Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

Jonas Klepper Rødningen

# Emotion Recognition from Speech and Instrumental Music: Proof of Shared Emotional Code Through Transfer Learning

Masteroppgave i Informatikk: Kunstig Intelligens
Veileder: Björn Gambäck
Juni 2022

NTNU
Kunnskap for en bedre verden

Jonas Klepper Rødningen

# Emotion Recognition from Speech and Instrumental Music: Proof of Shared Emotional Code Through Transfer Learning

Masteroppgave i Informatikk: Kunstig Intelligens
Veileder: Björn Gambäck
Juni 2022

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

NTNU

Kunnskap for en bedre verden

# Abstract

Emotion is a large part of the human experience—within ourselves, but also recognizable from e.g. affective sound like speech and music. Understanding how emotion is transmitted through sound is therefore highly relevant. Previous research has hypothesized and found supporting evidence of a shared emotional coding between these two forms of affective sound. The main goal of this thesis is to investigate the emotional coding overlap between natural (non-acted) speech and instrumental music. Instrumental music contains the least speech, and is therefore a stronger case to derive evidence from (than music with vocals).

A structured literature review of speech emotion recognition (SER) and (MER) is included, plus some additional transfer learning research between the domains. Two emotional *taxonomies* are compared in terms of suitability and potential. Moreover, a novel instrumental music dataset is compiled (available on Github), with static *valence* and *arousal* ratings. An optimized and graph-compatible Keras-layer implementation for a *dilated LSTM* was also made.

Experiments are done in large scale through direct transfer learning from SER (training) to MER (testing), which has never before been attempted. The second experimental setting is MER to MER, for comparison. Two customized neural network architectures are explored: *DCNN* (dilated CNN) and *ADCRNN* (attention dilated CNN RNN).

The DCNN managed 33.2% accuracy in the SER to MER setting, and 43.1% in the MER to MER setting. ADCRNN scored 30.7% and 49.2%, for the SER to MER and MER to MER settings (respectively). The experimental results are proof that at least some part of the domains' emotional coding are common—which is also analogous to previous neurological findings in the human brain. This proof is reflected by significantly stronger SER to MER performance than the random baseline (24% accuracy). More specifically, overlap has been proved for the emotional dimensions of arousal (stronger) and for valence (less). As the *true* overlap, in reality, is inconclusive based on the present results, future work is proposed for further exploration.

# Sammendrag (in Norwegian)

Følelser er en stor del av menneskelig tilværelse—både i oss selv, men også gjenkjennelig f.eks. i andre mennesker via stemme og i musikk. Å forstå hvordan følelser overføres via lyd er derfor veldig relevant. Tidligere forskning har hypotetisert og funnet bevis for delt emosjonell koding mellom disse to typene for affektiv lyd. Hovedmålet for denne avhandlingen er å undersøke overlappet for den emosjonelle kodingen mellom naturlig tale (ikke skuespill) og instrumentell musikk. Instrumentell musikk er den formen for musikk som inneholder *minst* tale, og er derfor den sterkeste settingen å utlede bevis fra (kontra musikk med vokaler).

En strukturert litteraturgjennomgang for SER ('speech emotion recognition') og MER ('music emotion recognition') er inkludert her, samt 'transfer learning'-forskning mellom domenene. To *følelses-taksonomier* blir sammenlignet, med fokus på egnethet og potensiale for denne konteksten. Videre kompileres et helt nytt datasett for instrumentell musikk (tilgjengelig på Github), med *statiske* følelsesrangeringer i form av de emosjonelle dimensjonene 'valence' og 'arousal'. I tillegg ble det laget en optimalisert og 'graph'-kompatibel 'Keras-layer'-implementasjon for 'dilated LSTM'.

Eksperimenter utføres i stor skala med transfer learning fra SER (trening) til MER (testing). Dette har aldri blitt gjort før. En ekstra eksperimentell setting gjøres for å kunne sammenligne; MER-til-MER. To skreddersydde nevrale nettverksarkitekturer (maskinlæring) utforskes: DCNN ('dilated CNN') og ADCRNN ('attention dilated CNN RNN').

DCNN klarte 33.2% 'accuracy' i *SER-til-MER*-settingen, og 43.1% for MER-til-MER. ADCRNN skåret 30.7% og 49.2%, for *SER-til-MER* og *MER-til-MER*, henholdsvis. Resultatene er bevis på at *minst* noen deler av domenenes emosjonelle koding er den samme—som også stemmer overens med resultater fra nevrologisk forskning i hjernen. Dette beviset er reflektert via signifikant bedre SER-til-MER ytelse enn en sammenligningsmodell som brukte ren tilfeldighet (24% accuracy). Mer spesifikt, så har overlapp blitt bevist for den emosjonelle dimensjonen arousal (sterkere) og valence (svakere overlapp). Siden den *sanne* mengden av overlapp, i virkeligheten, fremdeles er ukonkludert, så foreslås fremtidig arbeid for videre undersøkelse.

# Preface

This document is an artificial intelligence (AI) Master's thesis, in Informatics, at the Norwegian University of Science and Technology (NTNU). I give thanks for all the personal growth; personal healing; learning about sound and AI—that I have received throughout all the efforts connected to this research. I hope that all this and that the thesis will have beneficial future impact for many people. I also want to give thanks to my supervisor, Professor Björn Gambäck, for all his feedback and guidance during the past two semesters. I extend my appreciation to the researchers Lotfian and Busso (2019), Eerola and Vuoskoski (2011), Aljanaki et al. (2017) and Hung et al. (2021)—for the original datasets used in this thesis. To the reader: I hope you go for what you want in life, honor your needs, and remember that *you* are lovable.

Jonas Klepper Rødningen
Trondheim, June 1, 2022

# Contents

# List of Figures

# List of Tables

# Abbreviations

**ADCRNN** attention dilated CNN RNN. 62

**CNN** convolutional neural network. 61

**DCNN** dilated CNN. 62

**DL** deep learning. 49

**GAN** generative adversarial network. 19

**IC** inclusion criteria. 94, 95

**MER** music emotion recognition. 1–7, 12, 23–25, 31, 33, 36, 39, 40, 47–49, 93, 95

**MFCCs** Mel Frequency Cepstral Coefficients. 15, 36

**QC** quality criteria. 96

**RNN** recurrent neural network. 19, 20, 61

**SER** speech emotion recognition. 1–3, 6, 7, 12, 23–25, 31–33, 36, 37, 47, 50, 93, 95

**SLR** structured literature review. ix, 2, 7, 23, 25, 26, 31, 37, 47, 60, 93–96

**TFNN** tensor factorized neural network. 20, 37

# Chapter 1

# Introduction

Humans can perceive emotion both from other humans' speech and from music (among other mediums). As subfields of *affective computing*, the two research fields of speech emotion recognition (SER) and music emotion recognition (MER) seek to automatically recognize emotion from these affective sound modalities. Research in these domains has been going on for a while, but satisfying performance has not been achieved yet. In this chapter, the thesis is introduced; through its motivation, goal, research method and contributions.

## 1.1 Background and Motivation

Music is often considered the *language of emotion* (Weninger et al., 2013). The average music listening time per week has been reported to be 17.8 hours for adults (Nadon et al., 2021). As for speech, the vocal tract can for example be used to communicate *verbal language, with emotion*. Emotional information tend to be blended together with semantic verbal information in speech—and this non-verbal part is referred to as *prosody*. The personal motivation when starting this Master's thesis was a deep passion for both sound and emotion. At first, the intention was to explore music emotion recognition only—until it was discovered that *speech* emotion recognition was also a hot research topic. Later, a hypothesis was created of the different levels of 'speechiness' ranging all the way from natural speech to instrumental music (without vocals). This is presented in fig. 1.1. Further, a hypothesis was formulated that the underlying emotional code actually is the same thing through all of these forms of sound expression. Perhaps there exists an underlying universal language, or code, for emotion, and that this universal code is manifest in all forms of human emotional communication, e.g. speech, singing, instrumental music—and even non-vocal facial expression. One

way this hypothesis could be explored further was by testing transfer learning between SER and MER, or the reverse order. If transfer learning could be used to improve classification performance between the two domains with least similarity, which according to fig. 1.1 is regular speech and instrumental music, then that would serve as evidence of this universal emotional code—at least within the domain of affective sound. The speechiness-scale is a hypothesis, but it is hard to argue against that instrumental music is the sound-type (across speech and music) which is the least 'speechy' compared to natural speech. If vocals were included in the music, then it is possible that a model trained on SER would be able to take advantage of the emotional features carried in the voice of the singers, and potentially ignoring the features of the other sounds. Therefore it makes sense to explore the overlap between these two extremes of the scale—natural speech and instrumental music. As an additional motivation, Cañón et al. (2021) recommended future research to pre-train on emotional speech (as opposed to non-emotional speech) for transfer learning to MER, as this might result in transferable emotion-related features. Yet, a prerequisite to all of what has been shared above is to first explore and get familiar with *both* these sound emotion recognition domains: SER and MER. The main approach for this is a structured literature review (SLR), which was conducted during the preparation project (fall 2021). The results of this review is included in this thesis (chapter 3).

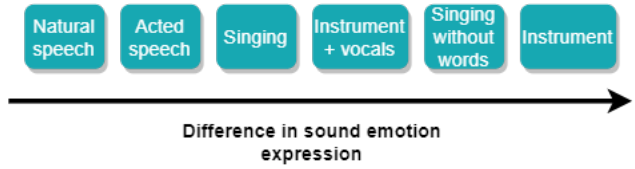



Figure 1.1: A hypothesized scale of the difference (speechiness) in emotional expression, from human to instrument

Before diving deeper into anything else it is useful to get a grasp of how *emotion* can be defined. Note that there has so far been little consensus around emotion within the field of psychology—and venturing to create the perfect definition will not be within the scope of this thesis. The 'sub-components' of emotion, however, are more agreed upon, according to Eerola (2018). These are: appraisal (how dangerous one assess a situation to be), expression (e.g. laughter), autonomic reaction (e.g. change of breathing rate), action tendency (e.g. flee the situation) and feeling (e.g. feeling the energetic quality of sadness within). In contrast, the personal standpoint taken here is that *emotion* is only the energetic activation of a certain quality/frequency, of which occurs in the plane of *feeling*, and that the other sub-components that Eerola mentions are simply other mechanisms that tend to occur simultaneously or as a response/reaction/interaction

to/with the current emotional state (which can also be more and less intense). Russell (2003) describes emotional life in a way that aligns more with my view. He offers that "emotional life consists of continuous fluctuations of core affect; a perception of affective qualities interacting with perceptual, cognitive and behavior processes. Occasionally, these components form one of the prototypical patterns [common and discrete emotions, like sadness or anger], just as the stars form constellations". However, righteousness aside, the reader is invited to make up their own mental map of *emotion*, and in this thesis *emotion* will be used as a broad concept that can encompass both all and some 'components' of *emotional life*, as described by Eerola.

Some of the applications of SER are:

- Improving mental and emotional health care, through assisting therapists in recognizing the emotional state of their patients (and also tracking mental health degradation), especially since explaining one's emotional state through words can be a challenge (Dossou & Gbenou, 2021; Feng & Chaspari, 2021).
- Improving customer-care services by classifying and prioritizing feedback based on the emotions of the customer's voice data (Dhondge et al., 2022).
- Improving human-computer interaction by automatic SER, with applications such as gaming and interacting with robots (Praseetha & Joby, 2021). Recognizing humans' emotional state accurately will make human-computer interaction more smooth and harmonious (Liu et al., 2018).
- Better classification of speech emotion enables better synthesis of (natural-like) speech, for instance through generative adversarial networks (GAN).

Some of the applications for MER are:

- Personality studies linked to music listening behavior (Anderson et al., 2021)
- Easier information retrieval for songs, like automatic playlist generation for the user, based on selected or current mood, as well as improving music recommendation systems in general (Agrawal et al., 2021; Preeti GUPTA, 2021; Xu et al., 2021). Additionally, music-based purchase behavior prediction (Xu et al., 2021). This also has relevance in exercise settings (Griffiths et al., 2021).
- To track the emotional progressions within soundtracks, and analyze the distribution of emotions of a musical composition (Grekow, 2021).

