podobnou problematikou boli vybrané štyri systémy na základe ich zverejnených výsledkov a použitých technológií. Ich popis s krátkym komentárom sú súčasťou práce.

Výsledkom analýzy súčasného stavu problematiky bol zoznam nedostatkov, ktorých odstránenie má potenciál zlepšiť výkon rozpoznávania. V rámci tejto práce boli stanovené ako hlavne ciele, na ktoré by sa malo riešenie zamerať, nasledovné body: využitie neurónovej siete typu Transformer (alebo jej derivátu), využtie predtrénovaných modelov neurónovej siete a sledovanie uhlíkovej stopy, ktorú tréning modelu zanechá.

Na základe stanovených bodov bol navrhnutý systém vychádzajúci z neurónovej siete Audio Spectrogram Transformer (AST), ktorá je derivátom neurónovej siete Vision Transformer (ViT). Sieť AST bola navrhnutá za účelom klasifikácie audio nahrávok na základe mel spectrogramu preto sa líši od pôvodnej ViT architektúry v premietaní vsupu na sekvenciu embeddingov. V rámci pôvodnej implementácie AST je dostupná predtrénovaná sieť na dátovej sade ImageNet.

Predložený návrh pozostával zo systému, ktorý sa skladal z vlastného extraktoru mel spektrogramu zo vstupnej nahrávky a predtrénovaného AST modelu. Návrh ďalej obsahoval zoznam troch databáz s nahrávkami označenými pomocou emočných tried (RAVDESS, Emo-DB, EMOVO), na ktorých mal byť AST model netrénovaný spoločne s plánom vyhodnotenia výkonnosti pozostávajúcim zo sady klasifikačných metrík. Súčasťou návrhu je aj postup pre meranie vypusteného CO_2, ktorý sa zameriava na spotrebu energie grafického procesora využívaného počas tréningu modelu.

Implementácia sa skladala z troch častí. V prvej sú implementované moduly pre načítanie, spracovanie a augmentáciu prvkov dátovej sady. Druhá časť sa zameriava na tréning neurónovej siete na vybranej dátovej sade. Ako tretí je vyhotovený testovací modul, ktorý má za úlohu zostaviť zoznam predikcií na testovacej časti dátovej sady. Tento zoznam je vstupom pre vzorce sledovaných klasifikačných metrík.

Experimenty a ich vyhodnotenie sú uskutočnené na každej dátovej sade samostatne bez ich miešania. Tréning modelov prebieha pomocou metódy desaťnásobnej krížovej validácie.

Vyhodnotenie je rozdelené do dvoch častí. V prvej sú interpretované výsledky predikcie natrénovaných modelov na jednotlivých dátových sadách. Za všetky uvažované metriky je prezentovaná nevážená presnosť, ktorá dosahuje hodnoty 84.5 %, pre RAVDESS, 91.6 % pre Emo-DB a 73.8 % pre EMOVO.

Druhá časť vyhodnotenia sa venuje uhlíkovej stope, ktorá bola zanechaná pri tréningu. Celkové množstvo vypusteného CO_2 na základe spotrebovanej elektrickej energie grafickým procesorom *Nvidia A40* dosiahlo hodnotu 1058.37 gramov.

Autor vidí ako jeden z perspektívnych smerov ďalšieho rozvoja tejto práce otestovanie AST modelu na väčšej dátovej sade, napríklad MSP-Podcast, ktorá obsahuje viac ako sto tisíc rečových nahrávok ľudskej reči získaných z prirodzeného prostredia ako sú diskusné relácie a podcasty. Ďalší smer, ktorým sa je možné v budúcnosti zaoberať je tréning AST modelu s inými vstupnými parametrami (napr. zmena vo veľkosti mel spektrogramu alebo úprava vzorkovacej frekvencie) za účelom skúmania ich vplyvu na výkonnosť.

Emotion Recognition from Analysis of a Person's Speech

Declaration

I hereby declare that this master thesis was prepared as an original work by the author under the supervision of doc. Aamir Saeed Malik, Ph.D.

. . . .
Martin Knutelský
May 16, 2023

Acknowledgements

I would like to thank my supervisor doc. Aamir Saeed Malik, Ph.D. for his guidance. Also, I would like to thank my family, girlfriend and friends who supported me while writing this thesis and throughout my entire university studies.

Contents

List of Figures

Chapter 1

Introduction

Currently, the field of human-computer interaction (HCI) rapidly grows due to the ubiquitous presence of technology in human life. The interaction between man and machine is on a daily basis, and there is demand for more sophisticated methods that would enable a machine to communicate with people as they do among themselves. To unlock such advanced technology, the machine needs to have human-like properties that allow perception and processing of the human stimulus and generation of an appropriate response.

Even though most of the interactions with machines are through touch devices (keyboard, mouse, smartphone display), the trend in the current HCI field is to adapt machines to communicate with people through speech as the main and most natural communication tool that is an essential part of their daily life. Its interactive and straightforward character makes it a perfect transmitter of information enabling people to express their thoughts in real-time. Through the speech, there are encoded also other non-verbal features that complete the context of the speaker's thoughts. One such non-verbal expression is the emotional footprint, which provides insights into the feelings of a speaker during communication. To create a machine that knows properly communicate, the awareness of emotional context is crucial. The machine can obtain this ability through analysis of speech in order to decode and obtain this context.

The research area that deals with gaining emotional information from human speech is called *speech emotion recognition*. The practical reasons to consider this topic are e.g. the ability to adjust machine behaviour to correspond with human emotional state, providing better service in call centres or better understanding of the nature of emotions. Despite trying to find a universal solution for more than twenty years, there is no present system that would even approach human abilities in emotion perception, analysis and classification. The main reason for a such situation is the unclear phenomenon of emotions and their interpretation even from the human point of view. This work aims to contribute with an innovative solution that follows current trends in the research and achieves a comparable performance with the best systems so far.

Chapter 2 is dedicated to the analysis of the speech emotion recognition field. It comprehends the basic psychological theory of emotions, emotional models and the impact of emotions on human speech. This chapter continues with the description of methods for preprocessing, feature extraction and classification with trends for each topic. The last sections are dedicated to current state-of-the-art solutions and a list of issues and problems that are not yet solved.

Chapter 3 contains a theoretical explanation of concepts necessary to understand the solution design based on the Audio Spectrogram Transformer neural network (AST). Then, Chapter 4 describes the implementation of the proposed solution. Chapter 5 explains the evaluation methodology, describes the set of performed experiments with obtained results and conducts discussion about them.

Chapter 2

Speech emotion recognition

Speech emotion recognition (SER) is a task, where the machine is able to infer emotion from the human voice. This task belongs to the field called *Affective computing*, which joins together computer science, psychology and cognitive science [42].

Systems performing SER are complex and they are assembled from components, which consider different parts of the SER process. The general structure of this system is depicted in Figure 2.1, which suggests there are several areas to study in order to know and understand the entire system, in particular, these: an underlying psychological theory of emotions, speech signal processing, and machine learning. This chapter is designated to describe each of these areas at a high-level overview and also in the context of the SER.

2.1 Emotion analysis

According to the Cambridge dictionary, emotion is: „*A strong feeling such as love or anger, or strong feelings in general*" [10]. However, from a scientific point of view, there is still no clear consensus on what emotion actually is. Scientists have been studying the emotional phenomenon for centuries and have a variety of theories about what causes it.

From the current point of view, the first serious study of emotions was conducted by Charles Darwin in the nineteenth century, when he treated emotions as biological processes [13]. He laid down the foundations for the next development of emotional research with his theory called **the Facial feedback hypothesis**, in which he described the connection of facial neurons and emotions. It explains that emotions are influenced by the activity of muscles responsible for facial expressions. Later, it was proven by experiments that the inhibition of facial neurons causes degradation of the brain part responsible for the interpretation of emotions [20].

Nowadays, there is no general agreement on whether emotions have a physiological or cognitive base or in other words, if humans feel emotions, because of physiological arousal or cognition. According to the Jeon [23], there are three fundamental theories explaining the origin of emotions: **James-Lange theory**, **Cannon-Bart theory** and **The Schachter-Singer two-factor theory**.

The James-Lange theory says that the first occurs the physiological manifest to external emotional stimulus. Emotional perception is a result of the cognitive awareness of particular physiological change [22]. It is one of the oldest theories, which is still not fully disproved.

The Cannon-Bard theory explains that physiological change and emotion occur simultaneously and independently [11]. Cannon proposed his ideas as the response to the James-

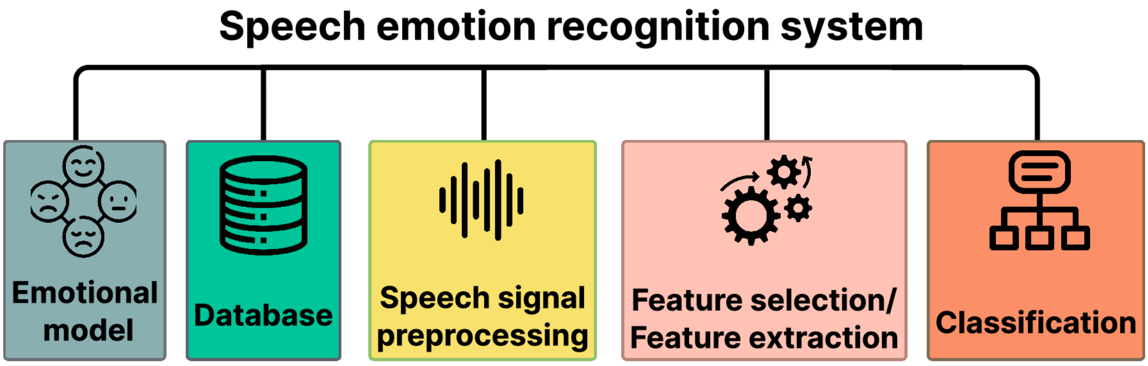

Speech emotion recognition system
Emotional model
Database
Speech signal preprocessing
Feature selection/ Feature extraction
Classification

Figure 2.1: SER system and its components. In order to design such system, it is required knowledge from psychology, signal processing and machine learning. It depicts its overall complexity.

Lange theory, where he argued that the same physiological change can cause different emotions, e.g. crying can lead to sadness emotion as well as to happiness; hence he wanted to show that James-Lange theory is unlikely.

The baseline for the Schachter-Singer two-factor theory is that mental processes such as information processing and the creation of thoughts are the main reason for the formation of emotions. It indicates that emotion is a result of the perceptual processing of emotional stimulus and physiological changes [38]. It explains that when the stimulus comes, the physiological change appears, but only after the cognitive process of mentioned stimulus together with physiological change, the man starts to perceive emotion.

This thesis aims to propose the SER system that recognizes emotions from a speech in a way that approaches the abilities of humans, and to do so there is a need to study how to categorize emotions like them as the fundamental brick for the following work. Two types of emotional models are interesting from the SER point of view: **discrete (categorical)** model and **dimensional (continuous)** model [4]. The following subsections are designated to these models with their definitions and descriptions of advantages and disadvantages.

2.1.1 Discrete model

The discrete model divides emotions into disjunct categories. Ekman and Oster proposed one of the most popular types of discrete models when they studied the expressions of different cultures. They concluded that there are six emotion categories: *anger*, *fear*, *disgust*, *happiness*, *surprise*, *sadness* [16]. Later, Ekman observed that the proposed categorization could be proved not only by visual factors but also by observing physiological changes when a man perceives emotion [17].

The study by Shaver et al. used a prototype-based approach, where they grouped 135 expressive words into prototypes or groups, where one word was characteristic. With this methodology, they found these emotional groups: *joy*, *sadness*, *fear*, *anger*, *love* and *surprise* [39].

Another discrete model was proposed by Robert Plutchik. This model is called the **emotional wheel** and is depicted in Figure 2.2. This model represents eight basic emotions joined in pairs, in which each pair contains mutually opposite positive and negative emotions e.g. joy and sadness, fear and anger etc. It is also apparent from figure 2.2 that the emotional wheel considers compound emotions such as optimism or love, which are made up of basic emotions.

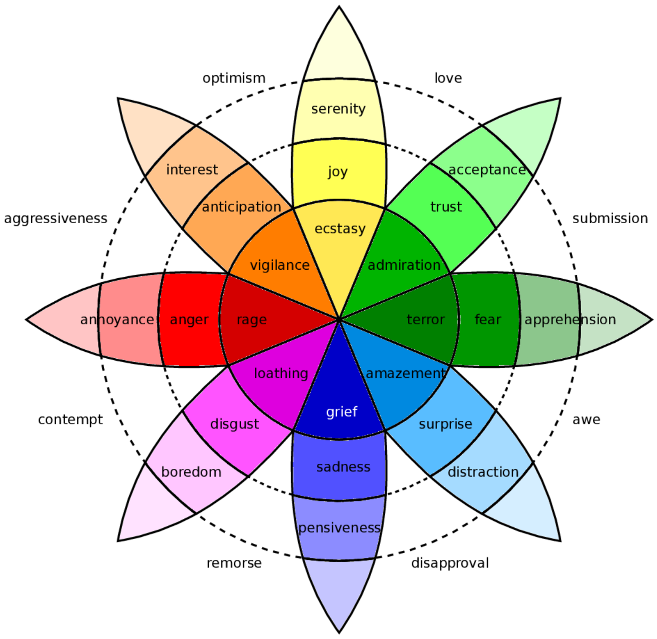


Figure 2.2: Plutchik's emotional wheel [36]. It contains eight atomic emotions grouped in the pair, where each contains mutually exclusive emotions, e.g. joy and sadness. The second-level or compound emotions are made of basic emotions, e.g. submission or disapproval.

The advantage of a discrete model is its intuitiveness and comprehensibility because emotions are often described in this way. However, it does not have such expressive power to depict complex emotions which occur in everyday life from a scientific point of view.

In the SER, the discrete models have been predominant since the beginning of its research [4], but it is not clear why researchers take into account this type of model in most approaches. In the author's opinion, one of the likely explanations is that most of the available databases for this area, examined in Section 2.2, consider data labelled in a way that follows the discrete model pattern. However, many of them do not strictly follow some of the mentioned categorizations. Mostly, they can be regarded as derivatives of Ekman's model with subtle domain-dependent adjustments.

2.1.2 Dimensional model

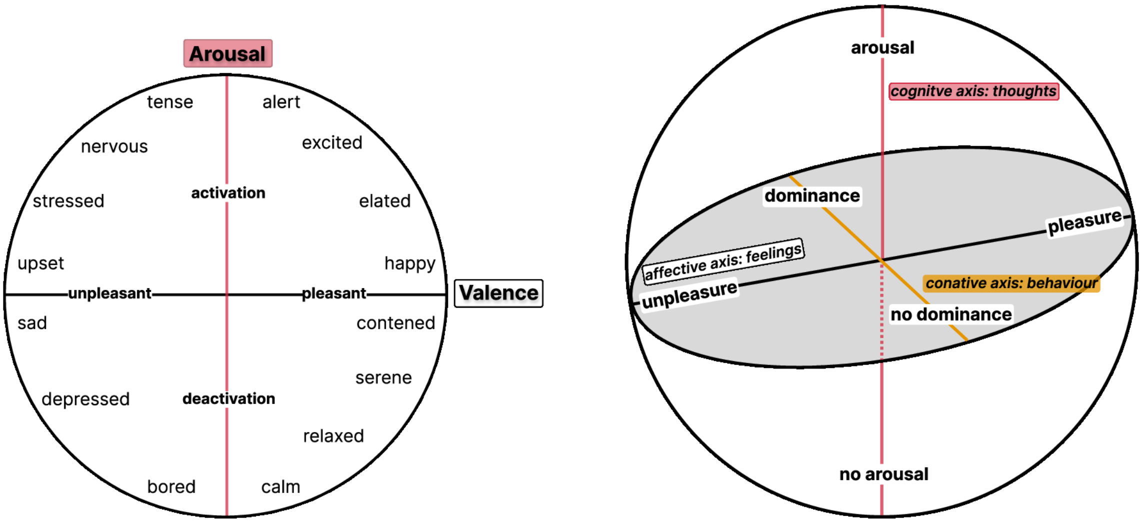
For more precise emotion categorization researchers proposed the dimensional model represented as an n-dimensional continuous space, where each axis denotes some emotional characteristics. With this model, it is possible to express emotion as a point defined by its coordinates.

The most relevant dimensional models in the context of SER [4] are the Circumplex model by Russell, depicted in Figure 2.3a [37] that contains the horizontal axis representing the valence domain and on the vertical axis is the arousal domain. Around these axes are organized particular emotion states in a circular shape expressed as a ratio of valence and arousal.

Another famous model is Mehrabian's Pleasure-Arousal-Dominance (PAD) three-dimensional model. Each of these terms represents one axis, as is shown in Figure 2.3b [32]. This model is comparable with Russell's Circumplex model because the only difference between them is that the PAD model has an extra dominance axis. It represents a level of the subject's control above the current state.

The advantage of the dimensional model is that it enables the expression of emotional state as a point in continuous n-dimensional space. It implies the ability to compute the exact similarity degree between emotions. The disadvantage of this model is its non-intuitive approach to emotions and the fact that some emotions cannot be precisely expressed, e.g. *surprise* in two-dimensional valence-arousal space, because its valence can be either positive or negative. Bakker et al. [6] consider as the disadvantage of the dimensional model that there is no unified axis labelling, and hence it is not clarified the interpretation of them. In other words, there is no standardized measurement unit for the dimensions.

According to the reviewed literature [4], the most commonly used dimensional model is the three-dimensional PAD model.



(a) Circumplex model [37].

(b) PAD model [6].

Figure 2.3: Dimensional models.

Model type	*Emotions interpretability*	*Emotions comparability*	*Datasets availability*
Discrete model	similar to human emotions perception	not present any exact metrics	many available databases (freely available, on demand)
Dimensional model	ratings in given number of dimensions, less intuitive	cosine similarity, Euclidean distance	a few available datasets

Table 2.1: Emotional models comparison.

2.2 Databases

Databases in the context of the SER serve for obtaining data that are necessary for further building a classification model. Database quality is crucial because it directly influences the system's performance. If the database contains low-quality, incomplete or data that are damaged in some way, the results of the system will not be satisfactory, so the choice of the database has to be performed carefully.

The categorization of the databases from the SER perspective includes three types: **natural speech (spontaneous)**, **acted**, and **elicited** [46]. Databases from the spontaneous category contain voice recordings from a natural environment, e.g. dialogues from talk shows or podcasts. This database type is the most valuable because it comprehends emotional expressions from real-life situations, which are hard to reproduce artificially. However, due to privacy laws, it is difficult to build and access these databases.