- Music can be used therapeutically (both in healthcare and personal settings), and emotional content of music then becomes relevant. Most people already listen to music for relaxation, inspiration, energy, to express feelings and emotions, find relief and reduce stress and agitation (Griffiths et al., 2021; Kumar & Gupta, 2021). Music may also lead to reduced symptoms of depression and anxiety (Brewer & Rahman, 2020).

- With good MER systems, auto-emotion-labeling of music data would improve training of automatic music composition models. Effective MER models can also enable better discriminators in GANs, both for music composition and other areas.

- Coaching systems for vocalists and instrumentalists, to give them live feedback on how their musical performance is conveying emotion.

- The level of arousal (energetic stimulation of the body) of music can modulate mood and affect the performance of cognitive tasks (Nadon et al., 2021). This is another argument for being able to predict the arousal level of songs through a successful MER system.

So what about the connection between emotional code in speech and music? The idea of a close relationship between music and human voice has a long history and is a well-supported idea also today (Cañón et al., 2021; Juslin & Laukka, 2003). Both domains (modalities) are nonverbal channels and rely on acoustic signals for transmitting emotional messages. A hypothesis, which is supported by many researchers, is that speech and music evolved from a common origin (Juslin & Laukka, 2003, p.770). The famous composer Richard Wagner said that "the oldest, truest, most beautiful organ of music, the origin to which alone our music owes its being, is the human voice" (Juslin & Laukka, 2003, p.774). Additionally, in the lines of an evolutionary perspective, vocal expressions of discrete emotions (e.g. fear or sadness) usually occur in similar types of life situations in different organisms. It is also obvious that the same sentence can be pronounced in several ways, and can be verbalized with multiple emotional colors (prosody). In principle it is possible to distinguish between the verbal message content and additional features of acoustic realization of the vocalized sound. Analogously, one musical piece can be played in numerous ways, and can convey different emotional content to the listeners. When it comes to differences, 'harmonic progressions' is a feature of musical expression that is believed to have no direct counterpart in speech (Juslin & Laukka, 2003). Juslin and Laukka conducted a literature review to compare speech and music performance, where 19/49 music studies used human singing voice as instrument (of such studies, some included words too). Recognition of discrete emotions had approximately *equal* accuracy

for intra-cultural speech, inter-cultural speech and music performance (no cross-domain in that result). Their results also strongly suggest that music performance and vocal expression use largely the same emotion-specific patterns of acoustic cues (features), and that these cues can be used to communicate discrete emotions (Juslin & Laukka, 2003). Human neurology is another interesting area with findings that support similarity between the expression of emotion in speech and music. A core neural network in the brain has been suggested, that facilitates the decoding of emotional information from all sources of affective sound (e.g. speech and music) (Frühholz et al., 2016). Yet the relative weights of sub-regions of this network varied across affective sound types, and some brain regions outside of this core net responded only to specific sound types, in addition to the core net. Phrased differently, the sound domains had some *common* and some *specialized* associated brain areas. Frühholz et al. suggested that future research investigates cross-domain questions and to transfer research paradigms across domains of affective sound.

## 1.2 Goals and Research Questions

The following goal is declared for this Master's thesis:

**Goal** *Use transfer learning to map from speech emotion recognition to instrumental music emotion recognition.*

What is the motivation for looking at both fields? As introduced earlier there is strong evidence for similarities. Additionally, based on the support, it is probable that transfer learning can be used to improve performance in the domain with less data and poorer results available, which is suggested to be music emotion recognition (MER) (result from the preparation project, and also echoed in Coutinho and Schuller, 2017, p.20). The reason for choosing *instrumental* (non-vocal) music specifically is that these have the least amount of similarity, in terms of speechiness (Cf. fig. 1.1). The focus here is to seek new evidence for a shared emotional representation in affective sounds, instead of the potential applications of a great speech or music emotion recognition system. Transfer learning from speech to music domains has been attempted before for music that contains vocals (Coutinho & Schuller, 2017), and for the time-continuous (as opposed to static) emotion recognition problem (more on these definitions later). It has also been attempted in one small-scale experiment for the case of acted speech and instrumental music (continuous ratings; Coutinho et al., 2014). When considering the setting of static emotion recognition, natural speech and instrumental music, this has not been attempted before. In general, large-scale transfer learning from emotional speech to instrumental music emotion recognition has not been done before this thesis.

Some research questions are defined in order to achieve the research goal, below. For context, an automatic sound emotion recognition study has some dependencies of choices: the choice of emotion taxonomy affects what datasets that can be used, and the labeling in the dataset influences the design of the machine learning problem. Next, the features to work with (extract from the sound-files) influences the design of machine learning architecture.

**Research question 1** *Are categorical or dimensional emotion taxonomies superior to the other, for emotion recognition from sound?*

Is any main emotional taxonomy (categorical vs. dimensional) superior in terms of validity and reliability? Validity here refers to its capacity for distinguishing and representing as wide range as possible of human emotional experiences. Reliability concerns whether a (preferably non-expert) user who wishes to map some *perceived* emotional content into model values, can do so in a way that enables highly agreeing values across attempts and across users. For instance, a model that has high validity can still result in no agreement when e.g. 10 annotators are to rate the perceived emotional content of a musical segment. A very detailed assessment of validity and reliability will not be prioritized, but a high-level treatment can still give useful directions for this project, as the choice of taxonomy strongly impacts the choice of datasets.

**Research question 2** *Can feasible datasets for both speech and instrumental music emotion recognition be acquired?*

A dataset will be needed for speech emotion recognition (SER) and music emotion recognition (MER) for this thesis, in order to do any machine learning (ML). There are plenty of SER datasets out there (of varying quality), but it is currently unknown, personally, if any instrumental MER datasets exist. A MER set needs to be either found or created. Feasibility here first and foremost needs to include the use of a suitable emotional taxonomy. These datasets should both be annotated with the same emotion taxonomy, which has to be applicable in both domains and preferably be as valid and reliable as possible. In addition, if efforts are required to create a feasible MER dataset, then it makes sense to choose to construct such dataset in accordance with the emotion model with the most potential applications.

For speech emotion recognition (SER), naturalness and variability in verbal content is of high priority for the data. For MER, a dataset needs to consist of samples that do not contain vocals (lyrics) nor singing voice. An original recording/song could be further split into multiple *clips* (this term is used interchangeably with *samples*, which represents the final data-units served to the ML-pipeline). All clips across SER and MER datasets will also need to be of a

length which enables an acceptable amount of information for decision making, yet is not so long that the emotional content of that segment varies too much.

Within MER there is a distinction between static and dynamic emotion recognition. The former entails one emotion annotation per clip, while the latter points to annotations as a time-series across each clip. The reason for this is that the emotional content within a song can vary drastically. In this project static annotations will be used for MER, since this is also most common for SER, and it is seemingly an easier ML-problem. If the experiments in this work are successful, it would be possible to explore dynamic emotion recognition in the future.

**Research question 3** *How does training on emotional speech affect recognition performance for instrumental music emotion recognition?*

The aim here is not to beat any state-of-the-art for MER (comparisons across datasets are difficult anyhow), but rather to look into how the transfer learning can affect performance. This means that model choice for the experiments could be based on or reproduced from any of the ones reviewed in related work, even though it may not be the best performing one.

## 1.3 Contributions

A structured literature review (SLR) was carried out in the preparation project, of which results are carefully analyzed and discussed in light of the research questions of that project. Additionally, some extra related work are added now. Discussions are provided around what emotional taxonomy is most suitable in this context. Further, a new MER dataset is compiled by carefully combining instrumental subsets of existing music datasets. A custom Keras-layer is implemented for a *dilated LSTM*, which is optimized and compatible with Tensorflow's graph compilation. ML experiments are conducted with two customized neural network architectures (DCNN and ADCRNN). The ML models are trained on SER data and directly evaluated on instrumental MER data (i.e. 'direct' transfer learning). Implications of the experimental results and future work recommendations are also included.

## 1.4 Thesis Structure

This work is structured as follows: Chapter 2 explains necessary background theory: emotion models, features and AI methods; Chapter 3 presents the SLR protocol and reports on the related work included in the final SLR set of articles, as well as additional articles; Chapter 4 describes the choice of emotional

taxonomy, the dataset and the experimental method; Chapter 5 presents the experiments; Chapter 6 discusses the results and implications; Chapter 7 concludes the thesis.

# Chapter 2

# Background Theory

Background theory is provided to support the reader with relevant concepts for this project. What is considered relevant is determined by what is used and of importance in the rest of the thesis. Throughout the entirety of this document it is assumed that the reader already has some fundamental experience in AI, such that only the more advanced of AI concepts will be explained, in general. Providing mathematical-details is only done when truly necessary, since I believe it is more useful to understand (or learn) the conceptual idea of *any* concept on a high level first, before potentially deciding to dive into more details later. This report also covers *many* techniques and concepts, and it would have been infeasible and irrelevant to have covered everything in the fullest detail. All of this chapter's sections are reused from the preparation project, though some have been refined.

## 2.1 Emotion Models

A foundation for emotion recognition tasks is the choice of how to define a space of which to recognize *emotion* (of *some* form) from. The different approaches to do so are called emotional taxonomies, or models. This choice also determines how a dataset is to be labeled, which signifies the significance of this choice, since bad data into an ML model means bad results output from it. There have been proposed many ways to do this, and the approaches can be grouped by two main paradigms: *categorical* and *dimensional* emotion models. In categorical approaches, the emotions are considered discrete (e.g. sadness, happiness). On the other edge, the dimensional view portrays emotional experience as something that is fundamentally representable through one or more continuous dimensions. Some popular emotion models are introduced below.

## Categorical Models

For categoricals, the key question would be which emotions are to be considered the (basic) emotional *families* of which all (discrete) emotional experience can be derived (either as nuances or combinations). According to evolutionary theory, these basic emotion types would have developed as important functions for survival, and is also common across all animals (Juslin & Laukka, 2003). Ekman et al. has proposed what is known as *The Big Six*—namely anger, disgust, fear, happiness, sadness and surprise. These might be the most agreed upon or referenced set of basic emotions (these have also been validated by cross-cultural studies; Ekman et al., 1987). That was up until 2016, at least... Today, Ekman and others (including the Dalai Lama)[1] seem to agree that the basic emotion families are anger, disgust, fear, sadness and enjoyment (renaming happiness and removing surprise; Ekman, 2016). Yet, the authors of emotion recognition work (among other fields) are free to choose whatever set of basic emotions they want, and also to include emotion labels in their datasets that they do not even consider to be *basic* emotions. Simultaneously, the upper limit set-size has usually been 14 among the studies that have proposed actual sets of basic emotions (Scherer, 2005).

## Dimensional Models

Moving over to the dimensional models, one very popular one is the 2D circumplex model (Russell, 1980). This consists of the dimensions *arousal* (sleepiness-arousal) and *valence* (displeasure-pleasure), and as such it is commonly referred to as the *VA-model* or *VA-plane*. Arousal represents the level of physiological arousal/excitation, and if high, a person experiences strong energetic bodily (physiological) sensations (Shuman et al., 2015). With high valence, the situation (that triggered the emotion) is "experienced as pleasant and enjoyable and/or is [believed to be] likely to have positive and desired consequences for the person" (Shuman et al., 2015). The two axis also form 4 quadrants, which is a popular way to convert the model into a 4-class classification problem, even though this perspective was not introduced in Russell (1980). The model is represented in fig. 2.1. Notice that concept of arousal is not the same as emotion intensity, and that emotion intensity is not a dimension itself, nor attribute, in the original model. The clear argument for this separation between arousal and intensity is that one could feel *intense* sleepiness, but the emotion of sleepiness (which, according to Russell, is an emotion) would have the lowest arousal of all emotions, according to the model. Russell admits his circumplex model of affect is far from perfect and is not the ultimate emotional model, yet it can account for a large

[1] http://atlasofemotions.org/#introduction/
(visited Feb, 2022)

part (58.9%) of the variability of 28 subject-reported affective states (emotions and feelings) (Russell, 1980). In February 2022, Russell's article has been cited over 16000 times.