The acted databases contain recordings produced by professional actors in recording studios, which implies the high quality of such voice recordings. Their advantage is that they are relatively easy to obtain, and most are even free. On the other hand, the expressed emotions may be exaggerated, possibly leading to degradation of recognition performance on real-life recordings [4].

The elicited databases have voice recordings created in a simulated emotional environment [46]. A speaker is placed into a situation that stimulates him in order to evoke a particular emotion. After this procedure, his speech is recorded.

Most of the available databases are from the category acted and are labeled with discrete emotional categories. Currently, according to the number of citations from the scientific portal Scopus[1], the most popular databases used for SER research are **Berlin Database of Emotion speech (Emo-DB)** [7], **The Interactive Emotional Dyadic Motion Capture (IEMOCAP)** [8] and **The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDEES)** [30]. We have provided additional information about mentioned databases in the table 2.2.

[1] https://www.scopus.com/

Name	*Type*	*Language*	*Emotions*	*Citations count*
Emo-DB [7]	Acted	German	Neutral, anger, sadness, fear, boredom, happiness, disgust.	1237 [1]
IEMOCAP [8]	Elicited	English	Anger, happiness, sadness, frustration, neutral.	1421 [2]
RAVDESS [30]	Acted	English	Surprise, anger, fear, disgust, sadness, neutral, calm, happiness.	558 [3]

Table 2.2: Brief description of databases for SER [46].

2.3 Speech and sound analysis

A speech is a voice communication method based on the rules of the chosen language. Speech production is a complex process that starts in the human brain, which sends neural signals transmitted into a series of physiological changes in the human body. The vocal cords placed in the larynx have the principal function in speech production. These instruments change the airflow from the lungs out of the human body in a way that results in speech signals propagating through an environment.

The following subsections are designated to the basic speech properties that characterize its perception, the speech processing, the useful speech features in the context of the SER and how to extract them. The end of the section belongs to classification methods and deep learning. The author emphasizes that the focus is given to offline methods that relate to SER systems retrieving recordings from databases, not from a real-time environment.

2.3.1 Speech acoustic properties

The acoustic properties of the speech signal influence its information value. The first examined property is **the fundamental frequency of speech** (F_0), which is the lowest frequency component of a complex sound wave or in other words, the first peak of the signal's power spectrum. In the case of speech, it is the frequency of oscillating vocal cords (number of times that vocal cords open and close per second) [9]. People perceive this frequency as a pitch of the voice and it is the main property for recognition of voiced and unvoiced parts of the speech.

The following inspected property is **the resonance of the vocal tract**. It describes how well is the voice from vocal cords amplified in the cavities above them. The resonance structure is described by *formants*. The most discussed formants are F_1 and F_2, also called the lowest resonance formants [9]. The whole resonance structure is obtained from the envelope of the speech power spectrum, as illustrated in Figure 2.4. In this figure, the blue curve represents the magnitude of frequencies, which is possible to observe in a speech wave and the red curve represents its spectral envelope. In this case, the fundamental frequency F_0 of this signal is around 150 Hz with 17 decibels intensity, formant F_1 has the frequency 700 Hz with intensity near 37 decibels and formant F_2 has the approximate frequency of 1200 Hz and intensity almost 30 decibels.

The last considered property is **the audio signal amplitude (intensity)**. It shows the deviation in air pressure in time (during the flow of the speech). The most common way

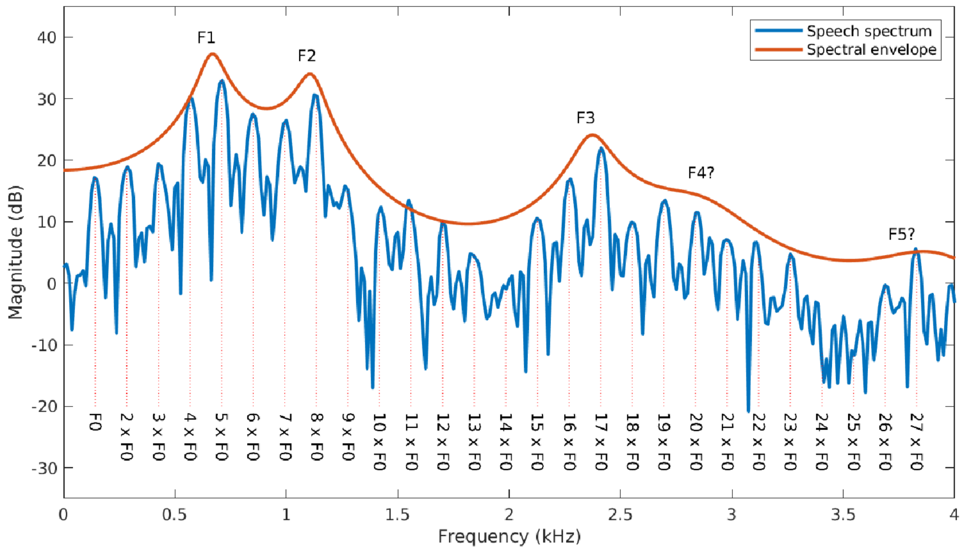
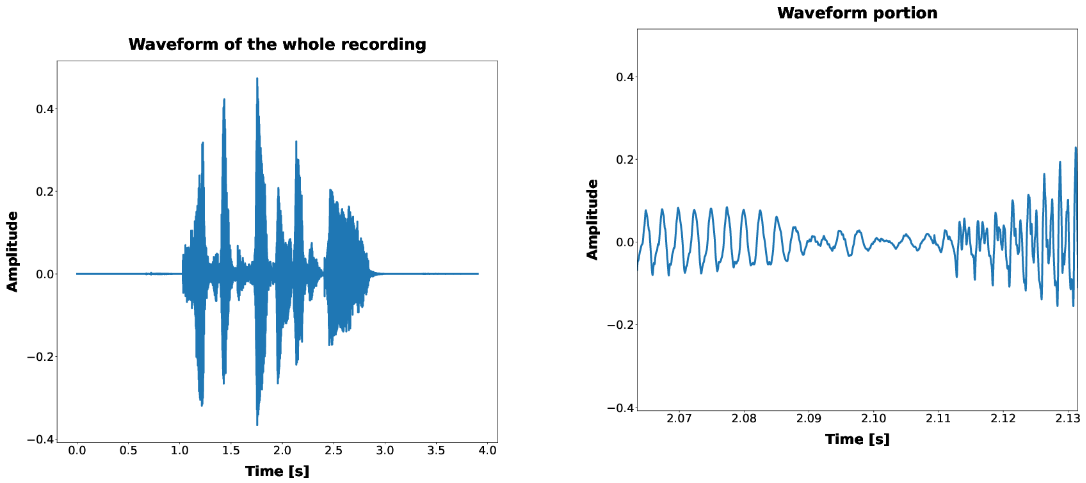


Figure 2.4: Audio signal power spectrum and its spectral envelope [9].

to plot this deviation is a waveform, whose examples are shown in Figure 2.5. Subfigure 2.5a shows a waveform of the whole recording, but due to the dense amplitude array, the peaks are not well recognizable and thus we provide a detailed portion of the waveform that is depicted in Subfigure 2.5b.



(a) Waveform of whole speech recording.

(b) Waveform portion. It is zoomed version of the original waveform.

Figure 2.5: Waveform examples.

2.3.2 Influence of emotions on the speech properties

This subsection inspects the impact of emotions on particular speech properties. Nwe et al. compiled a complex overview of this topic and described how emotions such as *anger*, *joy*, *fear*, *sadness*, *surprise* and *disgust* influence speech properties [34]. The following paragraphs provide a summary of this overview. All comparisons are relative to speech considered neutral (without any emotions).

For *anger*, the overview shows that the average pitch is higher and its range is wider. The intensity of the voice is raised, and the energy of the unvoiced speech part is higher.

The emotion of *joy* has similar parameters as anger; hence, it has a higher pitch with a wider frequency range and increased intensity however, in terms of the spectral properties, the energy is distributed into higher frequencies.

In the case of *fear*, the average pitch value with its range are higher, but the intensity of the speech is almost the same as in the neutral state. In terms of the speech spectrum, more energy is present in higher frequencies.

Sadness is the opposite case of the aforementioned emotions. The average pitch is lower and the pitch range is narrower. The intensity is significantly decreased and there are often present downward inflexions.

Surpise pitch has a wide range and its mean value is neither above nor below the normal state. It is caused by the character of emotion that allows expressing the positive, as well as negative emotions.

The last emotion discussed in this overview is *disgust*. It has a pitch a lot below the average of neutral, but the frequency band is slightly wider. The intensity of speech during the disgust emotion is rather low.

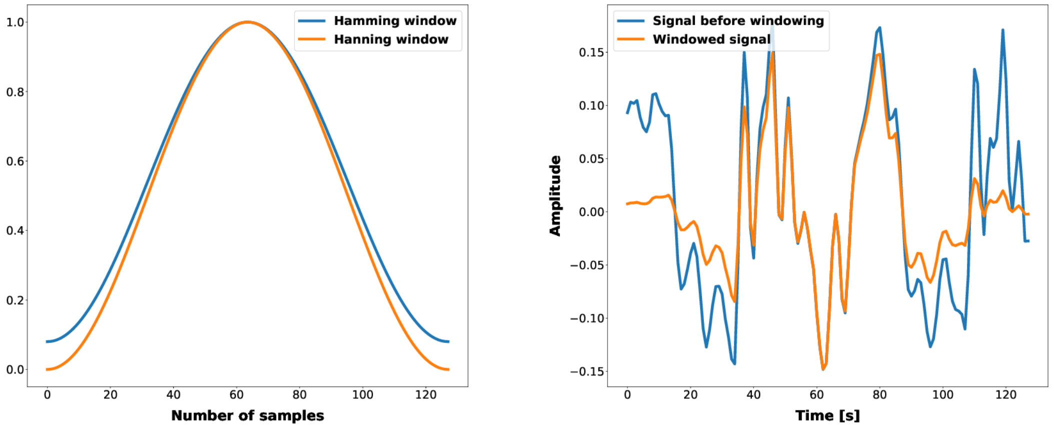
2.3.3 Speech preprocessing

Preprocessing includes methods that manipulate audio signals in a way that makes them more suitable for extraction and selection of the features. This subsection examines the preprocessing techniques relevant to the thesis topic.

The first examined method is called **framing** or **segmentation**. It divides the speech signal into smaller chunks (frames or segments) with lengths from 20 to 30 milliseconds. Framing is used because the expression of emotions tends to vary during the flow of the speech, which causes the audio signal to be non-stationary, but for this short time period is invariant and thus suitable for examining semi-fixed and local features [46]. This method allows Discrete Fourier Transform (DFT) to be performed and is also convenient for use with some classifiers [4].

Another processing method related to the above framing is **windowing**. It multiplies the speech signal with a window function (e.g. Hanning window, Hamming window), which preserves values in the middle of the signal and smooths values at both ends removing discontinuities between neighbouring frames [46]. The main reason for applying windowing is the preparation of frames for transformation from the time to the frequency domain. The negative aspect of windowing is information loss but it is possible to remove this drawback by setting sufficient overlay between frames during signal split.

Subfigure 2.6a shows how Hamming (blue curve) and Hanning (orange curve) windows look like for frames with a length of 128 amplitude values (samples). Both functions have a bell shape, and their peak is in the middle of the horizontal axis. From this subfigure follows that the biggest difference between these windowing functions is at both ends. Next, Subfigure 2.6b depicts the signal before (blue curve) and signal after (orange curve) windowing with Hamming window. The key property is saving the middle values of signals as much unchanged as possible, while completely eliminating marginal values that could cause discontinuities.



(a) Hamming and Hanning windows.

(b) Demonstration of windowing effect.

Figure 2.6: Windowing examples.

Another practical method is **normalization**, which reduces the influence of speaker and recording variability in audio signal values without loss of their discriminatory properties [4]. It is performed either at the frame or recording level. The most widely used normalization method is Z-normalization, which we compute for each signal value in sequence x according to equation 2.1, where μ is the mean and σ is of the audio signal.

$$z = \frac{x - \mu}{\sigma} \tag{2.1}$$

The last examined preprocessing technique is **voice activity detection (speech detection)**, which is useful for detecting if there is present some speech activity (either voiced or unvoiced) or if it is a silent part of the recording [4]. The purpose of this method is to eliminate speech frames that contain silent parts, which increase computational complexity and decrease the performance of the SER system. The most used algorithms for implementation of voice activity detection are **zero-crossing rate** and **auto-correlation method**.

2.4 Speech features

Speech features are a set of values that describe important speech properties. They are one of the key elements that have a big impact on the performance of the SER systems; hence, it is necessary to take care of their proper selection process. Although the SER has been a research-interesting topic for a long time, there is still no consensus about the universal set of features that we can reliably use in our SER system without consideration of their pros and cons [4].

Currently, a lot of features from different categories were utilized during the SER research. This section studies the most relevant features according to the reviewed literature. The first and most simple categorization is differencing between **segmental (local)** and **supra-segmental (global)** features [46]. Supra-segmental features are obtained from the whole speech signal (e.g. minimum, maximum, variance), while segmental are acquired from the individual frames of the partitioned recording. These frames contain the speech temporal dynamics that are important characteristics of speech features. Their study is significant because the expression of emotions is not uniform during the speech mentioned also in subsection 2.3.3.

According to the literature review for this topic, the following speech features categories are considered: **prosodic features**, **spectral features**, **voice quality features** and **teager energy operator (TEO) based features** [46], which are discussed in the following subsections.

2.4.1 Prosodic features

This feature type is connected to the prosodic (para-linguistic) speech properties, which man can perceive as intonation, tempo or rhythm. These properties are applied during the speech to units like syllables, words or sentences.

Speech phenomena suitable for prosodic feature gain are the fundamental frequency F_0 (defined in subsection 2.3.1), the energy of the speech (directly connected to the speech signal amplitude) and the speech duration. These features are called time-based and supra-segmental because they are related to the change of speech in time and are examined in the bigger portions of the recording, respectively.

Prosodic features are the results of the statistical methods such as mean, maximum, minimum and variance applied to the measured values of the above phenomena [4]. The performance level of the SER depends on the level which is for features computation (word level, sentence level, recording level).

2.4.2 Spectral features

On the opposite of the prosodic features, the spectral features are the result of a short-time analysis of the speech signal; hence, extracted features represent local changes in the speech, which are significant for emotion recognition. From the high overview, the spectral features represent the spectral energy distribution across different frequencies. This energy distribution is implied by the shape of the vocal tract [18].

To obtain spectral features, the Fourier transformation is performed on each frame of the speech signal, which results in frequency domain representation of the signal. It is possible to investigate spectral features after this kind of transformation or with further operations gain cepstral features. [41].

According to reviewed literature [41, 4, 46, 18], the most used spectral features are **Formants** (described in 2.3.1), **Mel Frequency Cepstral Coefficients (MFCC)**, **Gammatone Frequency Cepstral Coefficients (GFCC)**, **Log-Frequency Power Coefficients (LFPC)** and **Linear Prediction Cepstral Coefficients (LPCC)**.

MFCCs represent the short-term power spectrum of the speech signal expressed in the mel scale. They are defined as the result of the Discrete cosine transformation computed from the log of the power spectrum expressed in the mel scale. The argument for MFCCs utilization in the SER is that they mirror how the human auditory system works. The transition is performed from hertz to mel scale with a set of filters from the mel filter bank, which example is depicted in Figure 2.7. This bank contains six filters distinguished by colours.

GFCCs are another cepstrum-based features obtained with an almost similar method as MFCCs but it is used the gamma filter bank, instead of the mel bank. This filter bank represents changes in the inner ear and external middle ear [29].

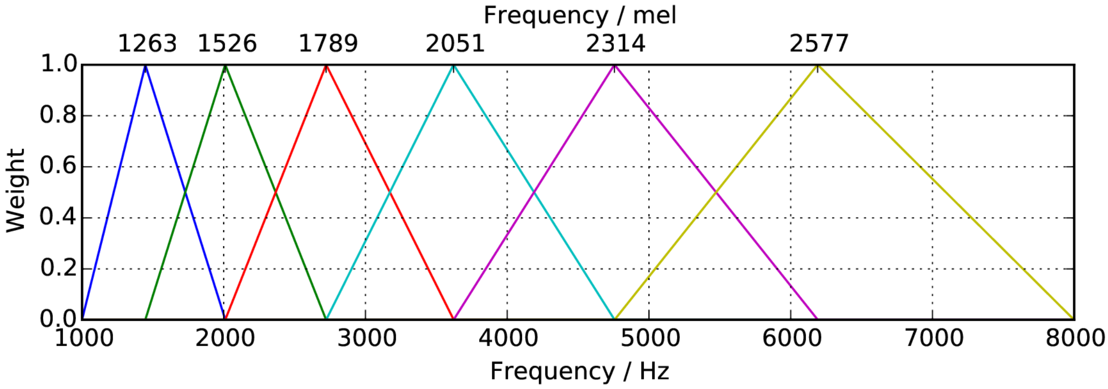


Figure 2.7: Mel filter bank example with 6 filters.

They are included in the family of cepstrum coefficients as well with the LFPCs, which reflect the frequency filtering property of the human auditory system expressed on the logarithmic scale. In the computational process is used a filter bank that contains log frequency bands [35].

LPCCs are Linear Prediction Coefficients (LPC) represented in cepstrum. The main idea behind LPC is a mathematical speech production model, where the vocal tract part is modelled with the all-pole filter with coefficients equal to LPC [48]. The reason for the usage of LPCC is they are more robust than LPC.

2.4.3 Voice quality features

The physical state of the vocal tract (e.g. condition of vocal cords) highly influences this feature type. The defects in speech signal cause effects called **jitter**, **shimmer** and **harmonics to noise ratio (HNR)**.

Jitter and shimmer are related to fundamental frequency F_0, and it is possible to describe them together. They are perceived as roughness, shortness of breath, or hoarseness in speech. In terms of the fundamental frequency, the jitter represents frequency instability (variation in vibrations of the vocal cords), and the shimmer is the amplitude instability [4]. Figure 2.8 depicts the speech signal affected by these effects.

Harmonics to noise ratio feature describes a ratio between periodic and aperiodic in the speech signal or also as a noise that is present in a voice, particularly in the vowels.