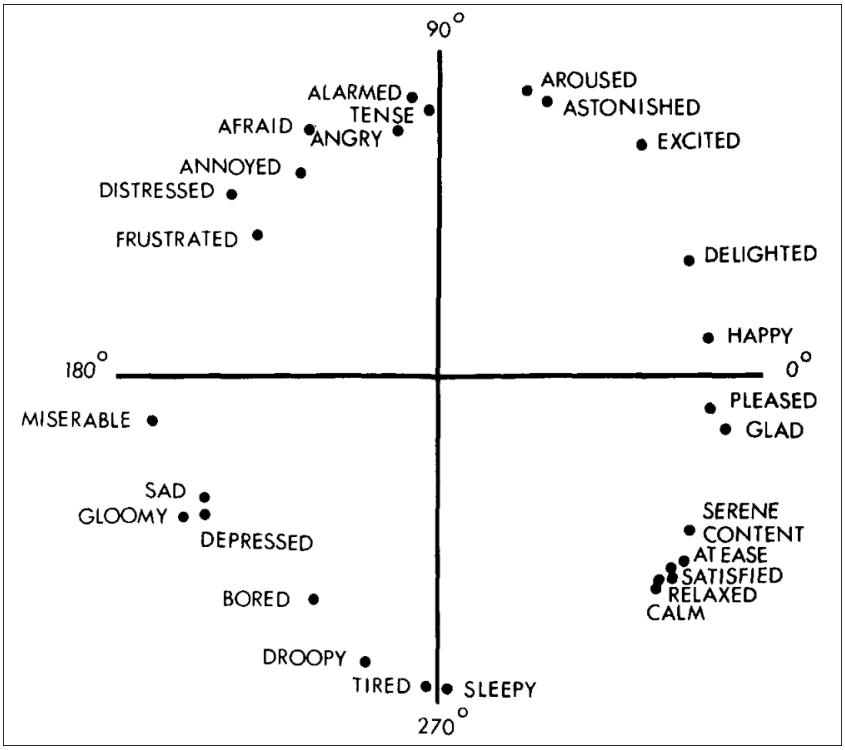



Figure 2.1: Russell's circumplex model. The affect terms are ordered in a circular order, counter-clockwise. The horizontal axis is valence, and the vertical is arousal. When referring to quadrants, the top-right quadrant (defined by the axis) is quadrant 1, and we count quadrants counter-clockwise.

Source: Russell (1980), with explicit permission.

Since one might like to use different words to explain the same type of emotion, but of varying strength (e.g. nervousness and terror), it is worth discerning between what is meant as an emotion and what is an *emotion family*, or in other wording, what is a *basic* emotion. This distinction is a foundation of another dimensional model: *The Geneva Wheel of Emotion* (fig. 2.2), which proposes 20 basic emotion families and 5 levels of strength per emotion family (Scherer, 2005). Scherer also arranges these families circularly in a 2D plane formed by the dimen-

sions of *control* (vertical and highest value at top) and *valence* (horizontal and highest value at the right). This model also can be segmented into 4 quadrants, like Russell's model. *Control* is sometimes called control/power and represents the degree a person with the emotion "believes that he/she can influence the situation to maintain or improve it" (Shuman et al., 2015). The fields "none" and "other" in center are meant for adding additional emotions perceived (e.g. in a song) that do not fit in with the other 20 emotions or if no emotion is recognized at all. Scherer (2005) also highlights that the proposed model is probably not the whole picture, and was intended to hopefully be a helpful contribution forward in the affective research fields.

## 2.2 Features

*Prosody* is the nonverbal aspects (features) of speech and music (and sound in general), which contain the emotional info, among other things. As the ultimate emotion recognition system from sound would be one that could represent the universal characteristics of emotions, there are countless proposals as to which extractable features (from audio) carry the information needed to do this. Originally these prosodic acoustic cues (features) have been investigated for each of speech emotion recognition (SER) and music emotion recognition (MER) individually. So far, it seems the ultimate feature combination has not been found, for any of the fields. Those features relevant for this report are explained here. Where appropriate, the main feature type (high-level category) will be focused on, and not the lowest detail level of the small sub-variations.

### Mel-spectrogram

*Mel-spectrogram* refers to a (acoustic) frequency vs. time vs. sound-intensity spectrogram (diagram), where frequency is in mel-scale and sound intensity is in *power* scale (normally). The mel-spectrogram representation will here be described top to bottom in terms of abstraction. The mel is a scale for pitch (frequency) that is scaled in a way that imitates how humans perceive differences in pitch, which is non-linear (Stevens et al., 1937). In standard pitch-scale (Hz), the interval between 1000Hz and 1000+$x$Hz sounds like a much larger gap than the interval from 10000Hz to 10000+$x$Hz. When the pitch is in mel-scale, each unit-distance of pitch sounds approximately equal to humans. An example of the *log* mel-spectrogram can be seen in fig. 2.3 (the log term is described later). Sound is represented digitally in audio-files as a waveform, which is a representation of how the sound pressure (sound-waves) changes over time, including its change of amplitude. A waveform (or sound) that carries multiple sound-frequencies is practically multiple waves traveling together, that oscillates at different frequen-

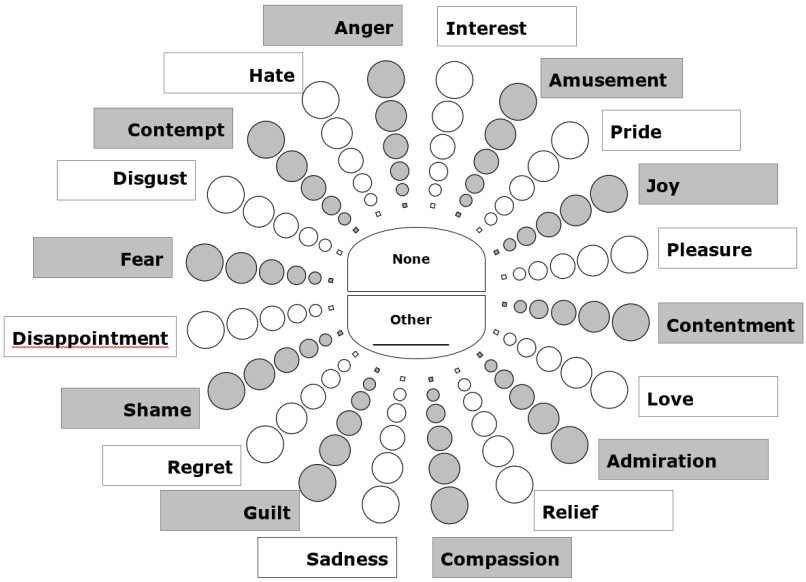



Figure 2.2: The Geneva Emotion Wheel

Source: Scherer (2005); Shuman et al. (2015), with permission[2]

cies. One question that remains to be explained here now, is how one can convert from a waveform to a spectrogram. A process named Short-time Fourier transform is a windowed process used to decompose the waveform into its amplitudes per frequency—over time. This process results in a time vs. frequency plane (spectrogram). Here, the amplitude is in energy-scale, which is the measured magnitude of the amplitude. This plane is then converted into power-scale (the squared of the amplitude-values) and the frequency dimension is converted into the mel-scale frequencies. The result is a mel-spectrogram! To be precise, when using the term mel-spectrogram, it means that power is the intensity scale, while *log* mel-spectrogram refers to the intensity being in decibel scale. Humans do not perceive the actual sound-intensity directly, but rather as *perceived loudness*, which is reflected by the decibel scale, which again is logarithmic.

## Mel-spectrogram Delta

It is possible to take the derivative of the mel-spectrogram in order to obtain a feature named (mel-spectrogram) delta. This can be compared to taking the

[2] https://www.unige.ch/cisa/gew/

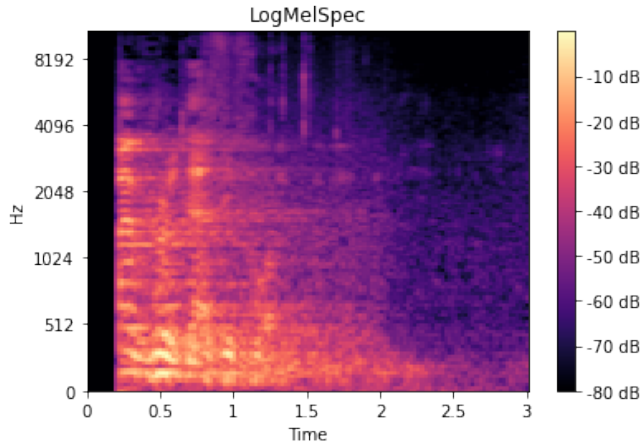



Figure 2.3: A log mel-spectrogram of a man saying "I've heard that before", with a small gap of silence in the beginning. The feature used for this plot has been mapped back to a log-scaled Hz on the y-axis, for human interpretability. Note that the color-dimension is in decibels.

derivative of any function. This delta is a useful method for extracting the mel-spectrogram's changes, in either of the frequency or time domain (axis). One usually analyzes changes in the frequency domain across time (which is to compute delta along the columns, when x-axis is time). The derivative of the delta obtains the double-delta (which behaves like any second-order derivative) (Tang et al., 2019).

## Modulation Spectrogram

*Temporal envelopes* of amplitude (volume) modulations are the perceived *changes* in amplitude of sound over time.[3] These temporal envelopes, which includes the envelopes' dynamics (e.g. patterns of changes in volume, like rhythm), can be measured across the whole frequency spectrum, but can also be done for separate frequency bins, like 'low', 'medium' and 'high' frequencies, or a large amount of bins which could yield a full spectrum modulation spectrogram. When separated into bins one can encode sound features that capture both spectral (frequencies) and temporal properties of the speech signal (Wu et al., 2011). A regular mel-spectrogram of course also displays spectral and temporal properties

[3] https://en.wikipedia.org/wiki/Temporal_envelope_and_fine_structure (visited Dec 9, 2021)

of the signal, but more high-level information like temporal envelopes are not as visible through this representation. There exist multiple types of modulation spectral features, and variants of *modulation spectrograms*. It is also possible to define modulation spectrograms of the temporal envelopes of (acoustic) *frequency* rather than amplitude, which would represent changes in frequency (e.g. pitch) over time. Wu et al. (2011) proposed some novel modulation spectral features, beyond a standard modulation *spectrogram*—and the reader is referred there for more info. Worth highlighting is that one of their features *alone* was able to almost completely separate 3 emotions across one well-known emotional speech database.

## Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) can be thought of as representing the shape of the vocal tract (which includes tongue, teeth, lips etc.), which is responsible for sound generation (of phonemes, which are distinct units of sound).[4] To get the *mel-frequency cepstrum*, one common derivation according to Jahangir et al. (2021) is to, in a summarized manner: for one segment of the sound do a Fourier transform (which gives a frequency vs. amplitude spectrum for that single segment only, not across wide time-step), then map this spectrum into mel-scale frequencies, then for each desired number of cepstral coefficients (which is a parameter) take the log of the magnitude of the mel-scaled spectrum, then finally do a *discrete cosine transform* (Ahmed et al., 1974) of this result. The final resulting spectrum is the mel-frequency *cepstrum*. This cepstrum is actually then a spectrum of a *transformed spectrum*, which means that the space is no longer about frequency, nor time, but rather something...intangible. The amplitudes of the cepstrum equal the *coefficients* (mel frequency cepstral coefficients). The whole process is in a way repeated with overlapping segments of the audio waveform, as in a short-time Fourier transform, in order to cover the full audio sample. This yields a time dimension in the final set of MFCCs. An example MFCC feature derived from the same audio as for the mel-spectrogram example is seen in fig. 2.4.