2.4.4 Teager energy operator based features

This feature category is based on **Teager energy operator** proposed by Teager and Teager [43]. The foundation of their research was that the speech production process is the vortex flow changed by the current shape of the vocal tract. The airflow is responsible for the energy transmission from speakers to listeners. Non-linear Equation 2.2 describes the airflow change during the speech, where Ψ is the sign of the TEO operator and $x(n)$ is the sampled speech signal. Kaiser et al. utilized this non-linear property and proposed a set of features based on TEO suitable for emotion recognition, particularly for stress recognition [51]. This features set consists of **TEO-decomposed FM variation (TEO-**

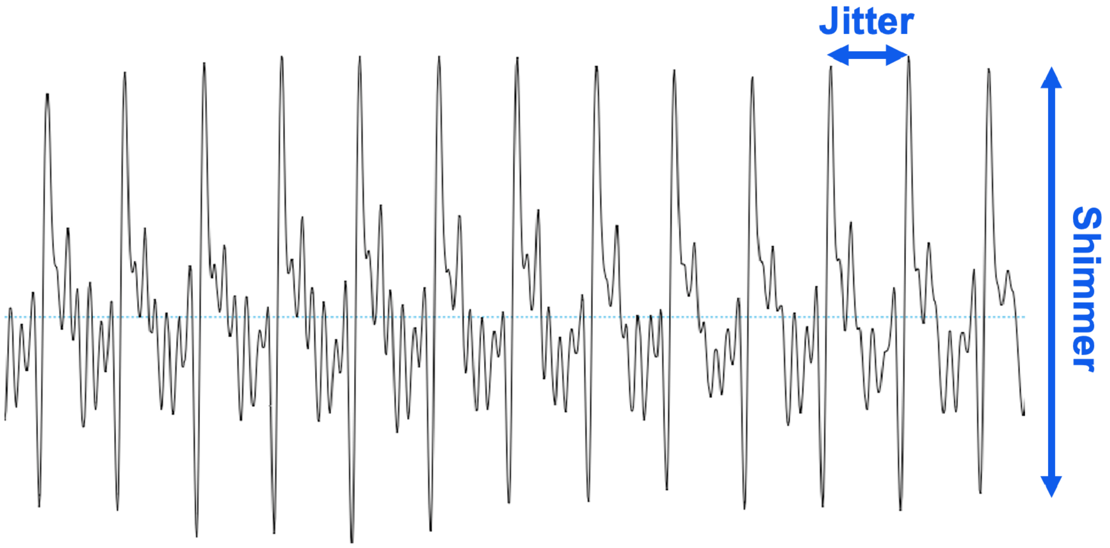


Figure 2.8: Jitter and shimmer example [9].

FM-Var), **normalized TEO autocorrelation envelope area (TEO-Auto-Env)** and **critical band based TEO autocorrelation envelope area (TEO-CB-Auto-Env)**.

$$\Psi\left[x\left(n\right)\right]=x^{2}\left(n\right)-x\left(n-1\right)x\left(n+1\right) \tag{2.2}$$

2.5 Classification methods

The last component of the SER system is classification responsible for assigning an emotion label or continuous value(s) to input recording. Since the start of the SER research, there were used, proposed and utilized different kinds of classifiers but none of them was sufficiently suitable and universal to cover all cases.

From a high-level overview of machine learning, classifiers need input in a structured form of the feature vectors, which consist of the values obtained from the feature selection and feature extraction. In the reviewed literature, SER research considers universal classifiers and there is no machine learning algorithm designed specifically for the needs of the SER [4, 46, 41, 21]. The most used classifiers for solving the SER task are **Hidden Markov Model (HMM)**, **Gaussian Mixture Model (GMM)**, **Naive Bayesian Classifier**, **Support Vector Machine (SVM)** and **K-Nearest-Neighbours (KNN)**. The rest of this section contains a description of the machine learning algorithm.

Hidden Markov Model is a supervised classifier that was comprehensively utilized in speech recognition, and the researchers successfully used it in the SER context, too. HMM is a sequential model, which can be drawn as the set of states and edges between them [4]. The transition process between states is based on probability distribution; thus, the choice logic of the following state is hidden from the observer. The current state at time step t depends only on one at the previous time step $t-1$.

Gaussian Mixture Model is a supervised generative model which consists of Gaussian components. Each component models the probabilistic distribution of one class, and they are described by their Gaussian parameters optimized during the fitting process. It is the version of HMM with only one state [18].

Naive Bayesian supervised classifier is based on the Bayes theorem [18]. It exploits the properties of the joint probability and assumes that each feature is independent of one another. Equation 2.3 represents the Bayes formula, where c represents a class label and $\mathbf{x}$ is a feature vector.

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \tag{2.3}$$

Support Vector Machine is the supervised linear algorithm that looks for a separating hyperplane with the maximal margin among the classes. The hyperplane lies in the middle of this margin and the samples that lie at the margin borders are called support vectors. If the problem could not be solved with a linear solution in the original representation, data are mapped to the new space with higher dimensionality where the algorithm can find the separating linear hyperplane.

K-Nearest-Neighbours is a parameterless supervised algorithm that assigns a class label to each sample, described by its feature vector $\mathbf{x}$, according to a majority vote of its k neighbours [46]. The distance between samples represented as vectors are measured, for example, with Euclidean distance (Equation 2.4a) or Cosine similarity (Equation 2.4b), where $\mathbf{x}$ and $\mathbf{y}$ are vectors in n-dimensional space and $\|\mathbf{x}\|$, $\|\mathbf{y}\|$ are their norms.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=0}^{n-1}(x_i - y_i)^2} \tag{2.4a}$$

$$\cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{2.4b}$$

2.6 Deep learning

The deep learning area is a relatively new approach but gradually begins to be ubiquitous. It is a subfield of machine learning based on the concept of the **artificial neural network (ANN)**, which tries to imitate human brain functionality [21].

The deep learning power lies in the automatic extraction of features. While traditional machine learning algorithms, described in Section 2.5, need a set of handcrafted features arranged in the appropriate structure, the deep learning models can process raw unstructured data and a feature extraction process taking place during the computation without the need of human intervention. However, in the SER context, it is a common approach to combine feature extraction methods together with deep learning to obtain better SER system performance.

This section provides a concise description of the models that are according to reviewed literature most popular: **Convolutional Neural Network (CNN)**, **Recurrent Neural Network (RNN)**, **Restricted Boltzmann Machine (RBM)**, **Deep Belief Network (DBN)**, **Deep Boltzmann Machine (DBM)**, **Autoencoders (AE)** and **Attention Mechanism** [24, 21, 46, 4, 28]. A model description includes the model type according to machine learning taxonomy (generative/discriminative, supervised/unsupervised), the main idea of the model and its basic structure.

Convolutional Neural Networks

CNN is a discriminative supervised model, whose main domain is the processing of the raw input data e.g. images and audio signals. [24]. CNN is built on a *feed-forward deep neural network architecture* that uses convolutional layers for extracting features from the input. These layers consist of filters that search for patterns according to their setting. To create a meaningful, but computationally efficient feature set, the downsampling method has to be performed in order to reduce the number of extracted features and at the same time to preserve discriminative information with a reasonable trade-off ratio. CNN uses for downsampling the pooling layers, where this trade-off is ensured. Usually, the last part of the CNN is made of dense layers dedicated to a classification of extracted features.

Recurrent Neural Networks

The RNN is a supervised discriminative model built on the *recurrent architecture*. Its purpose is to process sequential or time series data when there is a need to focus on the relationships and dependencies between them. The RNN mechanism works in a sequential fashion, which means that at each time step is processed one part of an input sequence. The output is computed as a product of the input vector and current context that is updated for further steps. The RNN structure is based on the recurrent unit, which computes output in the current time step and updates context information according to input. The improved versions of recurrent units used in the RNN are **Long Short-Term Memory (LSTM)** and **Gated Recurrent Unit (GRU)**. These methods improve the concept of RNNs and provide better performance and a more stable training process [21].

Restricted Boltzmann Machine

RBM belongs to the group of unsupervised models based on the generative approach. The main idea of the RBM is to model probability distribution over the neurons and by adjusting weights between them to learn the parameters of the distribution. RBM has two-layer architecture, where the input layer is called *visible* and the output layer is called *hidden*. The structure is represented as an undirected bipartite graph.

Deep Belief Network, Deep Boltzmann Machine

DBN and DBM models are built upon the idea of RBM. The DBN is deep learning architecture consisting of a cascade of RBMs, where the first two layers are connected with the undirected edges and the rest with directed edges in one way [24]. On the other hand, DBM is also made of RBMs cascade, but in this case, they are connected with undirected edges. DBM has one visible layer.

Autoencoder

Autoencoder is an unsupervised generative model whose architecture consists of two parts, *encoder* and *decoder*. The encoder realizes function f_θ, which encodes the input vector to a hidden structure or latent space of the model, and the decoder part represents function g_θ, which has the goal to reconstruct an original input from latent space [21]. The main aim of the AE is to compress input features with the encoder to as low dimensions as possible while preserving the reconstruction error of the decoder between its output and original input at a reasonable rate. AE is useful for feature extraction and dimensionality reduction because of its ability to compress input data with low information loss.

Attention Mechanism

The last described deep learning method is the Attention Mechanism. The idea behind this technique is the ability to look for the salient parts of the input data based only on the input data relationships. The model evaluates each input part with a weight according to its significance which is determined either by position in the input or by the input context. The Attention Mechanism was originally incorporated in the sequence-to-sequence type of RNN, where it helps to mark significant input parts in a given context [28]. Since the first proposal of the Attention Mechanism, researchers created various modifications that improved its original concept and enabled the emergence of other deep learning architectures, for example, the Transformer architecture [44].

2.7 Current SER state-of-the-art

Although the SER research is extensive for the past 20 years, there are still a lot of unresolved issues that even the current proposed solutions do not solve at a satisfactory rate [46]. This section analyzes the contemporary SER systems from the author's point of view; hence it focuses on the selected emotional database or corpus, the feature extraction and selection process and the classification model. Besides that, there is a discussion of the results of each system with the metrics used for the evaluation. Each system is reviewed with a critical debate to highlight its benefits and drawbacks. The final part of this section is dedicated to the challenges mentioned in reviewed literature and also the implications of described SER systems. This part serves as a guideline that should be followed by the SER system proposed in this work.

2.7.1 SER systems

This subsection comprehends recent methods that perform at the state-of-the-art level. There is a description of the overall system structure, what are system inputs (features), what is the basic working mechanism and what are the results with the provided evaluation methodology for each of the mentioned systems. After this description, the thesis author expresses his opinion if the results of the works fulfilled their objectives. He states what is missing, what is insufficient and how to make it better from our point of view as well.

Convolution-recurrent neural network

The first system with state-of-the-art results by Kumaran et al. [25] introduced a fusion of 40 MFFCs 39 with GFCCs that were fed to the neural network. The architecture of this network combines the advantages of the convolutional layers extracting salient input regions together with the LSTM units for modelling long-term relationships. The authors of this paper chose the RAVDESS dataset for training and testing purposes. In the evaluation, the authors provided total accuracy followed by a confusion matrix with precision, recall and F1 score values for each considered class. The accuracy of the system for the test data is above 70 %. The rest of the provided evaluation data implies that there is a high probability that the model was sensitive to *surprise* emotion and, on the other hand, the worst resolution shown for *fear* emotion. In the evaluation strategy, the thesis author lacks the comparison with other models, the implementation details of the model and the critical opinion of the proposed architecture from the authors of the paper.

From the author's point of view, the main acquisition of this work is the fusion of MFCCs and GFCCs that provided innovative ideas in the feature extraction step. On the other hand, there is no evidence in the evaluation part that exactly confirms that feature extraction is a key element of successful results. The author considers the disadvantage of this model the application of pure LSTM without any type of Attention mechanism, which enables to process only fixed-size sequences and forces a training process to be performed in a sequential way that implies an inability to use parallel computations to boost the speed of training.

CNN-BiLSTM with Stacked Transformers Layer model

Xianfeng et al. [45] introduced a SER system based on the CNN-BiLSTM and Stacked Transformer Layers (STL). The former-mentioned part is responsible for contextual representation extraction, and the latter part enhances the BiLSTM context vector resulting in emotional embedding. The output of the STL is forwarded to the fully connected neural network that generates discrete emotion labels. The authors used as the preprocessing method length unification of each recording to 7.5 seconds with the following framing, where each frame has a length of 25 milliseconds with a hop size of 10 milliseconds. The feature extraction and selection process consists of 40 MFCCs. The authors used the IEMOCAP emotion corpus for the evaluation of the proposed architecture with WA and UA metrics. According to provided results, the *weighted accuracy (WA)* score was 91.28 %, which is a 15.28 % improvement, and the *unweighted accuracy (UA)* score was 92 % with almost 20 % improvement compared with the second-best listed models. They also provided the confusion matrix, which implies a problem with a classification of the *neutral* emotion. For comparison, the authors provided WA for an ablated architecture without the STL that was equal to 72.79 % with which they demonstrated a significant impact of STL.

According to the author's opinion, the results of this work fulfilled stated expectations, and the application of the Transformer architecture played a key role in the proposed architecture. The evaluation process is comprehensive and well-structured. It provides a comparison of the proposed model results with previous state-of-the-art models, an examination of the number of the Transformers layers impact on the model performance, and an ablation study for the STL part together with how the model *grow up* with the number its part in terms of parameters number. Despite all these facts, the evaluation lacks other metrics such as precision, recall and F1 score that would depict which emotions are easily recognizable for the model and in contrast which of them make difficulties. Another drawback of this paper is the usage of only one dataset when the improvement for the IEMOCAP corpus was significant. The enhancement of evaluation with more datasets would bring more insights.

PCNSE-SADRN-CTC model

Zhao et al. [50] proposed an architecture of a neural network combining parallel convolutional layers and attention mechanism together with a dilated approach to classify emotions in a discrete fashion. They called it **PCNSE-SADRN-CTC** where PCNSE stands for *Parallel Convolutional Layers* with the squeeze-and-excitation technique, SADRN is an acronym for *Self-Attention Dilated Residual Network* and CTC is *Connectionist Temporal Classification.* The authors chose 40 MFCCs with delta-deltas values as input, which they obtained from recording partitioned to frames with a length of 25 milliseconds with a hop size of 10 milliseconds. According to the authors, the PCNSE block is responsible for extracting features from the input spectrogram, the SADRN block models relationships from extracted features and the CTC is used as a classifier. The model was trained and tested on datasets IEMOCAP and FAU Aibo Emotion corpus[2] evaluated with WA and UA metrics. The results were compared against the architecture from previous research papers that authors used at that time state-of-the-art architectures, particularly convolutional bidirectional LSTM (CNN-BLSTM), convolutional GRU (CNN-GRU) and bidirectional LSTM with CTC loss (BiLSTM-CTC). For the IEMOCAP dataset, the proposed architecture achieved WA 73.1 % and UA 66.3 %, which is an increase of 1 %, respectively 2 % compared to the second-best model (CNN-GRU). The UA of the PCNSE-SADRN-CTC model on FAU Aibo Emotion corpus was 41.4 %, which is 0.3 % less than the best-listed model (BiLSTM-CTC).

The novel idea of this architecture is the utilization of CNN with CTC in the SER. Although the results of the proposed model are not better compared with other listed models, the model architecture lacks sequential processing, which means that inference and training time demands are lower positively impacting the power consumption for both operations.

The evaluation part provides high-level information about the implemented model as a short table with all of the layer types with their parameters. The authors also include the results of the ablation study (e.g. *PCN*, *PCNSE*, *PCNSE-SADRN*) to examine the effect on individual components of the final model. These results prove that the incorporation of each part was reasonable. Another part to emphasize is the usage of the FAU-Aibo speech corpus, which is considered a natural speech database and provides a new point of view on how well could nowadays models perform in the real world. Overall, the author would welcome a more comprehensive and complex evaluation with more metrics and figures about the training process to show how well the model converged under the provided settings.

TIM-Net

One of the most novel experimental approaches was proposed by Jiaxin et al. [49], who created neural network architecture with the name **Temporal-aware Bi-direction Multi-scale Network (TIM-Net)**, to remove shortcomings in the long-range dependencies modelling and fixed temporal scale[3]. Its main building block is *Temporal-aware block*, which serves for feature extraction and modelling relationships between the input data. This block is a subnetwork, whose core element is the dilated causal convolutional layer modelling temporal-aware features, which are formed by temporal attention layers. The network contains an arbitrary number of these blocks. The term *Bi-direction* refers to the fact that the network process input from both ends to integrate complementary information from *past* and *future* together. Hence, for each block that processes data in a forward manner exists block processing data in reverse order and their results are connected to produce one temporal-aware feature. The last expression, *Multi-scale*, refers to the part of the network, where all temporal-aware features are fused as their linear combination. This network is designed to predict discrete emotion labels.

[2] https://www5.cs.fau.de/en/our-team/steidl-stefan/fau-aibo-emotion-corpus/

[3] The paper is still under review by the time of writing this thesis.

The network input consists of the spectrogram from 39 MFFCs that are obtained from framed recording. Each frame has a length of 50 milliseconds with a shift size between frames of length 12.5 milliseconds. On each frame is applied Hamming window.

The evaluation methodology proposed by the authors involves performance testing on six emotional datasets (Emo-DB, IEMOCAP, RAVDESS, SAVEE, CASIA and EMOVO) where they monitored the weighted and unweighted accuracy. For each dataset, there is a table with mentioned metrics for four models including TIM-Net. For every dataset, the TIM-Net architecture outperforms others. According to the authors, the average gain measured for all datasets in the UA is 2.34 % and in WA 2.64 % compared to the second-best listed models. For the datasets CASIA and Emo-DB, the authors provided also the confusion matrix. As part of the evaluation, the model was tested in a cross-corpus manner, or in other words, the authors trained the model on one database and tested it on the other for demonstration of generalization ability. For this purpose, the authors utilized only five emotions (*angry*, *fear*, *happy*, *neutral* and *sad*) that represent the label intersection of mentioned datasets. This model obtained the best results in used metrics (WA, UA) among all compared models. The improvement is around 2 %.

The authors also conducted the ablation study, where they examined the impact of removing individual components on the watched metrics. The result of this study shows that all components contribute to overall model performance with the almost same value (around 3.5 % for both metrics).

From the thesis author's point of view, the proposed SER system obtained the most promising results from the reviewed literature. Researchers evaluated the proposed architecture on six emotional databases, showing that the dynamic fusion of multi-scale features performs better than features with fixed scales. It is not clear how the authors were aiming to improve the modelling of long-range dependency since they did not define what consider long-range dependency. There are a few facts worth noting. First, the model testing was realized on more than two emotional databases followed by the cross-database multilingual test method used in this paper. The second is the ablation study, where the authors show the importance of each system component. On the other hand, there is a space for improvement in the selection of evaluation datasets by adding at least one based on natural speech.

The unanswered question is the complexity of the proposed model in terms of parameter count together with its resource needs for training. This issue arises from the fact that the authors did not provide this kind of implementation detail.

Model	*Database(s)*	*Preprocessing*	*Features*	*Results*
C-RNN	RAVDESS	-	40 MFCCs, 39 GFCCs	70 % / -
CNN-BiLSTM	IEMOCAP	Framing with the frame length of 25 ms and hop length of 10 ms	40 MFCCs.	92 % / 91.28 %
PCNSE	IEMOCAP, FAU-AEC	Framing with the frame length of 25 ms and hop length of 10 ms, Hamming window applied on each frame	40 MFCCs enhanced by their delta-deltas.	66.3 % / 73.1 % for IEMOCAP, 41.1 % / - for FAU-AEC
TIM-Net	EmoDB, IEMOCAP, RAVDESS, SAVEE, CASIA, EMOVO	Framing with frame length of 50 ms and hop length of 12.5 ms, Hamming window applied on each frame.	39 MFCCs	2.34 % / 2.36 % (avg. improvement)

Table 2.3: A brief overview of current state-of-the-art SER systems described in subsection 2.7.1. The *Results* column contains values in form of **unweighted accuracy* / *weighted accuracy**. Because TIM-Net used 6 datasets, there is provided only average improvement from all results.

2.7.2 Challenges in the SER research

Although results from the current state-of-the-art SER systems presented in the previous subsection 2.7.1 are promising, there is still a lot of work to even approach the human emotion recognition abilities. First, it is discussed the long-term problems that are in SER from almost the beginning of its research, and then there is a focus on the issues that come with deep learning.

Size of datasets

The size of the current datasets for SER is significantly smaller than for other recognition tasks. For example, ImageNet[4], one of the most used datasets for image classification tasks, contains more than 14 million images. For comparison, the EmoDB dataset contains 550 recordings, and the RAVDESS dataset contains 1440 recordings. After deep learning entered the SER field, this problem became even more recognizable because deep learning models need more training data than traditional machine learning algorithms.

Datasets labelling

The next challenge related to the previous idea is the creation of new datasets for SER needs. The biggest problem according to the reviewed literature is the human recognition accuracy, which is below 90 % [46, 21, 4]. This problem arises mainly from the subjective perception of emotions by humans, who labels data samples in the dataset.

[4] https://www.image-net.org/

Multilingual SER

Another unsolved task is overcoming differences in emotional expression in various languages or cultures. Each language has its own patterns in the expression of emotions. One of the discussed SER systems, TIM-Net, was evaluated on multiple databases that contained recordings in a different language, however, the performance was far below the results when only one language was considered.

Raw speech utterances as model input

This topic considers the SER systems that do not include preprocessing stage and therefore make the whole process more straightforward. It is interesting mainly with deep learning algorithms because some are designed to process raw input data. Although in other speech or audio fields, such as speaker recognition or audio sound classification deep learning algorithms reach state-of-the-art results with raw data, this approach is not successful in the SER.

Any of the listed state-of-the-art solutions did not consider raw speech as the input for the model. After specifically dedicated research to this topic separately, the following solutions were found. Latif et al. [27] implemented a deep learning model that processed raw audio speech. They evaluated implementation on the datasets IEMOCAP and MSP-Podcast [31] with unweighted accuracy as the metric. The obtained results were 59 % respectively 52 %, which we consider as average compared to models from the previous section. Mustaqeem et al. [33] created a model that processed normalized speech signals. They used databases IEMOCAP and Emo-DB for benchmarking and obtained 73 % and 92 % of unweighted accuracy. Even though the results are promising, this research did not find any other recent similar solutions with such results.

Transfer learning

This challenge relates to the deep learning paradigm and it is a way how to *transfer knowledge* from the previously trained large models to the contemporary ones [21]. There are good examples of transfer learning in other fields, where deep learning is dominant right now, e.g. image or speech recognition, where big models were trained on a huge amount of data and their training took weeks or months and further they were *fine-tuned* for more specific tasks. Again, this topic is also connected to the size of current datasets, which does not allow a pretraining general model for speech emotion recognition.

Deep learning models size

This topic is pragmatic because most of the SER use cases include real-time processing on a device with limited resources. The deep learning models are usually large and need a lot of space and computation power to make an inference. Due to this fact, the SER models based on this paradigm should consider the ratio between model resources necessary for proper functioning and model performance.

Deep learning environmental impact

This deep learning issue arises from the introduction of larger and more complicated models. With the increasing size of the models, the requirements for the training resources grow as well, which has an ultimate impact on power consumption and the production of greenhouse gases. According to Strubell et al. [40], the complete training of the large natural language processing model produces more greenhouse gases than the average car during its lifetime. Lacoste et al. [26] proposed a method for calculating that takes into account the hardware used for training, the number of hours needed to perform training, the cloud provider and world region, where is the training device. Anthony et al. [5] created the tool with the name *Carbontracker*, especially for deep learning purposes to track and predict the carbon footprint of the models.

2.8 Chapter summary and its implications

This chapter extensively analyzed all the necessary components to build a SER system. It included the emotion theory necessary for a complete grasp of the field followed by the taxonomy of emotions. Further sections were dedicated to speech signal overview, methods suitable for its preprocessing and features utilized for SER purposes. Then, the focus was redirected to machine learning and deep learning algorithms used in SER. Each algorithm was described to show its principal functionality.

After the overview of SER components, the chapter continued with the section about current state-of-the-art systems from this field to review the methodology of how are they designed, implemented and evaluated. For each system, the author provided his opinion, highlighted innovative approaches and ideas and mentioned drawbacks, which he considered as resolvable. This review implies current trends connect cepstral features with the chosen deep learning algorithm. The very popular system that combined convolutional and recurrent neural networks, however, there was also mentioned cases with the Attention mechanism and the Transformer neural network. The trend in the evaluation of SER systems is to choose one or two emotional corpora, perform the model fitting process and provide weighted or unweighted accuracy, sometimes enhanced with the confusion matrix.

The last part of this chapter stated a list of the contemporary challenges in SER. It indicates that this topic is still not entirely solved despite progress in the fields of psychology, signal processing and machine learning; hence there are unresolved problems and untested technologies that need future investigation. The next part of this thesis tries to target some of the mentioned challenges and make a contribution to their solution.

Chapter 3

Proposed methodology

This chapter analyzes a solution concept that is a blueprint for the implementation. The baseline of this concept is **Vision Transformer (ViT)** architecture, derivation of Transformer architecture. Therefore, for a complete understanding of the whole solution, it is necessary to study ViT as well as Transformer.

After the examination of mentioned deep learning concepts, further work is focused on the possibilities to use ViT in terms of audio and speech processing. The final section of this chapter devotes to the suggested plan for implementation part.

3.1 Transformer neural network

The idea of a Transformer was first proposed by the Vaswani et al. [44]. Its principal purpose is to process sequential inputs. The original model was used for the machine translation task (part of natural language processing), where it performed at the state-of-the-art level according to the results in [44].

The Transformer architecture is depicted in Figure 3.1. It is *encoder-decoder* deep neural network architecture, which does not contain convolutional or recurrent units. According to the mentioned figure, each block occurs N times in the architecture. Both blocks have common parts, particularly **positional encoding** and **multi-head attention mechanism**. This section is dedicated to the explanation of these two components, whereas they form the skeleton of each of these blocks. It is important to mention that each of the blocks can create a standalone neural network. The models that contain only encoders are primarily used for encoding and are used for another task (e.g. classification). On the other hand, models that consist only of decoders are used for the prediction of the next value in a sequence based on the previous input and thus are considered autoregressive.

The first step of the input processing is its encoding in a way that assigns each input part, called a token, at index i in a sequence the length of N, a value that determines its position. This step is called *positional encoding (PE)*, and its output is called *embedding.* The transformation process of the input values to the embeddings is expressed by equations 3.1 and 3.2. Each equation takes as the input the token denoted as *pos* with index i, which needs to be adjusted to create position awareness that determines its order in a sequence for further network processing. The parameter d_{model} is the size of the embedding, which was for the original Transformer model set to value 512. These embeddings are further connected to input matrix $I \in \mathbb{R}^{N \times d_{model}}$.

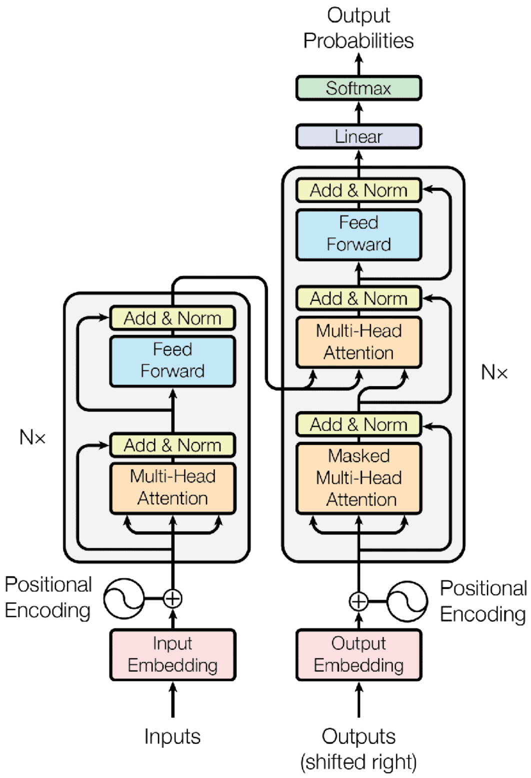


Figure 3.1: Transformer neural network architecture [44].

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{3.1}$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i+1/d_{model}}) \tag{3.2}$$

After the input is encoded to the suitable form, it is further processed by the encoder block, which uses **a multi-head self-attention layer** with a feedforward layer. The examination of self-attention is related to the prior work in Section 2.6, where the author briefly described the attention mechanism. The self-attention mechanism can be described as the computation process, which computes a score among the segments of input to express their relationships. In Figure 3.2a, there is a particular type of the self-attention mechanism, which is called **scale dot-product attention** and is used in the original model. Its inputs are the matrices $Q \in \mathbb{R}^{N \times d_{model}}$, $K \in \mathbb{R}^{N \times d_{model}}$ and $V \in \mathbb{R}^{N \times d_{model}}$, which are results of linear projections of input matrix I with projection matrices $W^Q \in \mathbb{R}^{d_{model} \times d_{model}}$, $W^K \in \mathbb{R}^{d_{model} \times d_{model}}$ and $W^V \in \mathbb{R}^{d_{model} \times d_{model}}$. The computation of these matrices is depicted by the equation 3.3.

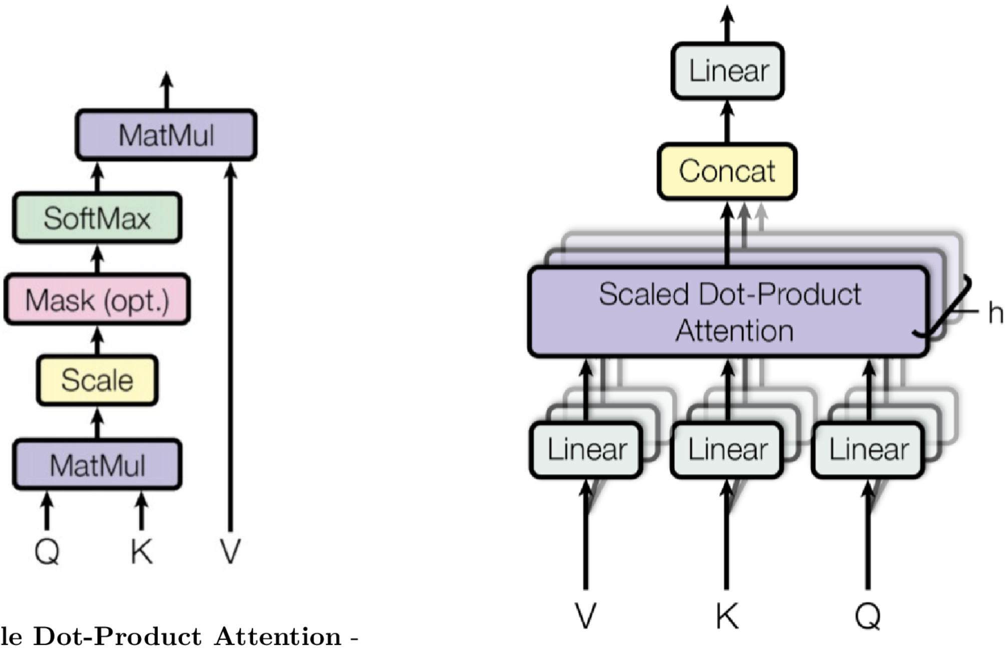
$$Q = I \times W^Q \tag{3.3a}$$

$$K = I \times W^K \tag{3.3b}$$

$$V = I \times W^V \tag{3.3c}$$

The final matrix $A \in \mathbb{R}^{N \times d_{model}}$ of the self-attention is expressed in the equation 3.4. The matrices W^Q, W^K and W^V are learnable parameters that are adjusted during the training of the neural network.

$$A = softmax(\frac{QK^T}{\sqrt{d_{model}}})V \tag{3.4}$$



(a) **Scale Dot-Product Attention** - version of the self-attention mechanism used in original Transformer architecture.

(b) Multi-head attention mechanism.

Figure 3.2: The Attention mechanism components [44].

The concept of the multi-head attention mechanism, depicted in Figure 3.2b, can be treated as the extension of the self-attention mechanism. The term *multi-head* refers to the fact that input is processed by multiple self-attention blocks enabling the model to focus on different parts of the input at the same time (in parallel). As shown in Figure 3.2b, there are h blocks of the self-attention called *heads*, and each of them processes the input independently of one another. In particular, there are h heads and each process the input of size that is equal to a proportion of the embedding size and the number of heads (d_{model}/h). It implies that the size of the projection matrices of each head is reduced by the factor of h.

After each head produces its results, they are concatenated together to the embedding of the original size d_{model}. The final output of the multi-head attention is produced by the fully-connected layer that has the purpose to link information among concatenated embeddings. In the original paper the authors used 8 heads implying that each one processed embedding of size 64 $(512/8 = 64)$.

3.2 Vision Transformer

Dosovitskiy et al. [15] presented a neural network architecture based on the Transformer architecture designed to handle computer vision tasks. The inspiration for ViT design came from the success of the Transformers in NLP tasks, especially a neural network called BERT [14]. Figure 3.3 depicts the original architecture of ViT, which consists only of the encoder part of the Transformer.

This architecture provides the linear transformation layer that projects the input image to fixed-size patches. They are converted to the $1D$ tokens in the next step. Another property that ViT brings is the extra learnable token that is denoted in Figure 3.3 as `[class]`. It serves classification purposes and is inspired by the BERT model. It is connected to the beginning of the token sequence and represents a summary of the whole image. At the start of the training, `class` is initialized with random values that are adjusted during the training. Its output plays a key role as it is connected to the classification head, which returns the final network result.

The first stage in the model inference is a partitioning of the input image with shape $H \times W \times C$ to patches $\mathbf{x}_p$ with shape $P^2 \cdot C$, where (H, W, C) represents height, width and number of colour channels of the image, P is the size of a patch, $N = (H \cdot W)/P^2$. After this step, each patch $\mathbf{x}_p$ is fed to the linear projection layer, where the shape of each patch is adjusted to the value D that is equal to the token size. This linear layer is composed of learnable parameters tuned during the fitting process. After projection, `[class]` token is inserted at the start of the sequence. Then to each token is added a positional encoding value, which results in the embeddings sequence ready to enter the encoder block. Its only used output is `[class]` token output that is further fed to the MLP head, which produces the final classification result.

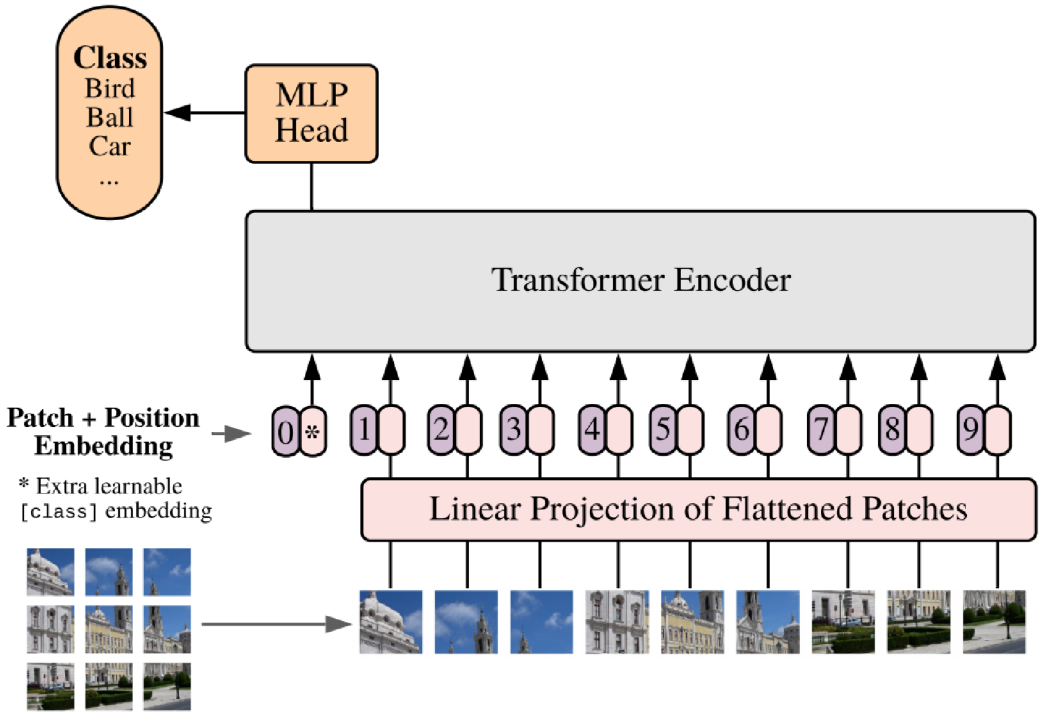

Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure 3.3: Vision Transformer neural network architecture [15].

The authors of [15] implemented and evaluated three types of Vision Transformer, which differ in the size of an image patch, the number of heads of multi-head attention, the number of encoders and the size of the final classification module. The evaluation was done on the image classification benchmarks such as ImageNet, CIFAR-100[1], Oxford-IIIT Pets[2] etc. The observed metric was accuracy. The obtained results were compared to the models **ResNet152x4** and **EfficientNet-L2**. Each of the ViT models reached better results than the other compared models. The authors also observed the training time measured in TPUv3-core-days, and each ViT model needed less training time than their opponents. In two cases was a difference in the order of magnitude.