## Linear Predictor Coefficients

Linear Predictior (Cepstral) Coefficients are also cepstrum-based (i.e. a transform of a spectrogram, and a spectrogram represents time and frequency dimensions), like MFCC. They similarly try to represent the vocal tract, but uses linear frequency scale, as opposed to the mel-scale. It is based on a technique named

[4] https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd (visited Dec 9, 2021)

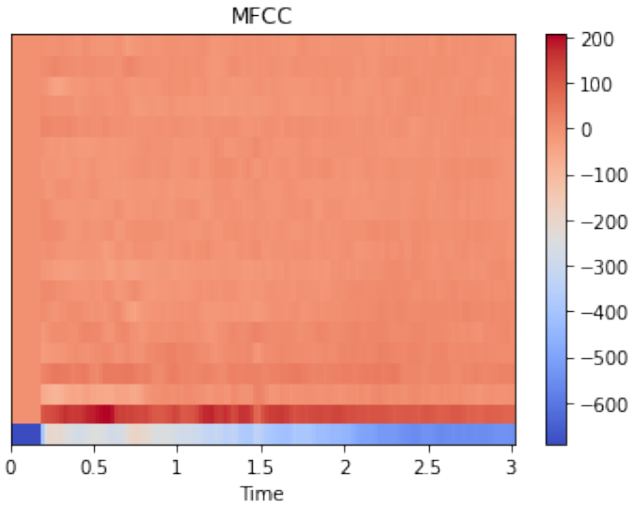



Figure 2.4: MFCCs of a man saying "I've heard that before". Except for the time dimension, it is not really that human interpretable. One can see that whatever 'activity' it shows in the bottom rows, is evaporating over time, similar to the power in fig. 2.3.

Linear Predictive Coding (LPC), which seeks to approximately predict the current sound-sample (or segment) based upon a linear combination of the previous samples (segments) (Jahangir et al., 2021).

## Tensor Factorized Mel-Spectrogram

When inputting a 2D image (or 2D array), such as a mel-spectrogram, into a neural network, one has to either vectorize (flatten) the image (which breaks spatial relationship of each pixel's value) or use a CNN (where the 2D representation needs to be flattened at some later point in the net) or use a tensor representation (this list of options is complete, to the best of my knowledge). Tensors for data representation can be used in e.g. a tensor factorized neural network, which is presented later in this chapter. A tensor can be viewed as a generalized extension of a matrix, which can be of any dimension (*order*, in tensor language), while a matrix can only be 2D. Some of the advantages of tensorized neural networks over standard CNNs are that it can lead to a model with significantly less parameters (e.g. up to 65% space savings without decreasing performance), it can maintain and even improve predictive performance (by better utilization of

the multi-dimensional connections in the data) and can be much more explainable (e.g. for tensor factorized neural networks) (Kossaifi et al., 2020; Pandey et al., 2022). So what are *tensor factorized* mel-spectrograms about? *Tensor factorization* (aka. *tensor decomposition*; De Lathauwer 2008) is a process that seeks to decompose a tensor into a combination of more 'meaningful' components, which results in a dimension-reduced tensor. In the application of this process on mel-spectrograms this could intuitively translate to components like meaningful patterns/features within the time dimensions (e.g. interesting changes over time) or frequency dimension. An example visualization of this process for mel-spectrograms can be seen in fig. 2.5. For comparison of the appearance of the *feature tensor*: if the output features of a *convolutional* layer were to be visualized, those activations would look very similar or highly correlated to the actual input spectrogram, while this is not the case for a feature tensor.

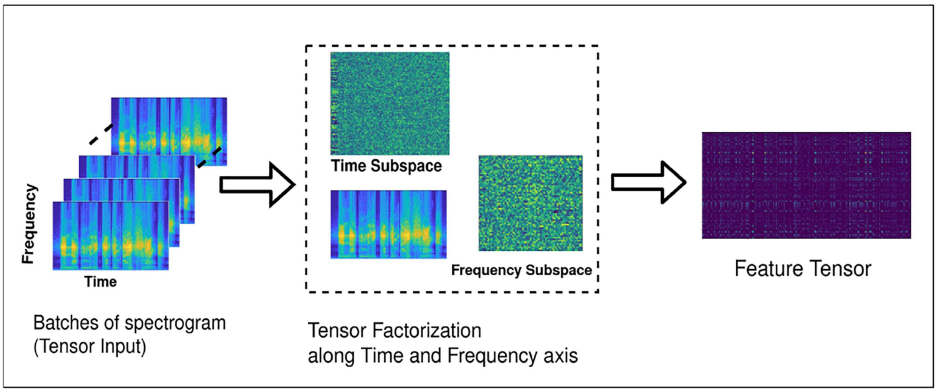



Figure 2.5: Tensor factorization of mel-spectrogram

Source: Pandey et al. (2022), with explicit permission.

## Low-level Spectral Features

The timbre, or 'color' of the sound can be described by several low-level *spectral* (i.e. frequency) features in combination, and some of them are listed below. For references and more detail on these features, please see the footnotes.[5,6]

[5] https://www.sciencedirect.com/topics/engineering/spectral-centroid (visited Dec 10, 2021)

[6] https://se.mathworks.com/help/audio/ug/spectral-descriptors.html (visited Dec 10, 2021)

- **Spectral centroid**: the center of 'gravity' or density of the frequency spectrogram (e.g. mel-spectrogram) in terms of amplitude, i.e. what frequency area is heaviest.
- **Spectral spread**: the deviation of the spectrum from the spectral centroid.
- **Spectral skewness**: measures the symmetry around the centroid (e.g. a signal can be very heavy in its lower frequencies).
- **Spectral entropy**: represents the quantity of information contained in a speech signal.
- **Spectral flux**: the variability of the spectrum over time.
- **Spectral roll-off**: the frequency where a certain percentage of the energy distribution (e.g. 80%) lies below.
- **Spectral flatness**: indicates the 'peakiness' of the spectrum. It can be used to distinguish between a noisy and a tonal signal.
- **Energy**: related to the total amplitude present across all frequencies, which is given by the area under the curve in a spectrum—or, for across the time domain, through the total area under the magnitude (simplified).
- **Chroma vector**: Chroma vector is a frequency-domain feature, which focuses on pitch classes (e.g. notes like C, C#, D, D# etc.). It is a measurement of how much energy that is present in each pitch class over time.[7]

## 2.3 AI Methods

Some conceptual explanations for the more advanced deep learning approaches are due.

### Convolutional Neural Network

Convolutional neural networks (CNN) (LeCun & Bengio, 1995) are popular for 2D and 3D images, but can deal with any 2D and 3D data. It is particularly useful for data that have spatial relationships, like pixels in an image. A high-level explanation is that a convolutional layer tries to 'explain' (e.g. through a dot product) one area of pixels (e.g. a 9x9 pixels filter) into a new and more abstract representation of that area of the image, which is done in a sliding-window fashion to cover the whole image. Through multiple convolutional layers

[7] https://en.wikipedia.org/wiki/Chroma_feature
(visited Dec 10, 2021)

one gets higher and higher level features (of abstraction). It is common to use a *pooling* layer in between a couple of convolutional layers. Their job is to slide over the image with a filter (e.g. 9x9 pixels) and summarize the content into only one scalar per area, which also reduces the resolution of that result, which can be negative if high detail is required for classification. One way to summarize (pooling) is to use *max pooling*, which yields the max pixel value found inside the filter.

### LSTM

An LSTM (Long Short-term Memory) (Hochreiter & Schmidhuber, 1997) is a recurrent neural network (RNN). An RNN is a network that inputs a time-sequence of feature-vectors, and uses the computational output of one input feature-vector to affect the computational output of the next feature-vector, and so on (horizontally across the time-steps). In this manner, one can say that the network incorporates memory across time. RNNs are common for time-series data and natural language processing. Imagine how one word in a sentence can affect the interpreted meaning of the next word. The LSTM cell is an improvement over a regular RNN, which in short adds mechanisms for forgetting information (like human short-term memory) and control over how much information to keep and when the memory information is to start affecting the LSTM cell's output.

### Generative Adversarial Network

A generative adversarial network (GAN) (Goodfellow et al., 2014) is an approach, which is popularly used for synthesizing images (other applications are also possible). One common application is image-to-image translation, for instance trying to convert one car into another type of car. But the more 'vanilla' GAN can also be used to simply generate not-in-the-dataset unseen cars. At a high level, a GAN consists of two networks: a *generator* and a *discriminator*. The generator's job is to generate 'translated' images (the output), and it is evaluated on its ability to fool the discriminator. The discriminator's job is to learn what a 'real' image is (e.g. a real car), and to judge whether a generated image from the generator is a real or fake image. These two components in the GAN compete and seek to improve based on their experience with each other, for instance when the generator generates something that the discriminator believes is a real image. It is common to use convolutional techniques together with a GAN, for its inner workings. A normal GAN needs labeled (supervised learning) 'paired images', where one input image is matched with a 'correct' paired output image, while an extension sub-architecture called *Cycle GAN* can work with unpaired images (e.g. having two independent and unmatched datasets of cats and dogs).

### Attention Layer

A common technique nowadays in deep learning is to include an *attention* mechanism (Bahdanau et al., 2014) in the network architecture. The attention mechanism has been one of the most influential ideas in deep learning, and especially natural language processing. In short, the attention layer enables giving each time-step of the input different importance in relation to the predicted class.

### Dilation

For CNN (dilated CNN), the method *dilation* (Yu & Koltun, 2015) is about expanding the width of the convolutional sliding filter (the *kernel*). By doing this, the *receptive field*, i.e. how large of an area in the image that affects the value of one node (pixel) in the next image (feature) layer of the neural network. With a larger receptive field, more information can be brought forward into the next feature layer. Some advantages of using dilation over e.g. pooling (which is used for similar purposes) is that dilation does not increase computation and memory costs and simultaneously preserves the resolution of the next feature layer. Pooling, on the other hand, decreases resolution, which means that information needs to be compressed. Just imagine how information gets lost when downsizing an image in general. As mentioned, this can be negative if high detail is required throughout the network for doing accurate classification.

Dilated RNNs have also been proposed (Chang et al., 2017, "DilatedRNN"). The idea of dilation here is to add skip-connections between RNN time-steps (horizontally across one layer, i.e. across time-steps i.e. input features, like words). For a dilation-rate of 2, each time-step will receive input from the second previous time-step (instead of the previous) and the current time-step of the previous layer. This technique tackles known RNN-challenges with: vanishing/exploding gradients (weight updates that disappear or explode because it needs to be propagated through many time steps), complex long term-dependencies (like between the beginning and ending words of an utterance) and efficient parallelization (Chang et al., 2017). The DilatedRNN reduces parameters needed in the network and improves training efficiency.

### Tensor Factorized Neural Network

Chien and Bao (2017) introduced a tensor factorized neural network (TFNN). To understand how this architecture works, it is useful to build upon what was shared in section 2.2 about tensor factorized mel-spectrograms. The concept of tensor factorization was introduced, and how this results in a *feature tensor* (fig. 2.5). The same procedure is repeated for each layer in the TFNN. The TFNN can be seen as a generalized neural network, as tensors *extend* the standard 2D matrix.

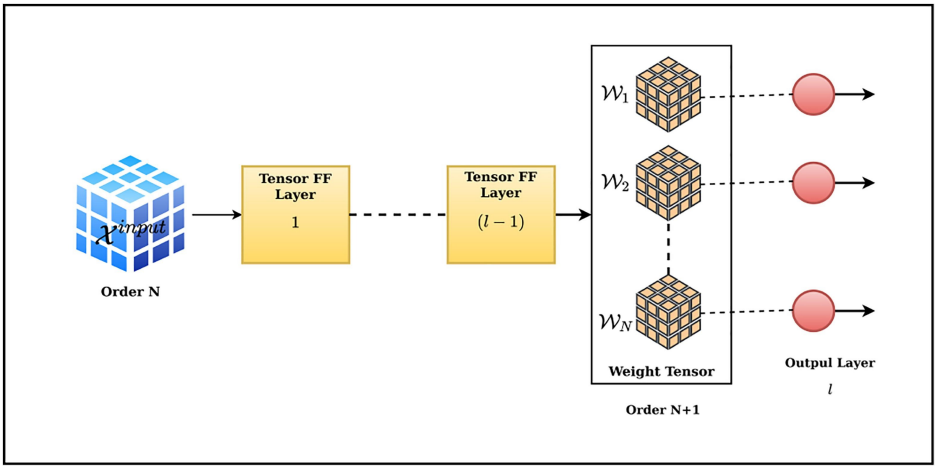



Figure 2.6: Tensor factorized neural network (TFNN)

Source: Pandey et al. (2022), with explicit permission.

Another difference is the use of *tensor-factorized error backpropagation* (Chien & Bao, 2017), which extends (generalizes) the stochastic gradient descent. fig. 2.6 visualizes a TFNN, which on a high level indeed looks similar to a standard feed-forward neural network.