3.3 Audio Spectrogram Transformer

Based on the ViT architecture, Gong et al. [19] created **Audio Spectrogram Model (AST)** utilized for the classification of audio recordings based on their Mel spectrograms. Figure 3.4 depicts its overall structure and outlines its working principle, which mostly matches the ViT. The original idea was to create a model based on transformer architecture that is able to process the audio and speech data without convolutions. Its additional interesting property is the transfer learning approach because the AST uses in its core the pretrained ViT on the ImageNet dataset.

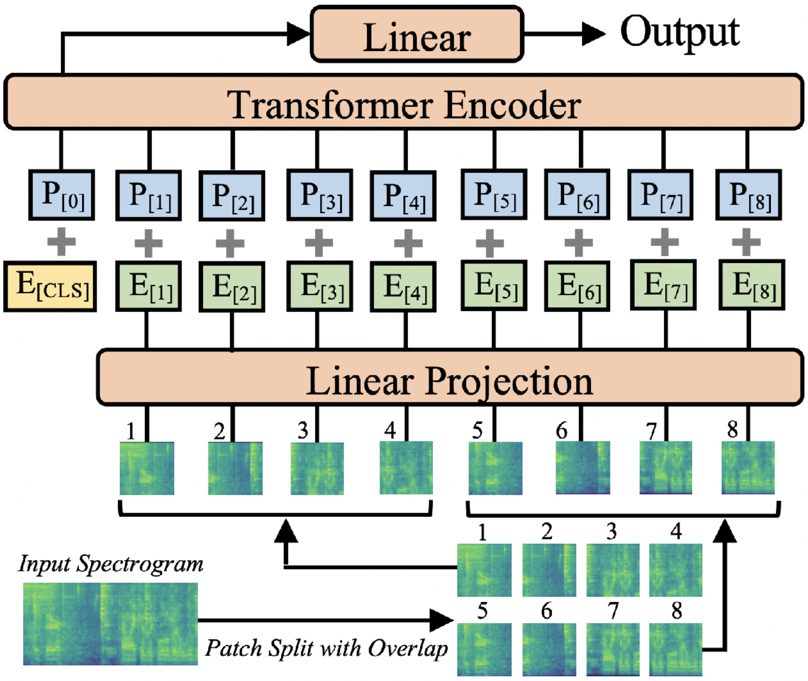


Figure 3.4: Audio Spectrogram Transformer architecture [19].

The ViT accepts the RGB images that are adjusted to fixed shape (e.g. 224 × 224 or 384 × 384) which is not typical for the mel spectrogram, which tends to vary in both dimensions. Therefore, the first step is the reduction of the input channel number from three to one. This is done by averaging the weights of the linear projection layer.

[1] cs.toronto.edu/~kriz/cifar.html

[2] https://www.robots.ox.ac.uk/~vgg/data/pets/

Another issue with the shape is connected to positional encoding. The ViT deals only with the sequences of fixed sizes even though it is a Transformer model. This is caused by the utilization of learnable positional encoding that is pretrained. As mentioned earlier, the mel spectrograms tend to vary and there is a need to choose the size of the mel filter bank and the number of frames that should spectrogram contain in order to adjust positional encoding to demanded shape. To account for this fact, the authors of AST proposed a technique, which modifies positional encoding with respect to the selected values, particularly, each dimension is reduced or lengthened with *bi-linear interpolation* method. Let's suppose for demonstration purposes ViT model with patch shape 16×16 that was trained with images of shape 224×224. Then, each image is partitioned to 14×14 and projected to a token of size 768. This implies that positional encoding generates value for 192 tokens to create position-aware embeddings. If the same model should be used as AST with inputs of size 1024×128 (64×8 patches), the positional encoding needs to be adjusted to be able to process sequence of size 512, where horizontal (x) part of positional encoding is reduced to the value 8 and vertical (y) part is interpolated to 64.

As mentioned in the section beginning, the Gong et al. [19] proposed model was pretrained on the ImageNet dataset and fine-tuned on the following datasets AudioSet[3], ESC-50[4] and Speech Commands V2[5]. While the results of the AudioSet did not reach state-of-the-art, the model performed better than state-of-the-art models on the rest two datasets. Besides these results, the model shows that it is possible to utilize the pure transformer model pretrained on the image dataset, fine-tune it on an audio dataset and obtain intriguing performance.

3.4 Architecture of the proposed SER system

This section comprehends the description of the proposed SER system together with the evaluation methodology. The system's explanation includes the emotional model, the database(s), preprocessing and feature extraction methods and the deep learning model.

The primary goal of the designed system is to tackle the following challenges mentioned in 2.7.2:

- **pure Transformer model** - use model that is based only on the Transformer architecture without any convolutions and recurrent cells as in other recognition fields that deep learning emerged in,
- **transfer learning** - use pretrained ViT model,
- **environmental impact** - track the greenhouse effect produced during the fitting process.

3.4.1 Emotional model and databases

The thesis author decided to choose a discrete (categorical) emotional model because of the properties such as good interpretability and comprehensibility. Another reason, for such a decision, is the wide offer of the datasets labelled discretely.

[3] https://ieeexplore.ieee.org/abstract/document/7952261

[4] https://dl.acm.org/doi/10.1145/2733373.2806390

[5] https://arxiv.org/abs/1804.03209

Regarding datasets, the author wanted to do a comprehensive evaluation that includes five different emotional corpora, particularly Emo-DB, EMOVO [12] and RAVDESS. This set contains three languages (English, German and Italian). The usage of more emotional corpora results in a more robust estimation of SER system capabilities.

3.4.2 Preprocessing and feature extraction methods

Due to using the skeleton of the deep learning model from [19], the feature extraction stage has to correspond with the process described in this paper. Its results are fixed-size mel spectrograms. In Figure 3.5 is this part depicted by the **Mel spectrogram extractor** part.

Each recording is divided into frames of length 25 milliseconds with 10 milliseconds hop size. On each frame is applied the windowing function with Hanning window. The Short-time Fourier transform (STFT) is used for the transformation of each frame from time to frequency domain and its results are filtered by a mel filter bank with 32 filters, which values are representing the mel spectrogram.

3.4.3 Deep learning model

The proposed deep learning model consists of two logical parts, particularly the pretrained AST model and the custom classification part. In Figure 3.5 are these parts depicted as **Audio Spectrogram Transformer** and **Classification head**.

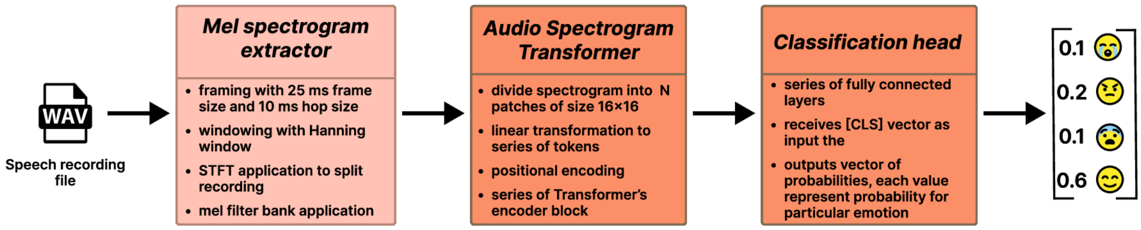

WAV
Speech recording file
Mel spectrogram extractor
• framing with 25 ms frame size and 10 ms hop size
• windowing with Hanning window
• STFT application to split recording
• mel filter bank application
Audio Spectrogram Transformer
• divide spectrogram into N patches of size 16×16
• linear transformation to series of tokens
• positional encoding
• series of Transformer's encoder block
Classification head
• series of fully connected layers
• receives [CLS] vector as input the
• outputs vector of probabilities, each value represent probability for particular emotion
0.1
0.2
0.1
0.6

Figure 3.5: The proposed SER system as a pipeline of modules. Its result is only illustrative.

The AST part is taken from the [19] as a pretrained model and it serves as a backbone of the model. The classification head is represented by a fully connected layer series with the output layer that returns the probabilities of emotion labels. The parameters of the classification head are fine-tuned during the fitting process on the mentioned datasets.

3.4.4 Evaluation methodology

The evaluation methodology is based on two dimension. The first dimension regards the quantity of dataset used for proposed model training. The second dimension regards the results quality of each dataset. As inspiration for quantitative approach, the review state-of-the-art model Tim-Net [49] was trained on six independent emotion databases. On the other hand, for each dataset is provided only weighted and unweighted accuracy and lacks the of other classification metrics, which causes insufficiency in qualitative dimension.

An opposite case is an approach in the paper that proposed C-RNN [25] that used only one dataset for evaluation, but on the other hand, it provides not only accuracy metric but also the whole confusion matrix, loss function progress curve and accuracy progress curve.

Chapter 4

Implementation

The proposed system implementation has its origin in the Audio Spectrogram Transformer architecture described in Section 3.3. This chapter is dedicated to the implementation details of the entire system, which includes speech recording preprocessing and augmentation, model implementation, model training techniques, and evaluation process. Python was selected as the main programming language due to its strong connections with deep learning. The author of the thesis chose the PyTorch library[1] and its ecosystem for input processing, training, and testing the model because of its stability, reliability, regular maintenance and development, a large ecosystem with a lot of implemented features and methods, and last but not least, its increasing popularity among academic works in recent years.

The implementation description is divided into sections that follow the flow of the resulting system. The description begins with the preprocessing of speech input data, which produces a dataset ready for use in model training. The next subsection is devoted to the model implementation details with a consecutive description of its training loop preparation. The last sections comprise the implementation of evaluation metrics

4.1 Speech recording preprocessing

The preprocessing of input data is divided into two parts: the transformation of the speech recording signal into the mel spectrogram, and data augmentation, which serves to add some random noise or obfuscation to the data samples that changes them in a way that preserves their core properties while enlarging the training dataset. The result is the implementation of a particular class that serves as the data source for the model during training.

The pipeline for mel spectrogram extraction consists of a few steps. The first step is to load the audio file with the appropriate sampling rate (expressed in Hertz [Hz]), which determines the number of samples (data points) of amplitude that represent one second of recording. The selected value for this purpose is 16000 Hz, as it strikes a balance between quality and the recording size (the higher the sample rate, the higher the quality but the higher the memory requirements). The second stage of the mel spectrogram construction pipeline is framing and windowing. Inspired by other works that used these methods, the implementation in this work partitions the loaded recording into frames of length 25 milliseconds with a 10 millisecond hop (shift) size, or frames of size 400 samples with 160 sample hop length when the sampling rate is 16000 Hz. The frame size value is chosen intentionally because research has shown that speech is invariant for twenty to thirty mil-

[1] https://pytorch.org/

liseconds [4]. The hop size value is set to 10 milliseconds to create a big enough overlap between neighbouring frames to prevent information loss during the windowing when the frame ends are smoothed by a windowing function to eliminate discontinuities that could influence the results of further processing (detailed explanation of framing and windowing is in Subsection 2.3.3). The implementation in this thesis utilizes the Hanning windowing function, one of the most widely used functions in other works during the speech preprocessing stage [4].

The third pipeline step consists of Short-time Fourier transformation (STFT). In this step, each frame is transformed from the time to frequency domain. The transformed frames are then filtered with a mel filter bank involving n mel band filters. The value n is a hyperparameter that can be set according to demand. After the mel filter bank is applied to the frames, the construction of the mel spectrogram is done. In this phase, it has a shape of $n \times m'$.

In the last stage, the spectrogram is adjusted to the selected number of frames (mel spectrogram columns), denoted as m. This modification is required to unify the length of each mel spectrogram to construct mini-batches with the same shape (more about training in Section 4.3). If $m < m'$, then the $m' - m$ number of columns is removed from the right side of the spectrogram. In the other case, when $m > m'$, new columns filled with neutral values are added to the right side of the spectrogram. Again, m is a hyperparameter that can be selected according to preferences. All in all, the final mel spectrogram has a shape of $n \times m$, where n represents the number of rows (mel bands), and m implies the number of columns (time frames).

To augment the dataset, there is implemented five augmentation types that work based on Bernoulli random variable X, where the probability for $X = 1$, denoted as p, is an arbitrary value that needs to satisfy the following criterion: $0 \leq p \leq 1$. This implementation considers $p = 0.25$. The value is chosen empirically and it is one of the topics for further discussion.

There are implemented five different augmentation strategies. Three are focused on raw speech recording and the remainder two modify the mel spectrogram. The recording augmentation includes adding random white (gaussian) noise, which is represented as a vector with the same shape as speech recording. Values in this vector are random numbers generated from the Gaussian distribution with zero mean and the variance argument is equal to the recording's variance. Before the noise is added to the original recording, it is multiplied by *noise factor* (small decimal number) serving as regularization of noise impact. The thesis implementation considers three noise factors (0.1, 0.15, 0.2). When augmentation is applied, one of the noise factors is randomly chosen as a regularization value.

The second augmentation technique is speed change. The main idea is to either stretch or shrink the original recording by chosen factor. The implementation considers four different factors (0.85, 0.9, 1.1, 1.15) and when the augmentation is applied, one of them is randomly chosen.

The next method for artificially increasing the number of recordings by editing their signals is *room impulse response* which adds an echo effect and creates the feeling of conference room sound. It is implemented according to the tutorial from the official PyTorch documentation page[2].

[2] https://pytorch.org/audio/main/tutorials/audio_data_augmentation_tutorial.html

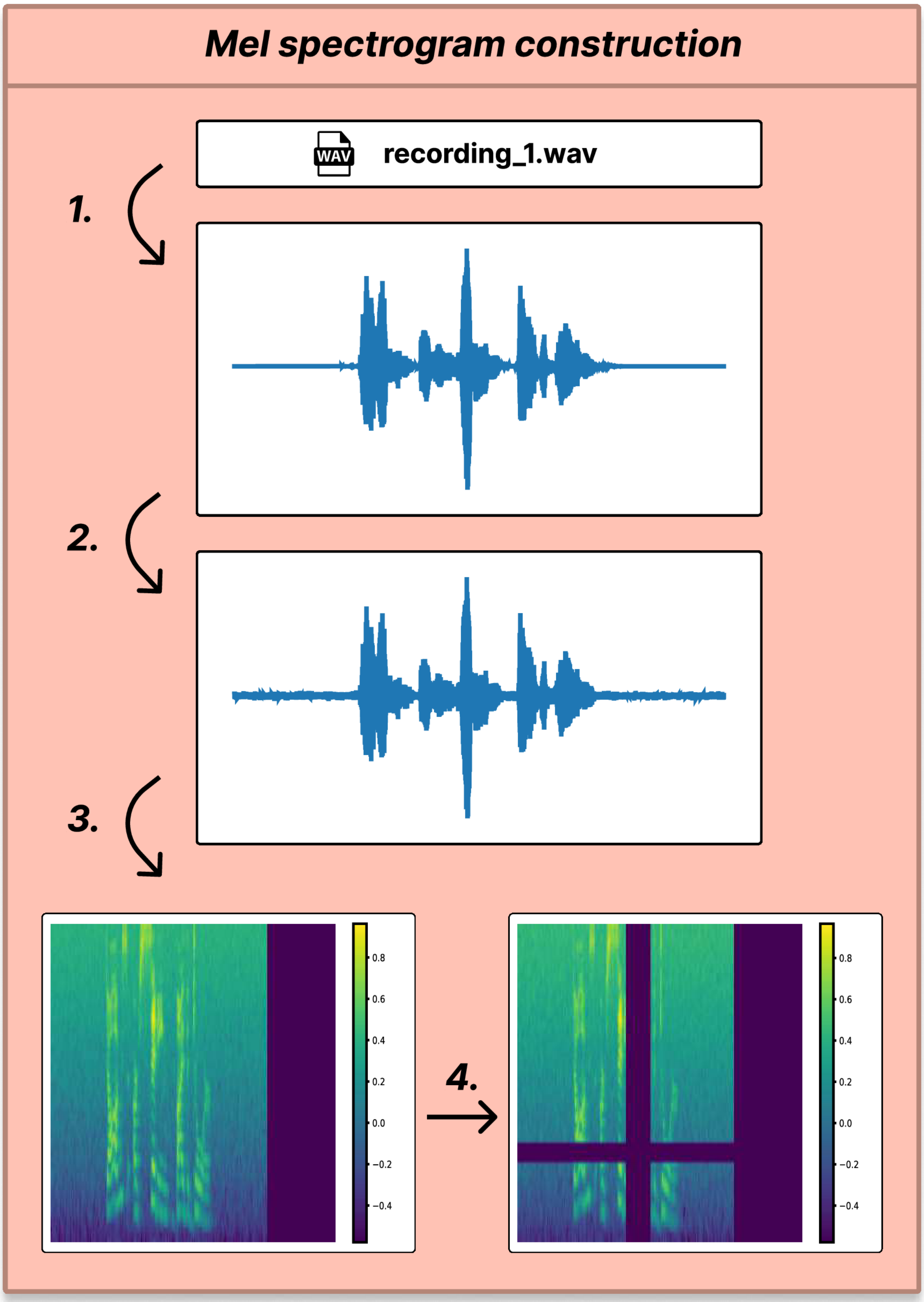


Figure 4.1: Flow of mel spectrogram cosntruction.
Step 1. represents loading recording with given sampling frequency.
Step 2. denotes signal augmentation of original recording.
Step 3. shows transformation of augmented signal to mel spectrogram.
Step 4. is mel spectrogram augmentation.

The fourth and fifth implemented augmentation methods are focused on adjusting the mel spectrogram. They are called time and frequency masking. Their purpose is to change a certain portion of the time respectively frequency domain to the value symbolizing that either in a given time is an amplitude of the original signal equal to zero or for given a frequency band that it is not present in the original signal. Both augmentations strategies receive *mask_param* argument determining the maximum value they can mask in a mel spectrogram. The actual value is computed in runtime and it is sampled from interval $< 0, mask\_param)$, which has uniform probabilistic distribution. The implementation of time and frequency masking is inspired by the PyTorch documentation[3].

4.2 Model implementation

Based on the description in subsection 3.4.3, the implementation utilizes the Audio Spectrogram Transformer (AST) by Gond et al. [19], which is freely available in GitHub[4] under the 3-Clause BSD license[5], which allows the author to use this code for purpose of this thesis. The thesis implementation contains minor changes without any fundamental impacts compared with the original AST implementation and therefore this section is focused more on the selection of one of offered pretrained ViT models.

According to the original paper that proposed AST, its implementation utilizes the standard ViT architecture proposed in [15], which was properly described in Section 3.2. In fact, this implementation uses the models from the library PyTorch Image Models (Timm)[6] . The provided code offers four types of pretrained ViT, which differ in their number of trainable parameters that directly implies their size and memory requirements. The list of offered models contains **tiny224** (≈ 5.8 million parameters), **small224** (≈ 22.2 million parameters), **base224** (≈ 86.8 million parameters) and **base384** (same number of parameters as base224). Obviously, the model name represents its relative size and the number denotes the image size that was used during ImageNet pretraining phase, e.g. type tiny224 was pretrained on the images of size 224 × 224. Since the utilization of SER models is focused on small or even embedded devices, the thesis author decided to choose the model of tiny224. Another reason to use type tiny224 is the size of available datasets that do not demand a bigger model type to obtain better performance. The further implication of this decision is the potentially less time-consuming training of the model, which is highly influenced by the number of trainable parameters.

The selected classificator for the pretrained model is a multi-layer perceptron that consists of one normalization layer and an output layer of shape equal to the number of considered emotions for a given dataset.

4.3 Training implementation

This implementation follows the backpropagation algorithm that is currently state-of-the-art in the training of neural networks. The training maintains this flow:

1. forward propagation of inputs,

[3] https://pytorch.org/audio/master/generated/torchaudio.transforms.FrequencyMasking.html

[4] https://github.com/YuanGongND/ast

[5] https://opensource.org/license/bsd-3-clause/

[6] https://github.com/huggingface/pytorch-image-models/tree/main/timm

2. compute loss (error) as the difference between predictions and expected values (ground truths),
3. backpropagate error throw the network in order to adjust parameters.

Loss function and optimizer definition

First, there is a need to define the function responsible for computing loss (error) as the difference between the outputs of the neural network and expected values. Because dealing with classification task, it is selected **Cross Entropy** function designed for this purpose. The algorithm **Adam** is selected as a learning rate optimizer that takes care of adjusting the learning rate value during the training process.

Dataset partitioning

Before the beginning of training, the dataset is divided into 10 equal parts called folds according to the method **stratified k-fold cross-validation**. Each fold contains two subsets of data samples split in a ratio of 9:1. The first subset contains data samples for training and the second for testing. The logic behind such partitioning is to have a subset of data samples that serve exclusively for training and the other one for testing purposes and final evaluation. In the code, there is used class `StratifiedKFold`[7] to achieve described partitioning.

The implication of this approach is the need to create 10 different models and run 10 training sessions independent of each other. The results of this way should secure enough evaluation data to decide, if the model performance is dependent on dataset partitioning or not.

Each subset is then further divided into mini-batches of size n, which is an arbitrary number. The purpose of the mini-batch is to tackle the problem of model parameter update frequency. The lower the n is, the higher frequency of parameter updates, but it can lead to some local optimum of loss followed by the suboptimal model results. On the other hand, the low frequency of updates can lead to slowing the training process and the high memory requirements. To handle the trade-off between these risks, the n is set to 32.

Training loop

The implementation of the training loop follows the steps mentioned at the beginning of this section. It includes an inner mini-batch loop, which first computes neural network inference for input data from the mini-batch, computes loss value and backpropagates the error via the network. Then the loss value together with reached classification accuracy for this mini-batch is stored to epoch evidence. Both values are preserved in the mentioned evidence and at the end of the epoch is logged their mean as model performance representation of the current epoch.

During the training loop is also tracked the carbon footprint since the most computationally-extensive operation are performed in it. For the collection of carbon footprint data is utilized the library Codecarbon[8]. The tracking frequency is set per epoch and values are stored to global evidence for evaluation.

[7] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

[8] https://codecarbon.io/

The tracker class provided by this library automatically detects the type of computational resources and measures their consumption of power in kilowatt-hours (kWh). The only watched computational resoruce is grahpical processing unit. The energy consumption of other computational units is not tracked because the training is run on computer with shared resources except GPU that is assigned exclusively to one user.

The measured consumed energy is multiplied by the coefficient, which is interpreted as the amount of emitted carbon dioxide mass necessary to produce this energy. The coefficient is determined automatically based on the computational resource location. For Czechia, the coefficient has the value of 386 grams of carbon dioxide per kilowatt-hour of consumed energy (gCO_2/kWh)[9].

Another implemented feature is model checkpointing, which monitors the value of loss function on training data after each epoch. At the end of the training, it returns the model with a setting that achieved the lowest loss value. This model is considered the best from this training and is stored for evaluation. Since there is performed 10 training runs for one dataset, the ten best models (one from each run) are saved for evaluation.

The implementation of training loop and model checkpointing is realized via the library PyTorch Lightning[10] that offers convenient interface and implementation guidelines. This library also takes care of the proper utilization of available computational resources (e.g. GPUs) with the possibility of configuration. To track, log and store the training data is selected library wandb, which provides an online tool for visualization of measured data with the possibility of its real-time monitoring[11].

4.4 Evaluation implementaion

After the training, each model makes predictions on its test part of the dataset, which results are stored and used as input for functions that compute metrics. The list of considered metrics includes: **accuracy** (weighted and unweighted), **precision**, **recall**, **f1 score** and **specificity**. Except for accuracy, the rest of the metrics are computed in three ways: per class, unweighted and weighted per dataset. For accuracy, there is provided only weighted and unweighted per dataset values.

Before metrics computation, the first step is to build the **confusion matrix**, which example is depicted in Figure 4.2. This matrix is starting point for further calculations. The formulas used to gain metrics values are shown in equations 4.1– 4.5. In each of these formulas TP (*true positive*) means the number of correctly classified samples into the *positive* category, TN (*true negative*) represents the number of correctly classified samples into the *negative* category, FP (*false positive*) means the number of incorrectly classified samples into the *positive* category and FN (*false negative*) denotes the number of incorrectly classified samples into the *negative* category.

To enumerate given formulas, there is need to substitute TP, TN, FP and FN with specific values. They are computed with an approach called *One-vs-Rest* (OvR) that selects one class as *positive* (referential) and others are considered as *negative*. Then, TP is equal to the value of correctly classified samples for the referential class, TN is equal to the number of samples that are marked as negative and their predicted class is negative too. FP is the sum of samples that are classified as positive, but their true label is negative. FN is

[9] https://github.com/mlco2/codecarbon/blob/master/codecarbon/data/private_infra/eu-carbon-intensity-electricity.csv

[10] https://lightning.ai/docs/pytorch/stable/

[11] https://wandb.ai/

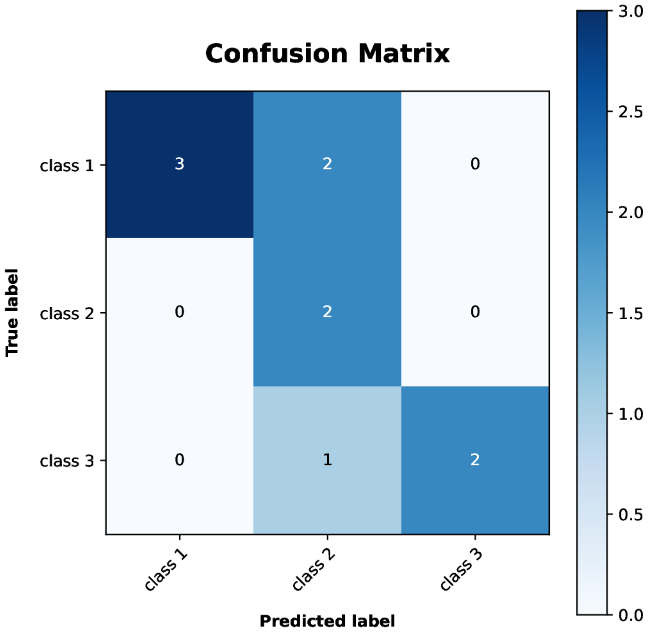


Figure 4.2: Example of the confusion matrix. The horizontal axis represents the predicted classes and the vertical target classes (ground truths). The values in diagonal cells represent the number of correctly predicted data samples for a given class.

the sum of cells that represents negative predictions for samples with a positive true label. This process is repeated for each class. The per-dataset values for precision, recall, F1 score and specificity are obtained as weighted and unweighted averages of per-class metrics. The implementation utilizes the Scikit-learn and Torchmetrics[12] libraries for metrics calculation.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1\,score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4.5}$$

This chapter contained implementation details of the program that performs speech recording preprocessing, model initialization and training. At the beginning of the chapter was described the dataset class that provides data samples for the AST model in demanded form. It includes the entire process from reading the speech recording data from the file to

[12] https://torchmetrics.readthedocs.io/en/stable/

editing them using augmentation techniques. The following section designates the model implementation details. At first, it was described the partitioning of mel spectrograms into the sequence of patches enriched by values determining their position in the original input. Then the inference process summary was provided, which outputs an vector with probabilities for each of the considered emotional classes. The last two sections contained a description of the training methodology and the evaluation of training outputs.

Chapter 5

Results

The model implementation described in chapter 4 was tested on a set of experiments with the principal intention of determining its performance on a speech emotion recognition task.

Each experiment comprehended model training on the particular emotional dataset with the following evaluation, which concept was proposed in Subsection 3.4.4. The proper description of evaluation methodology with particular goals is part of this chapter. All experiments were run with the following setting: *number of training epochs* = 200, *initial learning rate value* = 0.0001, *mini-batch size* = 32, *mel spectrogram columns count (time frames)* = 512, *mel spectrogram rows count (mel filter bank size)* = 64, *random seed* = 0 (to control random operation outputs), *number of folds* = 10. The training was performed on computer with one GPU of type Nvidia A40 and operation system Linux.

5.1 Results evaluation methodology

The evaluation was designed with the aim of testing the AST's ability to learn emotional patterns that are present in the input speech recordings. The methodology was inspired by reviewed state-of-the-art works [25, 49, 50, 45] in this field that is described in Section 2.7 with additional evaluation components that the thesis author lacked in these works.

The main idea is to evaluate trained models from a 10-fold cross-validation method and according to observed metrics (accuracy, precision, recall, F1 score and specificity) decide, if the model has enough generalization ability and is suitable for this task. This approach should ensure that the results are independent of dataset partitioning to train and test part.

The evaluation is performed for each dataset exactly once. From each fold, there is selected the best model that obtained the lowest value of loss function during the training. Each model then makes predictions on test data samples and their results are stored in evidence. After the last model finishes its predictions, this evidence is used for confusion matrix construction and metrics calculation. For each dataset, the measured values are provided in a tabular form together with metrics provided by other state-of-the-art solutions (if they used a particular dataset obviously). Moreover, there is provided confusion matrix and the progress curves of the loss function and training accuracy during epochs.

The originally-intended evaluation of carbon footprint with Codecarbon library and its environment, unfortunately, fails due to its buggy behaviour and incomplete documentation that lacks necessary details. In order to provide some kind of interpretation of measured carbon footprint data, there is a dedicated section on this topic. It comprehends the bar chart for each dataset with the amount of emitted CO_2 mass per each fold. To provide an interpretation that is more „human-friendly“, there is a figure, which depicts the frequency or quantity of popular activities that have the same environmental impact as the performed computation during this thesis. The chosen activities *driving the car*, *smartphone charging* and *coal burning*. For this purpose is selected Greenhouse Gas Equivalencies Calculator[1]. It is worth noting that this portal is primarily addressed to the USA region and therefore obtained values are just rough estimations but principal goal is to show the environmental costs of deep learning.

The last sections are dedicated to the evaluation of model complexity in terms of time and memory, discussion of obtained results and potential future work.

5.2 RAVDESS dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song [30] is one of the most used datasets for developing and testing models for speech emotion recognition tasks according to the number of citations, as shown in Table 2.2. This freely available dataset contains 1440 English speech recordings labelled by eight discrete emotion classes. The dataset is categorized according to taxonomy, introduced in Section 2.2, into the acted category because each recording was performed by a professional actor.

Figure 5.1 shows the confusion matrix. It implies that trained models recognize emotion *surprised* with the highest accuracy, while with *sad* and *neutral* emotions struggle. These facts are confirmed by values in Table 5.1, which contains the per-class metrics. It suggests that *neutral* emotion has more false positive predictions than false negative because precision is smaller than recall. The opposite case is *happy* and *emotions*, where the precision is higher than recall suggesting more false negative than false positive predictions. The metrics for other emotions indicate almost equal values for false positive and false negative predictions.

Emotion	*Precision*	*Recall*	*F1 score*	*Specificity*
neutral	0.709	0.813	0.757	0.976
calm	0.862	0.880	0.871	0.978
happy	0.836	0.770	0.802	0.976
sad	0.796	0.750	0.772	0.970
angry	0.880	0.885	0.883	0.981
fearful	0.853	0.844	0.848	0.978
disgust	0.873	0.896	0.884	0.984
surprised	0.888	0.906	0.897	0.982

Table 5.1: Class metrics for RAVDESS dataset

Table 5.2 contains values of global metrics compared with other SER models that used RAVDESS. This table shows that the AST model reached the second-best accuracy among compared models. Unfortunately, the other metrics are not provided.

Figure 5.2 shows statistics observed during the fitting of the models. Figure 5.2a shows a chart with the loss function progress, while Figure 5.2b depicts the accuracy progress for training data. Both figures provide curves for each fold but they are aligned so well due to performance consistency so it is hard to recognize them. This consistency indicates that for AST with initial conditions set in this work, the model is independent of dataset splitting.

[1] https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator#results

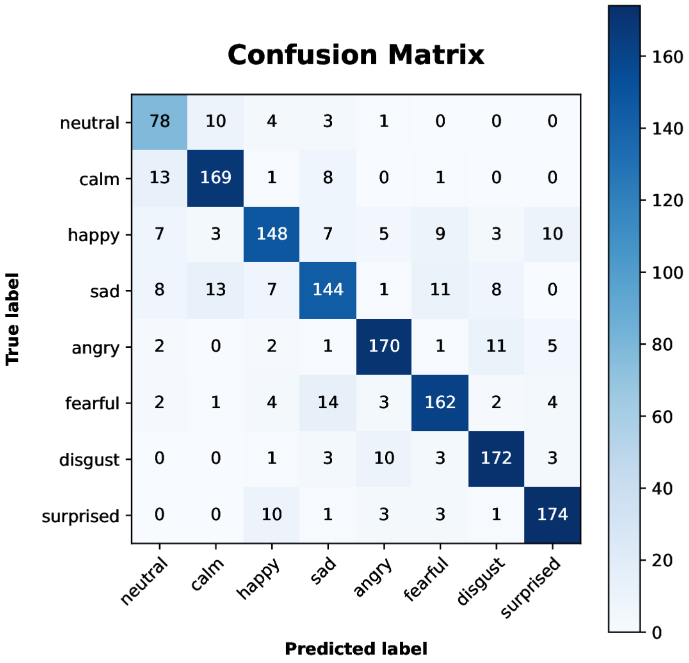
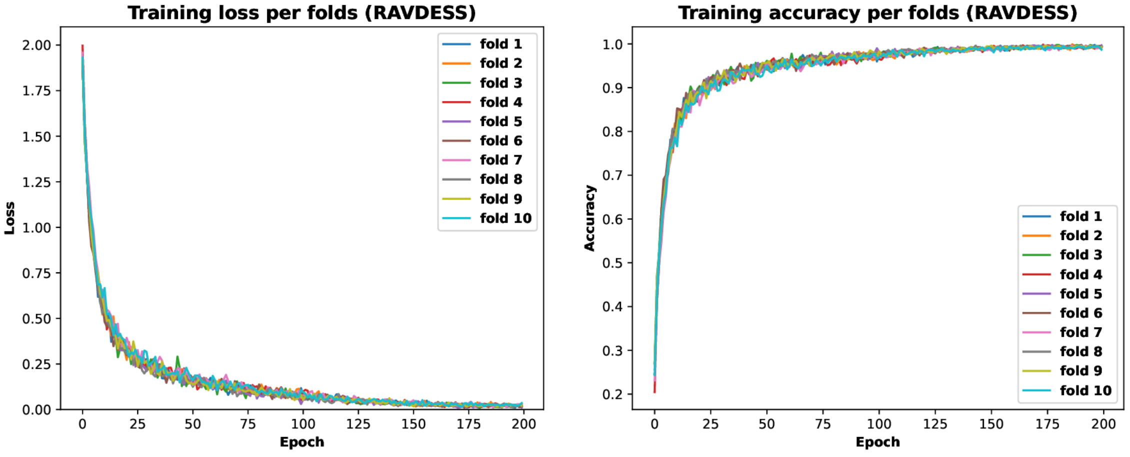


Figure 5.1: RAVDESS confusion matrix.

	Accuracy	*Precision*	*Recall*	*F1 score*	*Specificity*
C-RNN	0.7/-	-	-	-	-
TIM-Net	91.93 / 92.08	-	-	-	-
AST	0.845 / 0.851	0.837 / 0.846	0.843 / 0.845	0.839 / 0.845	0.978 / 0.978

Table 5.2: Global metrics for RAVDESS dataset compared with other solutions. The format for each cell is **unweighted value** / **weighted value**



(a) Loss function progress

(b) Training accuracy progress

Figure 5.2: RAVDESS loss and accuracy progress for 10-fold cross-validation. Both curves aligned so well that it indicates consistent behaviour no matter without the influence of dataset split.

5.3 Emo-DB dataset

Emo-DB [7] is the most popular dataset for speech emotion recognition according to the number of citations in academic papers, which demonstrates Table 2.2. It comprehends 535 utterances in the German language recorded by professional actors, so it is categorized as the acted dataset. It divides recordings into seven discrete emotional classes.

Figure 5.3 depicts the confusion matrix created from predictions of trained models on their test sets. It suggests that model performance for this dataset is good because of low false positives and false negatives occurrence frequency. The only odd one in false positive predictions is *happiness* emotion.