## 2.4 Evaluation Metrics

### Krippendorff's Alpha and Cronbach's Alpha

The Krippendorff's alpha and Cronbach's alpha are two metrics that can be used to measure for instance inter-rater agreement (e.g. multiple raters who rate the same set of objects), which is the case in the reviewed work in chapter 3 and chapter 4. A detailed walk-through of these are not needed, other than knowing that a perfect agreement results in a score of 1.0, and they both decrease in value together with worse agreement.

# Chapter 3

# Related Work

This chapter walks through the structured literature review (SLR) planning phase and execution, then reports the extracted data from the SLR-set of papers (i.e. the resulting set from the SLR), and finally some related work which are even more specific to this thesis (transfer learning between the domains) is reviewed. This final part did not go through a structured review phase, mostly due to the low amount of such work. However, the quality of these articles were also taken into account before deciding to include them. All of this chapter is from the preparation project, except section 3.2.5 and section 3.3.

## 3.1 Structured Literature Review Approach

As an unbiased way to get sufficient knowledge within the fields, and high chance of identifying the state-of-the-art of speech emotion recognition (SER) and music emotion recognition (MER), a SLR was carried out. The approach used here is based on Kofod-Petersen (2018). This process is split into three steps: planning, conducting and reporting. The SLR was conducted during the preparation project for this thesis.

### 3.1.1 Planning the Structured Literature Review

A review protocol was developed as part of the planning phase. This enables reproducibility, as well as guiding my own work. The full protocol (and more reasoning behind decisions) is included in appendix A.1, while a more concise version is given here.

The research goal and research questions (RQs) that applied to the fall project (PP) are as follows:

**PP.Goal** *Give an overview of the research fields of automatic music emotion recognition and speech emotion recognition.*

Another important aspect of the goal is to achieve a good understanding of the fields, in order to support the creation of anchored theme-suggestions for my upcoming Master's thesis. The experience from this current project will hopefully also serve as good decision foundations when developing the thesis.

Some RQs were created in order to achieve the PP's research goal:

**PP.RQ1** *What is state-of-the-art for music emotion recognition and speech emotion recognition?*

More specifically, what emotion categorizations were used, what datasets were used and which domain has the least high-quality data publicly available? How do the solutions proposed compare w.r.t. method and performance?

**PP.RQ2** *Do the findings to PP.RQ1 support the use of transfer learning between the two domains?*

Do the findings in PP.RQ1 show that one of the domains have achieved poorer results (state-of-the-art) than the other (perhaps due to too small amount of good data or less sophisticated methods)? In that case, if possible to use the same emotion taxonomy in both domains, or to translate between them, it would be interesting to attempt transfer learning to boost the performance in the weakest domain.

**PP.RQ3** *Do the findings motivate other future work that is suitable for a Master's thesis?*

Perhaps the findings in PP.RQ1 inspire new personal ideas or directly suggest further work, which are suitable for my Master's thesis (Spring 2022).

### 3.1.2 Conducting the Structured Literature Review

The conduction of the review contains five steps (Kofod-Petersen, 2018): identification; selection; quality assessment; data extraction.

#### Step 1: Identification of Research

Google Scholar was chosen as the search engine, due to its inclusion of 'articles that cites' one specific article, and the fact that it searches through multiple sources. Based on some 'snowballing' (exploring the network of references) from Djupvik (2020) a set of search terms were chosen. Since the SLR goal is divided

across two domains (SER and MER), the SLR was also divided in two. Therefore the two search queries used in Google Scholar (and filtered by "After 2021") were:

```
   allintitle:(Speech)
(Mood OR Emotion OR Affect OR Ambiance)
(Classification OR Recognition OR Detection)

   allintitle:(Music OR Musical)
(Mood OR Emotion OR Affect OR Ambiance)
(Classification OR Recognition OR Detection)
```

The search was conducted 27th of September 2021, and revealed 373 SER and 55 MER articles during the previous year. This alone shows that SER is the most popular field of the two domains. Even though the results were sorted by date and not relevance, the highly specific query string and using the "allintitle" command (searches only article-titles), yielded mostly relevant results. Note that the term *prediction* was not included, since it was overlooked. This probably excluded many relevant articles that had explored affective sound emotion regression, since those articles are likely using *prediction* in the title.

#### Step 2: Selection of Primary Studies

In line with the protocol (appendix A.1), only the first 25 articles from each of MER and SER searches which passed the primary inclusion criteria was brought forward. Similarly, only the first 12 articles that passed the secondary inclusion criteria was progressed into the quality assessment (step 3).

#### Step 3: Quality Assessment of Studies

The amount of articles from each domain in the *SLR-set*—the set that passes the final quality assessment—was affected by remaining time resources at execution time. The SLR-set contained 5 articles from each domain. The quality criteria ratings (scored 0, 0.5 or 1) are available in appendix B.1. Note that one of the SER articles was later excluded, since it was discovered that their method-part was not as reproducible as first judged to be, since it became clear that no features used in the experiment were specified.

#### Step 4: Data Extraction

For each article in the SLR result set, the following data points are extracted and summarized:

- Unique ID

- Author(s)
- Publication year
- Title
- Emotion taxonomy used
- Dataset (and whether acted/natural)
- Machine learning method(s)
- Features used
- Findings and conclusions (and whether cross-validated)

**Step 5: Data Synthesis**

The result of the SLR data extraction is displayed in table 3.1, and later analyzed thoroughly.

## 3.2 Structured Literature Review Results

The result of the SLR is summarized in table 3.1. The articles are represented row-wise and their ID is prefixed by SLR, then with MER or SER. *EA* means et al., *CV'ed* means final results stems from cross-validation. If tuning has been done without a separate test-set (which is the case when doing CV), the generalizability of the results is questionable. The following sub-sections further consider important themes related to the extracted data—i.e. emotion taxonomies, datasets, features and methods plus findings—for each article from the SLR. These aspects are arguably the building blocks for the findings of any affective sound study.

Table 3.1: SLR data extraction

| Article ID | Author | Year | Title | Emotion taxonomy | Dataset | Machine learning method(s) | Features | Findings and conclusions |
|---|---|---|---|---|---|---|---|---|
| SLR. SER1 | Pandey EA | 2022 | Attention gated tensor neural network architectures for speech emotion recognition | Categorical (anger, happiness, neutral, sadness) | Emo-DB, IEMO-CAP (both acted) | 3D attention gated tensor factorized neural network (3D AG-TFNN), Parallel AG-TFNN | log mel-spectrogram and 3D log mel-spectrogram tensor (with delta, double delta), modulation spectrogram | 3D AG-TFNN (for Emo-DB, acc=85.15%) and Parallel AG-TFNN (for IEMOCAP, acc=55.56%) reached approximately state-of-the-art results with less parameters and less computational complexity than the baseline CNN+LSTM (CV'ed) |
| SLR. SER2 | Throung Pam EA | 2021 | Hybrid Data Augmentation and Deep Attention-based Dilated Convolutional-Recurrent Neural Networks for Speech Emotion Recognition | Categorical (anger, boredom, disgust, fear, happiness,neutral, sadness) | Emo-DB (acted) | attention dilated convolutional recurrent neural network (ADCRNN) | log mel-spectrogram, delta, double delta | The data augmentation methods: time shifting, pitch shifting, WaveGAN and using a combination of softmax and CT-C loss resulted in acc=91.90 (CV'ed) on Emo-DB. The runner-up from the compared work was acc=85.39, which used a very similar model. Their results indicate that the data aumentation techniques had the greatest impact on the improvement, compared to the referenced runner-up. |
| SLR. SER3 | de Lope EA | 2021 | Speech Emotion Recognition by Conventional Machine Learning and Deep Learning | Categorical (neutral, calm, happiness, sadness, anger, fear, disgust, surprise) | RAVDESS (acted) | kNN, SVM, random forest, multi-layer perceptron, CNN | log mel-spectrograms, MFCC. They applied data augmentation | One experiment found a strong trend for accuracy improving when number of emotion classes decreased. For the main experiment, SVM with polynomial kernel achieved the best accuracy of 71.2% (CV'ed). This result is competitive with results reported for RAVDESS |

**Table 3.1 continued from previous page**

| Article ID | Author | Year | Title | Emotion taxonomy | Dataset | Machine learning method(s) | Features | Findings and conclusions |
|---|---|---|---|---|---|---|---|---|
| SLR. SER4 | Li EA | 2021 | Emotion recognition from speech with StarGAN and Dense-DCNN | Categorical (neutral, calm, happiness, sadness, anger, fear, disgust, surprise) and subsets of these | RAVDESS, SAVEE, Emo-DB, CASIA (acted) | Dense-DCNN | log mel-spectrograms | The log mel spectrograms were better input features than MFCC and raw spectrograms (with their method) across all datasets. They used StarGAN to expand each emotion to 2000 samples each. When expanding, one dataset reached 97% accuracy, compared to 83% when not using data augmentation. They beat the runner-up with a few percent on each dataset (weighted accuracy, CV'ed). |
| SLR. SER5 | Zhao EA | 2021 | Cross-Corpus Speech Emotion Recogntion Based on Sparse Subspace Transfer Learning | Categorical (anger, disgust, fear, happiness, sadness) | Emo-DB, eNTERFACE, RML (acted) | Sparse subspace transfer learning (SSTL) | Not specified, but one primary reference use 12 features, including MFCC, log mel-spectrogram, delta, double delta | SSTL was better than all compared methods in 4/6 cross corpus experiments, and is far superior on average, where it reached 47.11% accuracy (average over 50 runs). This shows SSTL can obtain a more transferable feature representation. |
| SLR. MER1 | Krishnaiah and Divakarachari | 2021 | Automatic Music Mood Classification using Multi-class Support Vector Machine based on Hybrid Spectral Features | Categorical (Indian ragas) (sympathy, seriousness, peace, sadness) | Hindustani music dataset (HMD), Carnatic music dataset (CMD) | SVM | Spectral spread, spectral centroid, spectral skewness, MFCC, linear predictor coefficients | The combined features and the SVM yielded 97.53% accuracy, better than all baseline models in experiments and previous work on the same dataset. (Not CV'ed) |

**Table 3.1 continued from previous page**

| Article ID | Author | Year | Title | Emotion taxonomy | Dataset | Machine learning method(s) | Features | Findings and conclusions |
|---|---|---|---|---|---|---|---|---|
| SLR. MER2 | Cañon EA | 2021 | Language-Sensitive Music Emotion Recognition Models: Are We Really There Yet? | Dimensional (quadrants in VA-plane) | 4Q, CH818 | SCAE, CPC | mel-spectrogram | Pretraining on a mixture of speech languages (not emotional) may improve emotion classification in music. Their method does not appear to learn emotion-related features from speech that are transferred to MER. The F score for valence classification (high vs low) was consistently better than for arousal. The SCAE models was better than CPC. (not CV'ed, but train/val/test split) |
| SLR. MER3 | Farris EA | 2021 | Musical Prosody-Driven Emotion Classification: Interpreting Vocalists Portrayal of Emotions Through Machine Learning | Categorical (Geneva Wheel of Emotion, 20 emotions) and 4 quadrants (valence/control-plane) | Custom dataset of vocalists only | kNN, SVM (linear), random forest, extra trees, gradient boosting, MLP | prosody related features (Zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, MFCC, Chroma vector and deviation) | Achieves good accuracies on 20 emotions taxonomy for datasets of single singers (49.1%) and datasets of multiple singers (43.8%). For the 4 quadrants on 3 singers the accuracy was 68.8%. Their feature experiment showed performance well maintained by using small subset of the original total features. (not CV'ed, but train/val/test split) |

**Table 3.1 continued from previous page**

| Article ID | Author | Year | Title | Emotion taxonomy | Dataset | Machine learning method(s) | Features | Findings and conclusions |
|---|---|---|---|---|---|---|---|---|
| SLR. MER4 | Griffiths EA | 2021 | A Multi-genre Model for Music Emotion Recognition Using Linear Regressors | Dimensional (VA-plane) | Custom dataset of 20 songs (modeling) and 40 songs (evaluation)both with diverse genres | Linear regression (1 per emotional dimension) | Arousal: energy, standard deviation energy, median energy. Valence: spectral spread, median spectral spread, spectral flatness | Using only 6 of the initial 45 audio features yielded $R^2$ scores of 0.85 for arousal and 0.78 for valence. These results were either the same or significantly better than previous work. |
| SLR. MER5 | Grekow | 2021 | Music Emotion Recognition Using Recurrent Neural Networks and Pretrained Models | Dimensional (VA-plane) | GTZAN (324 songs) | LSTM | Up to 529 features | Seperate nets trained for arousal and valence prediction. Using pretrained model as feature extractor (from the feature vectors) contributed significantly to their best performance of $R^2$ equal to 0.73 (arousal) and 0.46 (valence). This compared to their strongest baseline model SMOreg with $R^2$ 0.48 and 0.27 for arousal and valence (respectively and CV'ed) |

### 3.2.1 Emotion Taxonomies

Within the fields of MER and SER there are no agreed way on how to model emotion (taxonomy) nor what emotions are considered universal. The reviews revealed two approaches being used: categorical and dimensional. The publications within SER only used a categorical approach, while for MER all except Farris et al. (2021) used Russell's dimensional circumplex model with the valence/arousal-plane (Russell, 1980). Cañón et al. (2021) worked with the 4 quadrants of the circumplex model. Farris et al. (2021) on the other hand, used the 20 categorical emotions given by the Geneva Wheel of Emotion (Scherer, 2005), as well as a quadrant approach with the 20 emotions mapped to a valence-control plane (classification problem).

### 3.2.2 Datasets for Speech Emotion Recognition

Within SER there were several public datasets that were used across multiple studies. Also, most of the SER studies ran experiments on several datasets. This sub-section explores the variety of datasets used in the field, and how they were built, since this is especially relevant for assessing the data quality later.

*Emo-DB* (Berlin Database of Emotional Speech) was released in 2005 (Burkhardt et al., 2005). It consists of 800 acted utterances in German, by 10 actors (gender-balanced) and 10 different sentences (global for all emotions). The set covers 7 emotions, namely: happiness, anger, anxiousness, fear, boredom, disgust and neutral. The creators of the dataset aimed for high naturalness, even though the speech is acted. The actors self-induced emotion by remembering a situation where they had felt the emotion strongly (known as *Stanislavski method*). During development, the dataset was evaluated through a perception test. Following the test, only audio samples with at least 80% emotion recognizability and judged as natural by more than 60% of the listeners were included. The Emo-DB dataset was used in "Cross-Corpus Speech Emotion Recognition Based on sparse Subspace Transfer Learning" (n.d.); Li et al. (2021); Pandey et al. (2022); Truong Pham et al. (2021).

Another dataset is the *IEMOCAP* (Interactive Emotional Dyadic Motion Capture) (Busso et al., 2008). It is acted by 10 actors (gender-balanced) and consists of 12 hours (10039 utterances) of multimodal data (e.g. video with audio and images). Only the audio samples are used in the reviewed studies, as defined by the secondary inclusion criteria of the SLR, this applies for all the following multimodal datasets too. In the production of the dataset, there were specifically designed scenarios designed to elicit (authentic) emotional expressions. These scenarios were either improvised or scripted. Its labeled emotions are: anger, happiness, excitement, sadness, frustration, surprise, neutral and "other". The samples are also labeled with 3-dimensional values (valence, acti-

vation and dominance) (Mehrabian & Russell, 1974). 3 students annotated each utterance, and reached discrete emotion agreement in 74.6% of the utterances (66.9% in scripted sessions and 83.1% in spontaneous sessions). The continuous ratings got Cronbach's alpha (section 2.4) agreement of 0.8, 0.6, and 0.6 for valence, activation and dominance, respectively. Pandey et al. (2022) experimented with the IEMOCAP dataset, but only the categorical labels.

*RAVDESS* (Ryerson Audio-Visual Database of Emotional Speech and Song) is an acted SER dataset (Livingstone & Russo, 2018). It is acted by 24 actors, vocalizing two sentimentally neutral sentences. Its emotion labels for speech are: neutral, calm, happiness, sadness, anger, fear, disgust and surprise. For singing, the emotions are: calm, happiness, sadness, anger, and fear. It holds 1440 speech files (60 per actor) and 1012 singing files (44 per actor, for 23 actors). Regarding validity, each file was rated 10 times on emotional validity, intensity and genuineness. The validation tests reported high levels of emotional validity, interrater reliability, and test-retest reliability. The dataset was used in de Lope et al. (2021); Li et al. (2021).

*SAVEE* (Surrey Audio-Visual Expressed Emotion) is also an acted dataset (Haq & Jackson, 2010).[1] Only the audio part is considered here. 4 male actors acted 120 utterances, totaling a dataset size of 480. The emotion labels are anger, disgust, fear, happiness, sadness, surprise, neutral. These 6 basic emotions have earlier been supported as *basic* by cross-cultural studies in Ekman et al. (1987). The dataset uses 12 unique sentences per emotion, and 3 common (global) sentences which are used for all emotions. Notice that this means that 12/15 sentences are unique for each emotion, per actor. For evaluating the dataset, an experiment of emotion recognition was done by 10 humans. Here the audio alone resulted in 66.5% accuracy (±2.5%). Only Li et al. (2021) used this dataset (among the reviewed papers).

*CASIA* (Chinese Emotional Speech Corpus) (as cited in Li et al., 2021) is an acted dataset for Chinese emotional speech. The original publication URL of the dataset is not responding, as of December 2021.[2] Due to this, no deeper data validation information was found. It is performed by 4 actors (balanced gender). The emotional labels are anger, fear, happiness, neutral, sadness and surprise. Each of these 6 emotions is expressed through the same common 300 short phrases, and in 100 phrases uniquely per emotion, resulting in 9600 total utterances. Only Li et al. (2021) used this dataset.

*eNTERFACE* is an audio-visual emotion database with recorded subjects who elicited (hopefully natural) emotional reactions to 6 different situations. Two human experts evaluated whether the people had expressed emotion in such a way

[1] http://kahlan.eps.surrey.ac.uk/savee/
(visited Dec 14, 2021)

[2] http://www.chineseldc.org/resourceinfo.php?rid=76

that untrained observers could recognize the emotional message in the reaction without ambiguity. The emotional categories are: happiness, sadness, surprise, anger, disgust and fear, which are the same basic 6 supported by Ekman et al. (1987). Each of the 6 situations had 5 pre-defined answers. The final dataset contains 1166 video sequences. These are from a total of 42 subjects, but not all were able to produce believable results for all the emotions. Of all the actors were 23% women and 77% men, and all were research engineers (in unrelated fields) and nobody were professional actors. The subjects were from 14 countries (but all experiments in English). In the dataset is 1166 sequences, with roughly 195 per emotion. Additionally, 31% wore glasses and 17% had a beard (which may or may not be relevant for SER). The eNTERFACE dataset was used in "Cross-Corpus Speech Emotion Recognition Based on sparse Subspace Transfer Learning" (n.d.).

*RML* (Wang & Guan, 2008) is another audio-visual emotion database. The utterances were collected from 8 subjects, which spoke 6 different languages (English, Mandarin, Urdu, Punjabi, Persian and Italian). For this dataset also, the goal was to elicit natural emotional responses. To achieve this, the subjects were provided some sentences describing emotional situations, and the subjects were encouraged to recall a similar incident from their lives. 10 reference sentences (for elicitation) per emotion were given to support context independence of the speech data (and the subject could choose between them). The subjects were free to express themselves using the same sentences as they had used for elicitation, make their own variations of them or completely different sentences. The emotional labels are the same basic six as before: happiness, sadness, anger, fear, surprise and disgust. To validate the data, at least two participants who did not know the corresponding language was asked to report the emotion they perceived (and the sample was deemed valid if both recognized it correctly). For the English samples, it required the correct perception of all 8 subjects. The set contains 500 samples (emotion distribution unknown). "Cross-Corpus Speech Emotion Recognition Based on sparse Subspace Transfer Learning" (n.d.) used this dataset.

### 3.2.3 Datasets for Music Emotion Recognition

In MER, none of the reviewed publications used the same dataset, across studies. This sub-section explores the variety of datasets used in the field, and how they were built, since this is especially relevant for assessing the data quality later.

Krishnaiah and Divakarachari (2021) used the *Hindustani music dataset* (HMD) and *Carnatic music dataset* (CMD), both of which released by CompMusic.[3,4] In these Indian music datasets, the concept of *ragas* are used, and also the object of recognition. Ragas are not exactly the same as emotions, but rather a melodic framework related to Indian melodic modes, which again are associated to different 'colors' or emotions. A raga is a melodic structure and Indian tradition considers them to have the ability to "colour the mind and affect the emotions of the audience... Each raga traditionally has an emotional significance".[5] The ragas in Krishnaiah and Divakarachari (2021) and their associated emotions are: Sindhu Bhairavi, Darbari, Saveri and Sri (sympathy, serious, peacefulness, sadness). The authors state that these ragas have diverse melodic attributes, which leads to better recognition results. The study used 480 audio recordings from CMD (today CMD contains 1889 recordings across all ragas, i.e. not only these 4). The HMD dataset had 300 recordings used in the study, but today the full amount is 970 (across all available ragas).

Cañón et al. (2021) used the datasets *4Q* (Panda et al., 2018) and *CH818* (Hu & Yang, 2017). 4Q contains 900 clips of 30-seconds (one clip per song), which are annotated with the quadrants of Russell's circumplex model (the model with valence-arousal plane). The set contains 225 clips per quadrant. In its creation, the AllMusic API[6] was used to obtain the 30-second clip and metadata per song, including "mood tags", which results from an expert-made emotion tagging system (which is not fully documented). Panda et al. (2018) do not know if the mood tags originated from analysis of audio, lyrics or a combination of both. The way the 30-second segment for each song was chosen was not clear either, and some contained some applause and noise, which was raising suspicion about their representability. Further in the dataset creation, the 289 unique mood tags were intersected with a list of English adjectives and respective mappings to the VA-plane. The final dataset were only those songs of which subjects in a manual blind test, annotating the perceived emotions with Russell's quadrants, agreed with those quadrant labels that resulted from the AllMusic mood tags.

The CH818 is a dataset with 818 30-second clips of Chinese pop songs, annotated with valence and arousal values (Russell's VA-plane). It was also one of the ones used in Cañón et al. (2021). Each song's chosen 30-second segment was the one of all 30-second segments (sliding window) that yielded the highest

[3] https://dunya.compmusic.upf.edu/carnatic/info (visited Dec 2, 2021)

[4] https://dunya.compmusic.upf.edu/hindustani/info (visited Dec 2, 2021)

[5] https://en.wikipedia.org/wiki/Raga (visited Dec 2, 2021)

[6] http://developer.rovicorp.com/docs (visited Dec 2, 2021)

combined $valence^2$ and $arousal^2$ values from a regression model. In other words; the strongest affective content. Each clip in the dataset was annotated by 3 Chinese music experts. The Pearson's correlation coefficients of the annotations were 0.842 for arousal and 0.794 for valence.

Farris et al. (2021) created a custom dataset of singing speech (vocalists' portrayal of emotion). Three professional singers were asked to improvise (consciously) as many phrases as possible for each of the 20 emotions in the Geneva Emotion Wheel. No emotion elicitation strategy was used here. The authors collected 4 to 6 hours of recording per singer, each with approximately 15 minutes per emotion. The phrases lasted between 1 to 20 seconds. The authors wrote they instructed the singers to "Sing anything for each phrase that you believe matches the emotion except use words". Email correspondence with one of the authors confirms this means "do not use words". Other instructions given to the singers were: to not attempt to control for different intensities of emotion and to mark any phrase that they believed did not capture the intended emotion. The phrases are annotated with ground-truth labels by the singers themselves.