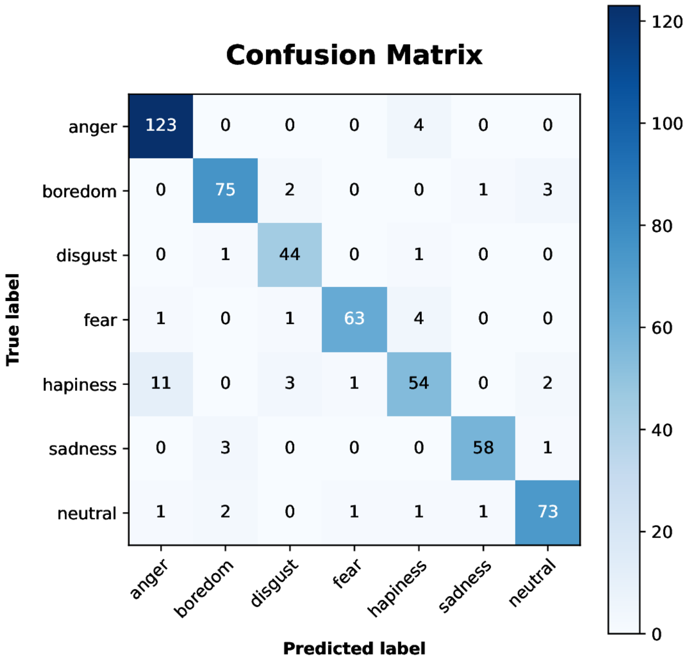


Figure 5.3: Emo-DB confusion matrix.

According to class metrics in Table 5.3, for five out of seven classes the model reached more than 90% for each observed metric. The biggest problem is noticeable with emotion *happiness*. Recall value under 80% confirms a high ratio of *false negative* predictions for this emotion, as mentioned earlier.

Table 5.4 shows global statics compared to other solutions mentioned in the thesis, but only TIM-Net utilized the Emo-DB as well.

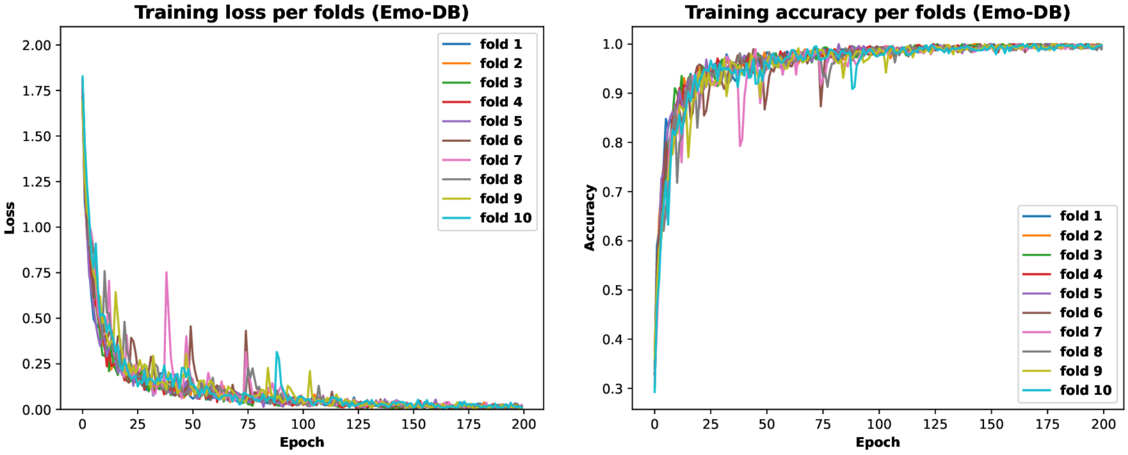
Figure 5.4 comprehends training statistics for loss function together with accuracy progress. The curves are again well-aligned, but the deviations are more frequent, mainly in the first hundred epochs compared to the RAVDESS data. These figures also imply the fast convergence of the loss function, which mostly decreases during the first thirty epochs. This fact is also noticeable in training accuracy, which reached 0.9 value in the first fifty epochs in each fold.

Emotion	*Precision*	*Recall*	*F1 score*	*Specificity*
anger	0.904	0.969	0.936	0.968
boredom	0.926	0.926	0.926	0.987
disgust	0.88	0.957	0.917	0.988
fear	0.969	0.913	0.940	0.996
happiness	0.844	0.761	0.800	0.978
sadness	0.966	0.935	0.950	0.996
neutral	0.924	0.924	0.924	0.987

Table 5.3: Class metrics for Emo-DB dataset.

	Accuracy	*Precision*	*Recall*	*F1 score*	*Specificity*
TIM-Net	0.952 / 0.957	-	-	-	-
AST	0.912 / 0.916	0.916 / 0.916	0.912 / 0.916	0.913 / 0.915	0.990 / 0.983

Table 5.4: Global metrics for Emo-DB dataset compared with other solutions. The format for each cell is **unweighted value** / **weighted value**.



(a) Loss function progress.

(b) Training accuracy progress.

Figure 5.4: Emo-DB loss and accuracy progress for 10-fold cross-validation. The curves contain occasional deviations that are stabilized after the first hundred epochs.

5.4 EMOVO dataset

EMOVO [12] is an Italian emotional corpus, which contains 588 recordings that belong to seven emotional classes. The utterances in EMOVO were recorded by professional actors. Therefore is EMOVO included in the family of acted datasets. The reason for the inclusion of this dataset is to expand the set of languages that were utilized during training and evaluation and observe their influence.

The confusion matrix depicted in Figure 5.5 indicates that despite EMOVO having almost the same number of recordings as Emo-DB, model performance is not the same or similar. The *joy* emotion seems the most problematic because of its high ratio of false positives as well as false negatives. The model predicted *sad* emotion most successfully.

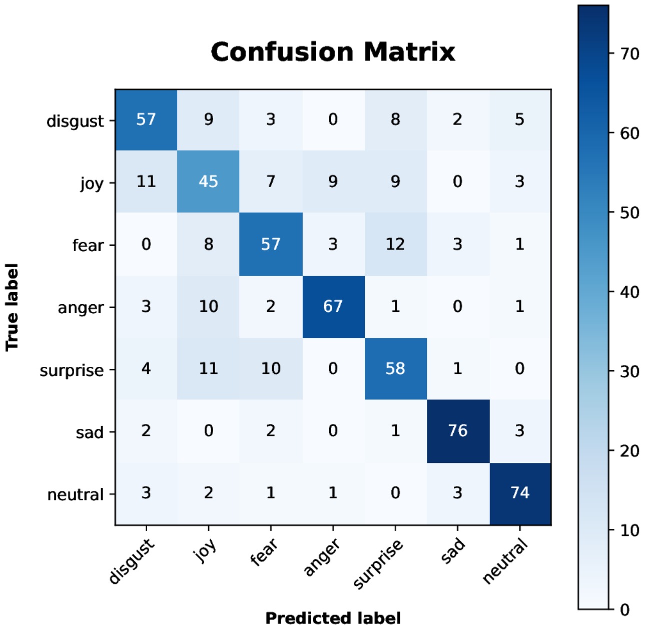


Figure 5.5: EMOVO confusion matrix.

Table 5.5 shows that model poorly recognizes *joy* and *surprise* emotions, but on the other hand it, *sad* and *natural* emotions classifies with relatively good results. *Surprise* and *neutral* emotions have a higher ratio of false positive than false negative predictions. The opposite cases are *disgust*, *anger* and *fear* with a smaller number of false positive than false negative predictions. Other emotions have almost the same ratio of observed metrics.

Emotion	*Precision*	*Recall*	*F1 score*	*Specificity*
disgust	0.713	0.679	0.695	0.954
joy	0.529	0.536	0.533	0.921
fear	0.695	0.679	0.687	0.950
anger	0.838	0.798	0.817	0.974
surprise	0.652	0.690	0.671	0.938
sad	0.894	0.905	0.899	0.982
neutral	0.850	0.881	0.865	0.974

Table 5.5: Class metrics for EMOVO dataset.

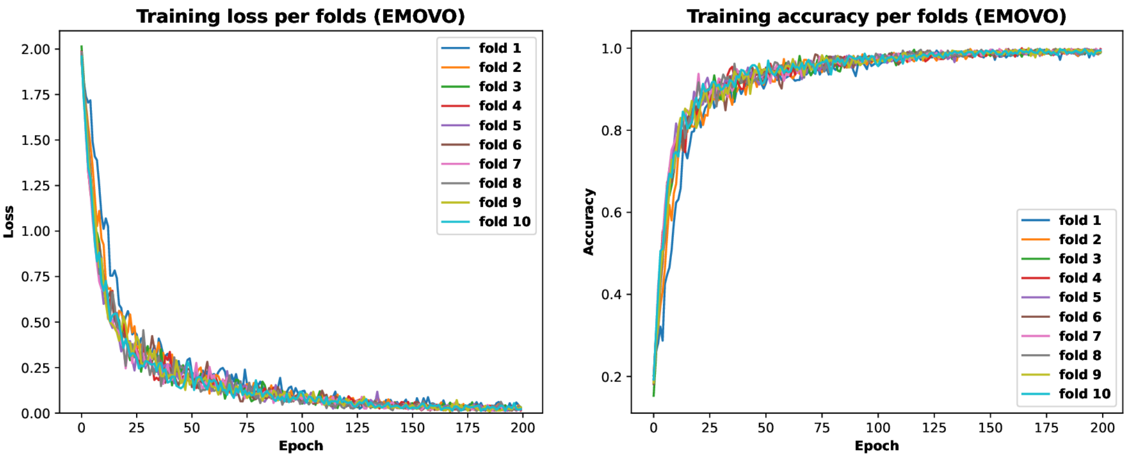
Table 5.6 contains a comparison of global metrics between AST and Tim-NET. Tim-NET with the same training and test methodology has better accuracy almost by 0.2. The rest of the models described in this thesis as state-of-the-art did not use the EMOVO as their training dataset.

Loss and accuracy progress depicted 5.6 shows an interesting paradox in terms of results during evaluation. Despite having the excellent convergence of loss function and training accuracy to 0 respectively to 1, the model performance on test data is not quite well. This fact implies the overfitting problem in spite of the data augmentation utilization.

	Accuracy	*Precision*	*Recall*	*F1 score*	*Specificity*
TIM-Net	0.92 / 0.92	-	-	-	-
AST	0.738 / 0.738	0.739 / 0.739	0.738 / 0.738	0.738 / 0.738	0.956 / 0.956

Table 5.6: Global metrics for EMOVO dataset compared with other solutions. The format for each cell is **unweighted value** / **weighted value**.

The progress curves show the stability of training for each fold without any notable deviations, similar to RAVDESS dataset. The fold curves are aligned, which means the convergence speed is almost the same for each fold, as in the case of previous datasets.



(a) Loss function progress.

(b) Training accuracy progress.

Figure 5.6: EMOVO loss and accuracy progress for 10-fold cross-validation. The curves indicate steady training without siginificant deviations.

5.5 Carbon footprint evaluation

This section is dedicated to the evaluation of the carbon footprint that was produced during neural network training. The information worth emphasizing is that provided CO_2 data are computed from energy consumed by GPU (detailed explanation of reasons in Section 4.3) during training measured in kilowatt-hours. This data were according to the used methodology multiplied by the carbon intensity index.

Table 5.7 shows statistics of emitted CO_2 with total training time for each dataset. According to the data in this table, the most-consuming training was with the RAVDESS dataset. This fact follows from the size of this dataset, which is bigger in order of magnitude than other used datasets.

The second dataset in terms of measured emissions as well as training time is EMOVO. Despite a comparable number of recordings to Emo-DB, during training with EMOVO was emitted vastly more CO_2 than training with Emo-DB. The argument for this is the length of recordings, which are longer in EMOVO datasets in the unit of seconds. This implies also the training time difference between these two datasets, which is almost three hours.

Dataset	*Emitted CO_2 (grams)*	*Training runtime length*
RAVDESS	521.3366 g	16h 17m
Emo-DB	198.2820 g	5h 20m
EMOVO	338.7516 g	8h 17m
Total	1058.3703 g	29h 54m

Table 5.7: Carbon footprint and training time statistics.

Figure 5.7 shows the amount of emitted CO_2 in grams per fold for each dataset. Training with Emo-DB (Figure 5.7b) and EMOVO (Figure 5.7c) shows relatively steady emission production without any significant variance across folds. The opposite case is in the training with RAVDESS (Figure 5.7a), where the variance between fold with the lowest amount of emitted CO_2 (fold 1) and fold with the highest (fold 6) amount is almost ten grams. The source of such variability is unknown.

Figure 5.8 depicts activities that have a similar environmental impact as all training sessions performed during the thesis writing. This figure implies that for such training it is needed to produce energy that has the same carbon footprint as driving an average gasoline car for circa 4.34 kilometres, charging a smartphone 129 times and burning roughly 0.5 kilograms of coal with respect to the location of Czechia (dependent on used carbon intensity index).

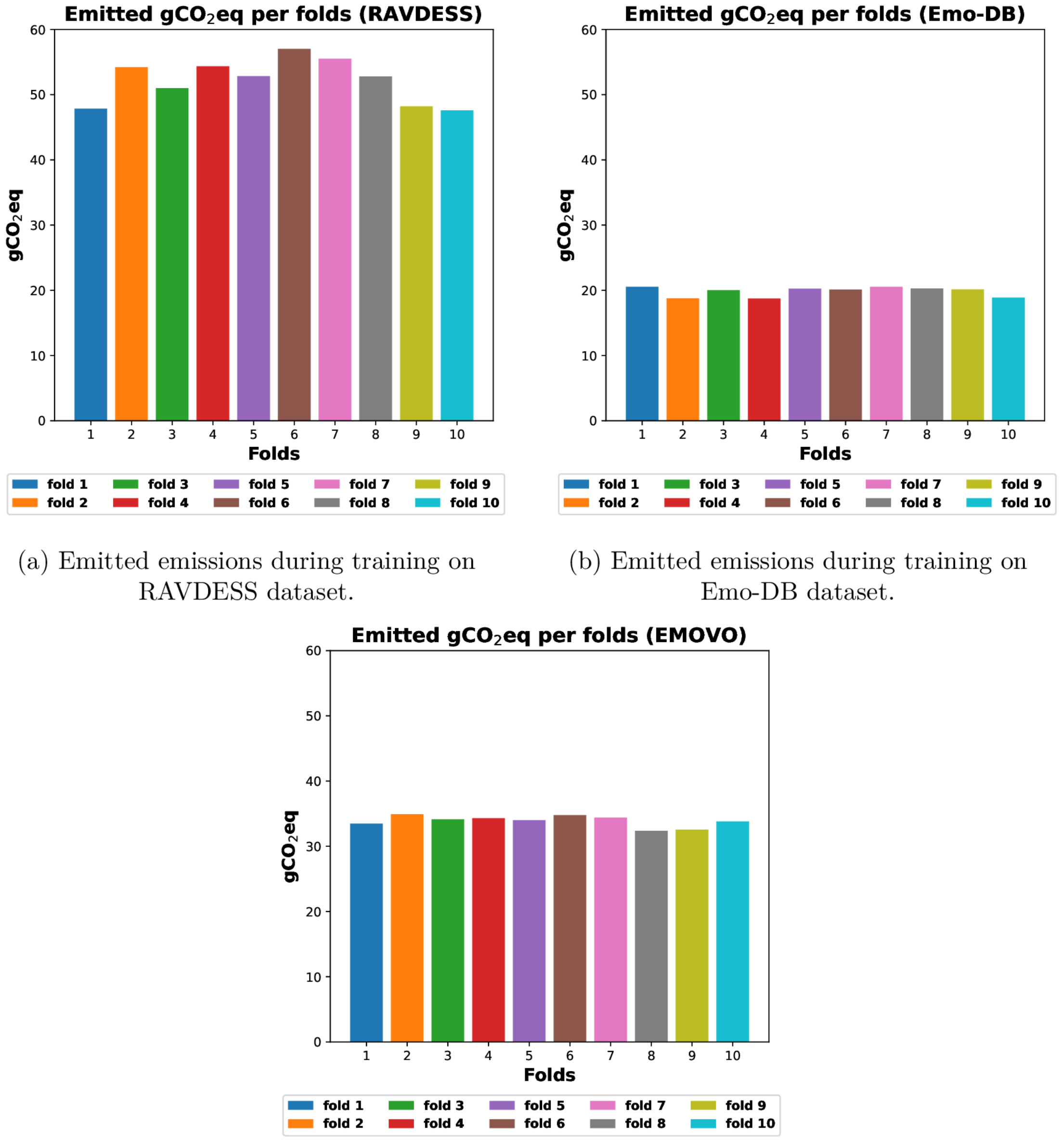
5.6 Model analysis

This section is dedicated to the model analysis in terms of its space requirements, steps required to perform one inference and average time to train one fold in 10-fold cross-validation in the used environment.

The implemented model has $\approx$ 5.8 million trainable parameters and takes 23.1 MiB of memory. To complete one inference, the implemented model needs to perform the following steps:

1. divides input mel spectrogram into patches - implemented convolutional layer,
2. transforms patches to embeddings and add positional encoding - implemented with matrix multiplication,
3. passes through encoder blocks - implemented as a series of self-attention blocks and fully connected layers,
4. computes of output in the classification head - implemented as one fully connected with layer normalization.

The required time for model training differs according to the selected computational resources and values of hyperparameters. In the case of this work (exact hyperparameters setting is at the beginning of Chapter 5), the average training time to complete one fold in 10-fold cross-validation for RAVDESS was 2 hours and 5 minutes, for Emo-DB 31 minutes and for EMOVO 52 minutes.

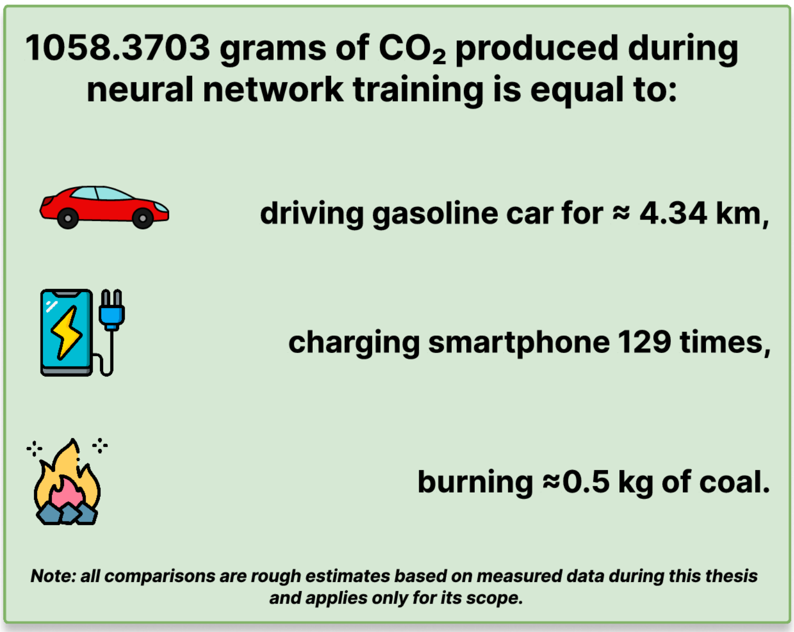


(a) Emitted emissions during training on RAVDESS dataset.