Griffiths et al. (2021) collected a small multi-genre song corpus of 20 songs (actually 60, as explained below) and 20 genres for developing two regression models (valence and arousal). The songs are from Western culture. The small size was due to the study being a proof-of-concept. The motivation for developing a cross-section corpus was to facilitate generalizability of the system. 44 participants were recruited (mainly from an email-list specialized in auditory perception), and they each rated all 20 songs with distinct emotion (nominal) and emotional strength (ordinal) values for perceived and induced emotion. 8 emotion-options were given (which covers all quadrants of the circumplex model), but only 6 were selected by subjects: sadness, happiness, excitement, relaxing, anger, miserability. Those emotions not picked were fear and tiredness. Induced emotion means emotion induced in the listener, which can be different from the one perceived in the song. They were presented with a 1-minute sample of each song, from its mid-point of duration. Ratings for each song were statistically significant ($p < 0.05$), except one song in the induced ratings, and the emotions vs. strength were verified for independence via Pearson's Chi-Square test. Yet, songs were presented to the subjects in a fixed sequence, which was suspected to have created some bias due to participant fatigue. The final ratings per song were the mode of the ratings. The emotion terms were then placed around the circumplex model spaced with 45°, and the strength ratings were used as the radius from origin. In this way, valence and arousal values—which are values along the axis—could be calculated. Further, a validation (test) set of 40 new songs (2 per genre) was developed in the same way. So, the dataset in Griffiths et al. (2021) totals 60 songs. Note that Q2 and Q4 in the VA-plane had none to few samples represented in the validation set.

Grekow (2021) created an annotated dataset from the audio dataset *GTZAN*[7] (Tzanetakis et al., 2001), which totaled 324 songs. The samples are of 6 seconds of songs derived from the genres: classical, jazz, blues, country, disco, hip-hop, metal, pop, reggae and rock. 6 seconds was the shortest length that experts could detect emotions for a given segment. Data annotation with valence and arousal values was done by 5 experts with university musical education (who dealt with creation and analysis of emotion in music daily). Their annotations had very high agreement. The quadrants were similarly represented with samples, with an average of 81 per quadrant, and the valence-arousal values were not correlated.

### 3.2.4 Features

Most features are used across both SER and MER domains, so they are not separated in this description. Some studies utilized a very large number of features, like Grekow (2021), who tested up to 529 of them. This is too many to list here and to cover in detail in chapter 2, so only a selection of features will be treated here, prioritizing those which were used in several of the papers.

Pandey et al. (2022) used log mel-spectrogram, 3D log mel-spectrogram tensors (a 3D tensor of mel-spectrogram, deltas and double deltas) and a modulation spectrogram. Truong Pham et al. (2021) used log mel-spectrogram, deltas and double deltas. de Lope et al. (2021) used log mel-spectrogram as well as Mel Frequency Cepstral Coefficients (MFCCs). 3 data augmentation methods were here done in order to get more training samples; time warping (e.g. speeding up a sample), frequency mask (hide a part of the frequency spectrum) and time masking (mask out a time-section of a sample). Li et al. (2021) utilized a log mel-spectrogram. Krishnaiah and Divakarachari (2021) used spectral spread, spectral centroid, spectral skewness, MFCC and linear predictor coefficients. Cañón et al. (2021) used mel-spectrogram (not log). Farris et al. (2021) extracted some prosody-related features: zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral rolloff, MFCC, Chroma vector and deviation. Griffiths et al. (2021) made one predictive model for arousal and one for valence regression, where both only included features that achieved Spearman's correlation coefficient of above 0.85 with the respective dimension to predict. They selected energy, standard deviation energy and median energy for predicting arousal. For valence prediction they used spectral spread and spectral flatness. Finally, Griffiths et al. (2021) experimented with up to 529 features, including most of the features mentioned earlier in this section.

[7] https://www.tensorflow.org/datasets/catalog/gtzan
(visited Dec 3, 2021)

### 3.2.5 Machine Learning Methods and Findings

Both traditional machine learning approaches and advanced deep-learning architectures were used in the reviewed work. The vast majority of concepts used in this section are explained in chapter 2 or in this section. Yet, the most fundamental AI concepts and methods are assumed to be known to the reader. The reviewed work is grouped under the best describing category of the succeeding subsections. "Cross-Corpus Speech Emotion Recognition Based on sparse Subspace Transfer Learning" (n.d.) is not included here due to too low reproducibility, as previously explained; their utilized features were not mentioned in the paper. This was discovered a long time after finishing the SLR, so the dataset they used was still included above, since one can never know about too many datasets.

In SER there are two fundamentally different experimental settings: *speaker-dependent* and *speaker-independent*. Speaker-dependent emotion recognition denotes that each speaker's samples can be present in both training and test sets. Conversely, speaker-independent settings imply that the same speaker cannot be present in both datasets. The latter demands a more generalizable model in order to perform well.

#### Tensor Factorized Neural Network

Pandey et al. (2022) (SER) proposed two extensions to the tensor factorized neural network (TFNN): 3D attention gated TFNN and parallel attention gated TFNN. The former is a TFNN where the input is a 3D tensor composed of log-mel spectrogram, deltas and double deltas. The attention layer is used to give more importance to the parts of the input mel-spectrogram that are more relevant for the target label. The other architecture is a parallel network of mel-spectrogram on one side and modulation spectrogram on the other—both tensorized. Their cross-validated results yielded 85.15% accuracy for Emo-DB and 3D attention gated TFNN. With parallel AG-TFNN they achieved 55.56% on IEMOCAP. In both cases they reached approximately state-of-the-art results, using less parameters and less computational complexity than the baseline CNN+LSTM.

#### Traditional Machine Learning Methods and Ensembles

de Lope et al. (2021) (SER) tested kNN, SVM, random forest, multi-layer perceptron (a feed-forward neural net) and a CNN—on RAVDESS. For their main experiment, SVM with a polynomial kernel gave the best results: 71.2% accuracy (cross-validated), which is competitive with results reported for the same dataset. They also found a strong trend for accuracy improving when the number of emotion classes to classify decreased.

Krishnaiah and Divakarachari (2021) (MER) tested a multi-class SVM, i.e.

multiple binary SVM classifiers (each with linear kernel) that are combined into a majority vote ensemble, where the class with most votes is the final prediction output. They worked on the Indian datasets HMD and CMD, classifying ragas. Using their selected features, they reported a performance of 97.53% accuracy, which was better than the runner-up of the compared work which scored 95%. The results were not reported to be cross-validated.

Farris et al. (2021) (MER) tested a kNN, SVM (linear kernel), random forest, extra trees, gradient boosting and MLP (multi-layered perceptron). Extra trees and gradient boosting (e.g. *XGBoost*) are ensemble models that combines the output of many simpler learners, like decision trees, into one final prediction. Using their custom dataset of singing vocalists, they achieved accuracies like 49.1% (testing one single singer, SVM) and 43.8% (using 3 singers, and gradient boosting and random forest got equal score) on the 20 emotions taxonomy from the Geneva Wheel of Emotion. Moreover, for the 4 quadrants task and multiple singers (i.e. 3 singers) the accuracy was 68.8% (gradient boosting). Their feature selection experiment found well-maintained performance when using a small subset of the original feature set they had chosen. The results were not cross-validated and one single train/validation/test split was used.

Griffiths et al. (2021) (MER) developed one linear regression model for predicting arousal and one for predicting valence values. Using only 3 features per model lead to $R^2$ (coefficient of determination) scores of 0.85 and 0.78 for arousal and valence, respectively. These results were either the same or significantly better than previous work. It is important to underline that their dataset had only 20 songs for training and 40 for testing, as well as the compared work not using the same dataset.

#### Long Short Term Memory

Grekow (2021) (MER) used an LSTM (long short term memory) for training separate networks for arousal and valence prediction. Since an LSTM is used, the input audio sample (6 seconds) was segmented to 2-second pieces (no overlap proved to be best). One feature vector was extracted per segment and these were concatenated to form a vector of feature vectors, which were then input to the next step of the system. They used a pre-trained model as a *feature extractor* to derive a more intelligent and abstracted representation of the input vector—faster. They trained one simple feature extractor of one dense layer for each of arousal and valence. The feature extractor's output activations were then used to create new feature vectors (the feature extractor was not fine-tuned during RNN training), which were further given as input to the RNN network which consisted of one LSTM layer, which next was input to a dense layer. $R^2$ scores was reported of 0.73 and 0.46 for arousal and valence, respectively and cross-validated. These results can be seen in comparison to their strongest baseline model (regression

SVM) with $R^2$ of 0.48 and 0.27 for arousal and valence.

### Convolutional Neural Network

Li et al. (2021) (SER) combined DenseNet and DCNN (deep CNN) to form "Dense-DCNN". DenseNet (Huang et al., 2017) is a CNN where each layer is densely connected to each succeeding layer, compared to the normal CNN where one layer is only connected to the next. Note that as of today the DenseNet article has been cited over 20000 times. DenseNet's jump-connections between layers mitigates the problem of gradient disappearance of deep neural networks and strengthens feature propagation, which further reduces model parameters (Li et al., 2021). DenseNet allows CNNs to be deeper, more accurate and more efficient to train. Li et al.'s resulting combined model can, according to the authors, learn high dimensional features and yield accurate and fast classification. Another highlight is their adoption of StarGAN (Choi et al., 2018) to generate augmented samples, for any emotion class of their choice. They (seemingly) train *one* StarGAN model based on all the datasets in focus (SAVEE, Emo-DB, CASIA) and all their available emotion labels. The Dense-DCNN, with StarGAN-expanded dataset of 2000 samples per emotion, showed good cross-validated generalization ability. When using the augmented dataset the model demonstrated 97.36% weighted accuracy on RAVDESS, compared to 83% without augmenting the data. For SAVEE, Emo-DB and CASIA the proposed architecture achieved 92.97%, 91.06%, 92.86%, respectively and weighted accuracy. The performance on all the datasets (one Dense-DCNN model per set) was a few percent higher than the runner-up, across all of the datasets (weighted accuracy and cross-validated). All in all, Dense-DCNN showed great performance and robustness, as well as great performance in multiple noise environments, from experiments where they blend in noise with the audio samples. They also found that log mel-spectrogram gave better model performance on all datasets, compared to MFCC and raw audio-spectrogram.

Cañón et al. (2021) (MER) tested transfer learning between unlabeled speech data (e.g. read speech from audiobooks, not necessarily emotional) to MER (music with lyrics), in intra- and inter-linguistic settings (e.g. speech in English to English music). Their input feature was log mel-spectrogram, or specifically, partially overlapping segments of a sample represented by such spectrograms. They explored two methods for unsupervised representation learning (pre-training), referred to as encoders: *CPC* and *SCAE*, that both utilize convolution. CPC (Contrastive Predictive Coding) learns by trying to predict future sample segments (e.g. the next 2 seconds of a sample) in a latent space. It is an unsupervised approach to extract representations from high-dimensional data (Oord et al., 2018). SCAE (Sparse Convolutional Denoising Autoencoder) (Cañón et al., 2020) is a deep and advanced CNN architecture. They later implemented transfer

learning by using one of the pre-trained encoders (CPC or SCAE) as input to a feed-forward net (for MER) with densely connected layers (all nodes in one layer connected to all nodes in the next). Here they tested both freezing (i.e. *one-step transfer learning*) and not freezing the encoders' weights while training the rest of the classification network. The full classification system, for both versions, is a multi-output classifier, which predicts quadrant (out of 4 quadrants), valence (positive vs. negative) and arousal (positive vs. negative, i.e. high vs. low). Some findings reported are: pre-training on a mixture of speech languages improved showed a strong trend of giving better performance for MER, compared to training on the same language of speech as music, but it was not always this case. Their approach did not appear to learn emotion-related features from speech that were transferred to MER. The f-score for valence (high vs. low) was consistently better than for arousal. The SCAE models performed better than CPC, across all experiments. For the 4Q dataset, the best result for the 4-quadrants/classes problem was f-score 57%, which was the SCAE model of English speech to English music, with unfreezing the weights of the encoder. Unfreezing the weights allows the supervised training process to adjust the encoders further for the MER problem. Their results were not cross-validated, and a train/validation/test split was used.

#### Long Short Term Memory + Convolutional Neural Network

Truong Pham et al. (2021) (SER) experimented with a dilated convolutional recurrent neural network. Combining the data augmentation methods of time-shifting, pitch-shifting and WaveGAN—together with using the combined softmax and CT-C loss resulted in an accuracy of 91.90% on Emo-DB (cross-validated). CT-C (contrastive center) loss is a loss function that compares the network output vector of a positive example to an output of the same class and to a negative class. The loss is low if the cosine distance to the same class vector is low and the distance to the negative vector is high. CT-C loss is also proved to outperform *center loss* for deep learning (Qi & Su, 2017). The runner-up from the compared work (*ADRNN*) yielded 85.39%. Truong Pham et al.'s ablation study showed that the data augmentation techniques deserved most of the credit for their performance improvement.