(b) Emitted emissions during training on Emo-DB dataset.

(c) Emitted emissions during training on EMOVO dataset.

Figure 5.7: Emitted CO_2 per datasets. Each subfigure depicts the amount of emitted CO_2 in grams per folds.



1058.3703 grams of CO_2 produced during neural network training is equal to:

driving gasoline car for ≈ 4.34 km,

charging smartphone 129 times,

burning ≈0.5 kg of coal.

Note: all comparisons are rough estimates based on measured data during this thesis and applies only for its scope.

Figure 5.8: Activities with a comparable amount of emitted CO_2 as the produced amount throughout the whole training.

5.7 Discussion and future work

This section is dedicated to the discussion of obtained results in terms of stated challenges that implemented system should address. The obtained results have several implications for further work.

The first challenge this model addressed was if the model based on Transformer architecture without convolutions or recurrent units is able to classify emotions based on the speech input. The results give answers that it is possible with relative success. Although the model's performance did not reach the highest state-of-the-art score, it is possible to confidently claim, based on the measured metrics, that the model with this architecture is capable of learning emotional patterns in speech recordings and classifying them in a manner that cannot be considered coincidental.

The second addressed challenge was to answer the question if it is possible to transfer knowledge from other domains, in this case, image classification, and use it for emotion classification. The evaluation outcomes indicate that the Transfomer-based model (ViT), pretrained on the image dataset, is able to use obtained „knowledge" in speech emotion recognition when the input recordings are adjusted to the appropriate form.

The last thesis solution challenge was to measure the environmental impact of model training. As pointed out in the summary of challenges, this topic is neglected, despite its importance for the whole deep learning paradigm. This thesis showed that some methods and tools for measuring the environmental impact of deep learning are already present but not sufficiently tuned to be satisfactorily used even for academic-level work. Apart from this fact, throughout the training were tracked carbon footprint as the rough estimates of emitted CO_2. Their interpretation provides insights into how the carbon footprint was developed during the training and which activities have similar environmental impacts.

For future work, the emphasis should be put on three principal goals. The first is to train the AST with different initial settings (mel filter banks size, number of mel spectrogram frames, patch size etc.) to find out if it impacts results.

The second objective is to train the AST on a more comprehensive and robust dataset in terms of data sample count and speech recordings origin, for example MSP-Podcast [31], which contains more than hundred thousand of natural speech recordings. Utilization of this dataset should bring more insight into the capabilities of AST in speech emotion recognition.

The last goal of future work should be to perform cross-corpus evaluation. The idea of this approach is to train the model on one dataset and evaluate it on completely different (e.g. to train on the RAVDESS dataset and evaluate on the EMOVO dataset). It should address the generalization question that is crucial for speech emotion recognition. There is already an early methodology for this topic in the paper by Cheng et al. [47].

Overall, the implemented Audio Spectrogram Transformer model showed promising results that make it a suitable candidate for future improvement in the speech emotion recognition task. One of the significant benefits of this work is the comprehensive evaluation methodology, which was not present in any of the reviewed works. The last highlight of this thesis is the measurement of the environmental impact of the selected deep learning approach, which was not implemented in any of the reviewed works and therefore bring a novelty with future development potential.

Chapter 6

Conclusion

The aim of this thesis was to design and implement a system that is able to recognize the emotions from the input speech recordings. Although this topic is more than 20 years old, there is still no solution that would reach human-level performance, and none of them is even approaching. Based on this fact, a comprehensive analysis of methods, techniques and contemporary state-of-the-art solutions was required in order to get a good grasp of this topic. The analysis results showed that the current trend in the solutions performing at the state-of-the-art level is the utilization of deep learning models, particularly convolutional and recurrent neural networks.

Because there were a lot of solutions founded on mentioned neural network types and none that utilized the Transformer neural network, which is cutting-edge technology in natural language processing and image classification tasks, the author decided to propose a solution founded on Transformer. Throughout the research was found Audio Spectrogram Transformer (AST) neural network. It is a pure Transformer model (without any convolutional layers or recurrent units), originally proposed for the classification of audio sounds, which accepts mel spectrogram extracted from input recording. It is a derivative of Vision Transformer, pretrained on the ImageNet dataset, with a slightly adjusted linear transformation layer and positional encoding.

The proposed solution is a pipeline that loads speech recording and transforms it to the mel spectrogram with the required shape defined by the size of the mel filter bank and the number of time frames. This spectrogram is processed by AST neural network outputting the probabilities for considered emotion classes.

The implemented neural network was trained on three datasets: RAVDESS, Emo-DB and EMOVO. For each dataset was performed 10-fold cross-validation. The evaluation was conducted for each dataset separately. The goal was to find out how well the proposed neural network is able to classify emotions and check if dataset split influences model performance. Therefore, the methodology consisted of standard classification metrics computation: accuracy, precision, recall, F1 score and specificity. The input for metrics formulas was test predictions evidence that contained predictions for a particular dataset. Despite not achieving state-of-the-art performance, the obtained unweighted accuracy values are promising: 84.5 % for RAVDESS, 91.6 % for Emo-DB and 73.8 % for EMOVO.

Throughout the training were tracked their environmental impact as a rough estimation of CO_2 mass based on the consumed energy of the used GPU. According to the measured values, there was emitted 1058.3703 grams of CO_2 in total. The tool used for measurement, however, failed in the evaluation and therefore it was conducted in a custom way.

The directions for future work are in testing the AST model with different settings that would allow studying their impact on performance and the utilization of the bigger emotional corpus with natural-based speech recordings that would show model capabilities on real-world data.

Bibliography

[1] *Berlin Database of Emotion speech citation count* [https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=A+database+of+German+emotional+speech&sid=402df430e1babea8efc8fe60a5851fc3&sot=b&sdt=b&sl=44&s=TITLE%28A+database+of+German+emotional+speech%29&origin=searchbasic&editSaveSearch=&yearFrom=Before+1960&yearTo=Present]. Accessed: 2022-11-26.

[2] *IEMOCAP citations count* [https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=IEMOCAP%3a+interactive+emotional+dyadic+motion+capture+database.&sid=30d7b4add7fcdfcccd175ed77604be39&sot=b&sdt=b&sl=69&s=TITLE%28IEMOCAP%3a+interactive+emotional+dyadic+motion+capture+database.%29&origin=searchbasic&editSaveSearch=&yearFrom=Before+1960&yearTo=Present]. Accessed: 2022-11-26.

[3] *RAVDESS citations count* [https://www.scopus.com/results/results.uri?sort=plf-f&src=s&st1=10.1371%2fjournal.pone.0196391&sid=a1f7ca1af0074a72f5f97e8f0f553fd7&sot=b&sdt=b&sl=33&s=DOI%2810.1371%2fjournal.pone.0196391%29&origin=searchbasic&editSaveSearch=&yearFrom=Before+1960&yearTo=Present]. Accessed: 2022-11-26.

[4] Akçay, M. B. and Oğuz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*. 2020, vol. 116, p. 56–76. DOI: https://doi.org/10.1016/j.specom.2019.12.001. ISSN 0167-6393. Available at: https://www.sciencedirect.com/science/article/pii/S0167639319302262.

[5] Anthony, L. F. W., Kanding, B. and Selvan, R. *Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models*. arXiv, 2020. DOI: 10.48550/ARXIV.2007.03051. Available at: https://arxiv.org/abs/2007.03051.

[6] Bakker, I., Voordt, T. van der, Vink, P. and Boon, J. de. Pleasure, Arousal, Dominance: Mehrabian and Russell revisited. *Current Psychology*. Springer Science and Business Media LLC. june 2014, vol. 33, no. 3, p. 405–421. DOI: 10.1007/s12144-014-9219-4. Available at: https://doi.org/10.1007/s12144-014-9219-4.

[7] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F. and Weiss, B. A database of German emotional speech. In: *Proc. Interspeech 2005*. 2005, p. 1517–1520. DOI: 10.21437/Interspeech.2005-446.

[8] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E. et al. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*. Springer Science and Business Media LLC. november 2008, vol. 42, no. 4, p. 335–359. DOI: 10.1007/s10579-008-9076-6. Available at: https://doi.org/10.1007/s10579-008-9076-6.

[9] Bäckström, T., Räsänen, O., Zewoudie, A., Zarazaga, P. P., Koivusalo, L. et al. *Introduction to Speech Processing*. 2nd ed. 2022. Available at: https://speechprocessingbook.aalto.fi.

[10] Cambridge Dictionary. Emotion definition. In:. October 2022. Available at: https://dictionary.cambridge.org/dictionary/english/emotion.

[11] Cannon, W. B. The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory. *The American Journal of Psychology*. University of Illinois Press. 1927, vol. 39, 1/4, p. 106–124. ISSN 00029556. Available at: http://www.jstor.org/stable/1415404.

[12] Costantini, G., Iaderola, I., Paoloni, A., Todisco, M. et al. EMOVO corpus: an Italian emotional speech database. In: European Language Resources Association (ELRA). *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. 2014, p. 3501–3504.

[13] Darwin, C., Ekman, P. and Prodger, P. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 1998. ISBN 9780195158069. Available at: https://books.google.cz/books?id=TFRtLZSHMcYC.

[14] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019.

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv, 2020. DOI: 10.48550/ARXIV.2010.11929. Available at: https://arxiv.org/abs/2010.11929.

[16] Ekman, P. and Oster, H. Facial Expressions of Emotion. *Annual Review of Psychology*. 1979, vol. 30, no. 1, p. 527–554. DOI: 10.1146/annurev.ps.30.020179.002523. Available at: https://doi.org/10.1146/annurev.ps.30.020179.002523.

[17] Ekman, P. An argument for basic emotions. *Cognition and Emotion*. Routledge. 1992, vol. 6, 3-4, p. 169–200. DOI: 10.1080/02699939208411068. Available at: https://doi.org/10.1080/02699939208411068.

[18] El Ayadi, M., Kamel, M. S. and Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*. 2011, vol. 44, no. 3, p. 572–587. DOI: https://doi.org/10.1016/j.patcog.2010.09.020. ISSN 0031-3203. Available at: https://www.sciencedirect.com/science/article/pii/S0031320310004619.

[19] Gong, Y., Chung, Y.-A. and Glass, J. *AST: Audio Spectrogram Transformer*. arXiv, 2021. DOI: 10.48550/ARXIV.2104.01778. Available at: https://arxiv.org/abs/2104.01778.

[20] Hennenlotter, A., Dresel, C., Castrop, F., Ceballos Baumann, A. O., Wohlschläger, A. M. et al. The Link between Facial Feedback and Neural Activity within Central Circuitries of Emotion—New Insights from Botulinum Toxin–Induced Denervation of Frown Muscles. *Cerebral Cortex*. june 2008, vol. 19, no. 3, p. 537–542. DOI: 10.1093/cercor/bhn104. ISSN 1047-3211. Available at: https://doi.org/10.1093/cercor/bhn104.

[21] Jahangir, R., Teh, Y. W., Hanif, F. and Mujtaba, G. Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications*. july 2021, vol. 80, no. 16, p. 23745–23812.

[22] James, W. What is an Emotion? *Mind*. [Oxford University Press, Mind Association]. 1884, vol. 9, no. 34, p. 188–205. ISSN 00264423, 14602113. Available at: http://www.jstor.org/stable/2246769.

[23] Jeon, M. Chapter 1 - Emotions and Affect in Human Factors and Human–Computer Interaction: Taxonomy, Theories, Approaches, and Methods. In: Jeon, M., ed. *Emotions and Affect in Human Factors and Human-Computer Interaction*. San Diego: Academic Press, 2017, p. 3–26. DOI: https://doi.org/10.1016/B978-0-12-801851-4.00001-X. ISBN 978-0-12-801851-4. Available at: https://www.sciencedirect.com/science/article/pii/B978012801851400001X.

[24] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H. et al. Speech Emotion Recognition Using Deep Learning Techniques: A Review. *IEEE Access*. 2019, vol. 7, p. 117327–117345. DOI: 10.1109/ACCESS.2019.2936124.

[25] Kumaran, U., Radha Rammohan, S., Nagarajan, S. M. and Prathik, A. Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN. *International Journal of Speech Technology*. Jun 2021, vol. 24, no. 2, p. 303–314. DOI: 10.1007/s10772-020-09792-x. ISSN 1572-8110. Available at: https://doi.org/10.1007/s10772-020-09792-x.

[26] Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T. *Quantifying the Carbon Emissions of Machine Learning*. arXiv, 2019. DOI: 10.48550/ARXIV.1910.09700. Available at: https://arxiv.org/abs/1910.09700.

[27] Latif, S., Rana, R., Khalifa, S., Jurdak, R. and Epps, J. *Direct Modelling of Speech Emotion from Raw Speech*. arXiv, 2019. DOI: 10.48550/ARXIV.1904.03833. Available at: https://arxiv.org/abs/1904.03833.

[28] Lieskovská, E., Jakubec, M., Jarina, R. and Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics*. MDPI AG. May 2021, vol. 10, no. 10, p. 1163. DOI: 10.3390/electronics10101163. ISSN 2079-9292. Available at: http://dx.doi.org/10.3390/electronics10101163.

[29] Liu, G. K. Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech. *CoRR*. 2018, abs/1806.09010. Available at: http://arxiv.org/abs/1806.09010.

[30] Livingstone, S. R. and Russo, F. A. 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*. Public Library of Science. may 2018, vol. 13, no. 5, p. 1–35. DOI: 10.1371/journal.pone.0196391. Available at: https://doi.org/10.1371/journal.pone.0196391.

[31] Lotfian, R. and Busso, C. Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings. *IEEE Transactions on Affective Computing*. October-December 2019, vol. 10, no. 4, p. 471–483. DOI: 10.1109/TAFFC.2017.2736999.

[32] Mehrabian, A. Basic dimensions for a general psychological theory : implications for personality, social, environmental, and developmental studies. In:. 1980.

[33] Mustaqeem and Kwon, S. MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach. *Expert Systems with Applications*. 2021, vol. 167, p. 114177. DOI: https://doi.org/10.1016/j.eswa.2020.114177. ISSN 0957-4174. Available at: https://www.sciencedirect.com/science/article/pii/S0957417420309131.

[34] Nwe, T. L., Foo, S. W. and De Silva, L. C. Speech emotion recognition using hidden Markov models. *Speech Communication*. 2003, vol. 41, no. 4, p. 603–623. DOI: https://doi.org/10.1016/S0167-6393(03)00099-2. ISSN 0167-6393. Available at: https://www.sciencedirect.com/science/article/pii/S0167639303000992.

[35] Nwe, T., Foo, S. and De Silva, L. Detection of stress and emotion in speech using traditional and FFT based log energy features. In: *Fourth International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*. 2003, vol. 3, p. 1619–1623 vol.3. DOI: 10.1109/ICICS.2003.1292741.

[36] Plutchik, R. *Emotions in the practice of psychotherapy: Clinical implications of affect theories.* American Psychological Association, 2000. Available at: https://doi.org/10.1037/10366-000.

[37] Russell, J. A Circumplex Model of Affect. *Journal of Personality and Social Psychology*. december 1980, vol. 39, p. 1161–1178. DOI: 10.1037/h0077714.

[38] Schachter, S. and Singer, J. Cognitive, social, and physiological determinants of emotional state. *Psychological Review*. American Psychological Association (APA). september 1962, vol. 69, no. 5, p. 379–399. DOI: 10.1037/h0046234. Available at: https://doi.org/10.1037/h0046234.

[39] Shaver, P., Schwartz, J., Kirson, D. and O'Connor, C. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*. American Psychological Association (APA). 1987, vol. 52, no. 6, p. 1061–1086. DOI: 10.1037/0022-3514.52.6.1061. Available at: https://doi.org/10.1037/0022-3514.52.6.1061.

[40] Strubell, E., Ganesh, A. and McCallum, A. *Energy and Policy Considerations for Deep Learning in NLP*. arXiv, 2019. DOI: 10.48550/ARXIV.1906.02243. Available at: https://arxiv.org/abs/1906.02243.

[41] Swain, M., Routray, A. and Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*. Mar 2018, vol. 21, no. 1, p. 93–120. DOI: 10.1007/s10772-018-9491-z. ISSN 1572-8110. Available at: https://doi.org/10.1007/s10772-018-9491-z.

[42] Tao, J., Tan, T. and Picard, R. W., ed. *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 2005. Available at: https://doi.org/10.1007/11573548.

[43] Teager, H. M. and Teager, S. M. Evidence for nonlinear sound production mechanisms in the vocal tract. In: *Speech Production and Speech Modelling*. Dordrecht: Springer Netherlands, 1990, p. 241–261.

[44] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. et al. Attention is All you Need. In: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R. et al., ed. *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017, vol. 30. Available at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[45] Wang, X., Wang, M., Qi, W., Su, W., Wang, X. et al. A Novel end-to-end Speech Emotion Recognition Network with Stacked Transformer Layers. In: *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021, p. 6289–6293. DOI: 10.1109/ICASSP39728.2021.9414314.

[46] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M. and Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access*. 2021, vol. 9, p. 47795–47814. DOI: 10.1109/ACCESS.2021.3068045.

[47] Wen, X.-C., Ye, J.-X., Luo, Y., Xu, Y., Wang, X.-Z. et al. *CTL-MTNet: A Novel CapsNet and Transfer Learning-Based Mixed Task Net for the Single-Corpus and Cross-Corpus Speech Emotion Recognition*. 2022.

[48] Wong, E. and Sridharan, S. Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In: *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No.01EX489)*. 2001, p. 95–98. DOI: 10.1109/ISIMP.2001.925340.

[49] Ye, J., Wen, X., Wei, Y., Xu, Y., Liu, K. et al. *Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition*. arXiv, 2022. DOI: 10.48550/ARXIV.2211.08233. Available at: https://arxiv.org/abs/2211.08233.

[50] Zhao, Z., Li, Q., Zhang, Z., Cummins, N., Wang, H. et al. Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition. *Neural Networks*. 2021, vol. 141, p. 52–60. DOI: https://doi.org/10.1016/j.neunet.2021.03.013. ISSN 0893-6080. Available at: https://www.sciencedirect.com/science/article/pii/S0893608021000939.

[51] Zhou, G., Hansen, J. and Kaiser, J. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*. 2001, vol. 9, no. 3, p. 201–216. DOI: 10.1109/89.905995.

Appendix A

CD content

The attached CD contains following items:

- `thesis_source/` - Latex source code of the thesis,
- `implementation/` - implementation source files,
- `thesis.pdf` - final thesis pdf file.