#### Performance per Dataset

It is not possible to directly compare which domain of SER and MER has the best performance, especially because performance within each is highly influenced by which dataset was used. From the lower amount of data in MER, one could argue that deep learning models would perform worse due to lack of training data, assuming speech and music emotion is similarly complex to classify. However,

the best performance per dataset (based on the SLR) is concluded in table 3.2. Note that Griffiths et al. (2021) dataset was poorly represented in quadrant 2 and 4, which gives a misleadingly good and generalizable performance. The domain of ragas (HMD + CMD data) is suspected to be a much easier ML problem, as ragas usually include distinct sets of notes. In the table we can see strong performance across the board, especially for the SER domain.

Table 3.2: Best performance per dataset

| **Domain** | **Dataset** | **Used by more than one paper?** | **Article** | **Method** | **Score** | **Cross-validated?** |
|---|---|---|---|---|---|---|
| SER | Emo-DB | Yes | Throung Pam EA | ADCRNN | 91.90% | Yes |
| SER | IEMOCAP | No | Pandey EA | Parallel AG-TFNN | 55.56% | Yes |
| SER | RAVDESS | Yes | Li EA | Dense-DCNN + StarGAN | 97.36% weighted accuracy | Yes |
| SER | SAVEE | No | Li EA | Dense-DCNN + StarGAN | 92.97% weighted accuracy | Yes |
| SER | CASIA | No | Li EA | Dense-DCNN + StarGAN | 92.86% weighted accuracy | Yes |
| SER | eNTERFACE | No | Disqualified | | | |
| SER | RML | No | Disqualified | | | |
| MER | HMD + CMD | No | Krishnaiah and Divakarachari | SVM | 97.53% accuracy | No |
| MER | 4Q | No | Cañon EA | SCAE (CNN) | 57% f-score | No |

**Table 3.2 continued from previous page**

| Domain | Dataset | Used by more than one paper? | Article | Method | Score | Cross-validated? |
|---|---|---|---|---|---|---|
| MER | CH818 | No | Cañon EA | SCAE (CNN) | 52% f-score | No |
| MER | Farris | No | Farris EA | Gradient boosting | 68.8% accuracy for quadrant problem | No |
| MER | Griffiths | No | Griffiths EA | Linear regression | $R^2$ of 0.85 (arousal) and 0.78 (valence) | No |
| MER | GTZAN | No | Grekow | LSTM | $R^2$ of 0.73 (arousal) and 0.46 (valence) | Yes |

## 3.3 Transfer Learning between SER and MER

This section reviews some additional related work that explores transfer learning between the domains of speech and music emotion recognition (both ways usually). The motivation for including these here is to exemplify some attempts to generalize between the domains of speech and music, using machine learning—although these were actually all the articles that was found after extensive research. As mentioned, these additional articles did not go through the structured literature review process (which was done during the preparation project), mostly due to the low amount of such work. However, the quality of these articles were also taken into account before deciding to finally include them. These articles are the results of snowballing (exploring networks of article-references) from early influential articles in this niche topic, like Juslin and Laukka (2003). Exploring Google Scholar was also attempted.

None of the articles that attempted transfer learning between the domains mentioned any distinction between music containing or not containing vocals (instrumentals), nor the possible impact this can have for the degree of overlap between the emotional feature space. Large-scale transfer learning from speech to *instrumental* music has not been attempted before this thesis.

#### Transfer Learning Reveals Shared Acoustic Codes

Coutinho and Schuller (2017) compared intra-domain (e.g. train on speech and test on speech), and inter-domain emotion recognition between speech and music. For the inter-domain experiments, two strategies were evaluated: direct supervised transfer learning and feature-representation-transfer based on denoising autoencoders. For the former, the model is trained on a source domain and tested on the target domain directly (referred to as $CD$). The supervised problem they worked with was continuous emotion regression in the VA-plane (valence/arousal; 1 value per second). In relation to the latter, unsupervised representation learning is done through *denoising auto encoders* (Vincent et al., 2008), which learns higher representation of input features through trying to reconstruct corrupted input data (Coutinho & Schuller, 2017) (referred to as $CD_{TL}$). Here, the autoencoders are trained on unlabeled emotional speech *and* music data. The motivation for this is that unlabeled data is more accessible than labeled data, and the hope is to be able to learn a common representation of the input which limits the gap between the two domains. When carrying on to the supervised training (on one domain only), the difference for $CD_{TL}$ from $CD$ was to initialize the first layer with the learned weights and biases (kept fixed during training) from the denoising autoencoder, instead of initializing randomly. The authors did not mention *any* distinction between music with or without vocals (instrumental) and the possible dissimilarity this can have in relation to shared

codes. Analyzing the music data they used, at least 11% of the data—in terms of duration—was classical music or film music (I here assume that film music equals instrumental, but this is not necessarily the case).

The results showed equal performance for music prediction, whether the model was trained on speech or on music. In the case of speech emotion prediction, the models trained on speech (intra-domain models) outperformed those trained on music (Coutinho & Schuller, 2017). Interestingly, the knowledge transfer approaches ($CD$ and $CD_{TL}$) both lead to the statistically same results for both speech and music. This is confirming that there is strong overlap between the emotional representation in speech and music (music including vocals). The findings also shows that computer scientists can utilize cross-domain data to improve their model performance (Coutinho & Schuller, 2017). An earlier article, Coutinho et al. (2014), also found good cross-domain generalization performance from speech to music. Note that this article is also doing a time-continuous problem and utilizes a set of 8 *instrumental* (clarified through email correspondence) film soundtracks, of only about 15 minutes total duration. Their speech data had a total duration of 9 minutes and stemmed from 8 speakers in total (mostly enacted emotional speech).

**Transfer Learning and Emotional Commonalities between Speech, Music and Environmental Sound**

Weninger et al. (2013) used one dataset for spontaneous (natural) emotional speech, one for acted speech, one for music (including vocals) and one for ambient sound events. As to the music, at least 73% contained vocals. They executed feature relevance analysis and automatic regression (support vector regression) for continuous VA-ratings across all these domains—also cross-domain where trained on one domain and tested on another.

If focusing on speech and music, their results show higher arousal performance when trained on music and tested on speech (both for spontaneous and enacted cases) than when trained on speech and tested on music. As to the valence performance, the trend was the opposite. Still, considerable generalization capability was demonstrated in all of these cases, and the aforementioned trends also holds regardless of using the general feature set (a common one identified for all the domains, ambient sound incl.) or task-specific set (e.g. optimized for generalization between spontaneous speech and music only). If relating to all of the domains in their study, there was a high degree of correlation (performance) in all cross-domain experiments both with task-specific features and generic features. The generic features awarded less correlation, on average, than the task-specific. Regardless, the study suggests that it might be possible to identify a common code (features) for emotion signaling across all the domains of spontaneous- and enacted speech, music (with vocals) and ambient sound. Ad-

ditionally, cross-domain arousal and valence regression have been proven feasible across all the combinations of domains (Weninger et al., 2013).

# Chapter 4

# Method

This chapter deals with the choice of emotion taxonomy, establishment of datasets for speech emotion recognition (SER) and music emotion recognition (MER), choice of features and finally architectures—to be used for the experiments in this work.

## 4.1 Choice of Emotion Taxonomy

Research question (RQ) 1 asks what emotion taxonomy is best to use in this project (section 1.2). More specifically, it asks whether the dimensional or categorical paradigm is better than the other in terms of validity and reliability. The answer to RQ 1 will be guiding the *choice* and/or the *creation* of datasets for this thesis (RQ 2), and the chosen taxonomy must fulfill the constraint of being applicable to both SER and MER. If efforts are required to create a feasible MER dataset, then it also makes sense to choose to construct such dataset in accordance with the emotion model with the most potential applications.

Section 3.2.1 reveals the following about the usage of emotion taxonomies according to the structured literature review (SLR): Firstly, the work in the SER domain only used categorical emotion taxonomies. They differed in *which* emotions that were used, but common for all of them were: anger, happiness and sadness. None of the SER authors explained their selected set of discrete emotions here, except their selection perhaps being the only common emotions across multiple datasets, or the only ones present in one dataset. Burkhardt et al. (2005) share that using a categorical approach as opposed to a dimensional makes more sense for SER, especially in acted settings, as discrete emotions are more easily understood by the performer and the listener. On the other hand, other methods to find research than SLR, has identified various SER datasets that also

provide dimensional labels (examples are given later). For the MER articles in the SLR, 2 of 5 used dimensional models, with regression tasks (for valence and arousal of the Russell's circumplex; VA), one used quadrants in the dimensional VA-plane, one used Indian ragas and the final one tried the categorical Geneva Wheel of Emotion with 20 discrete emotions, as well as the 4 quadrants defined in that model for the valence-control plane. It is indeed possible to use both categorical and dimensional emotion models (taxonomies) for SER and for MER.

Why does it even matter to put thought into the choice of emotion model for automatic emotion recognition from affective sound—why not make a random choice of emotion paradigm and perhaps also the set of emotion-labels? Juslin and Laukka (2003) argue that researchers should pay close attention to this choice (in their case they referred to choice of a set of discrete emotions), and base it on theoretical grounds—and the results (especially recognition accuracy, and feature-commonalities between speech and music) can differ greatly. Juslin and Laukka, ironically, did not provide any arguments for why they chose to investigate a categorical emotion approach over a dimensional one. Still, the authors provide a wide range of evidence for the existence of *basic and universal discrete emotions*, and that each emotion type is like an algorithm, and is the result of evolution. One can easily agree to the existence of some basic emotion-families (e.g. Ekman, 2016), yet one aspect that a plain and discrete emotion paradigm misses out on—especially in the context of SER and MER—is the strength of each emotion. Especially if the strength is low, it can be difficult to discern what distinct emotion is actually being portrayed (Eerola & Vuoskoski, 2011). It is in fact rare that emotions are portrayed strongly and prototypically in the wild, at least for SER (Burkhardt et al., 2005; Lotfian & Busso, 2019). Additionally, the borders between instances and non-instances of one discrete emotion can be very fuzzy, and therefore the discrete model has lower resolution than a dimensional model (Cowen et al., 2019; Eerola & Vuoskoski, 2011; Russell, 2003). Moreover, emotion categories have been found to drive speech emotion recognition across two cultures—more than valence and arousal—but were not seen as discrete clusters and rather as gradient transitions (fuzzy and continuous) from one cluster to another (Cowen et al., 2019). When it comes to validity, one can say that a discrete model that includes the main emotion families—fear, enjoyment, anger, sadness, disgust (Ekman, 2016)—covers most of emotional experience, and in that way has high validity. On the other hand, the disability to classify ambiguous emotional examples decreases the validity significantly. This all serves as strong support for not working with plain discrete emotion labels. If strength was included as another label dimension, it would mitigate some of the issue, but still the discrete choice, of a perceiver, of an emotion category when the strength is low—or the emotion is ambiguous for other reasons—still remains a challenge.

It has been demonstrated that a dimensional model (VA-plane) provided higher inter-rater consistency than the discrete model, especially when concerning moderate emotional strength (Eerola & Vuoskoski, 2011; Gideon et al., 2021). A downside of the dimensional (VA-plane) model is that the dimensions might be abstract for non-experts, and some training is likely necessary (Aljanaki et al., 2017). In comparison, identifying discrete emotions is something we are already doing in daily life. Imaginably and by my experience, the amount of training needed for doing VA-annotation is negligible. In relation to validity, when discrete labels have been mapped on top of the VA-plane (samples labeled with both label types), it has been shown that the clusters formed by fear and anger have significant overlap, even for samples of high emotional strength (Eerola & Vuoskoski, 2011). This suggests and supports the notion that the ultimate dimensional model (utmost validity) requires additional dimensions (Collier, 2007; Russell, 1980; Russell & Mehrabian, 1977), also considering that humans easily can distinguish between anger and fear. Another argument for VA-ratings, specifically, is that it makes sense to have access to the predicted arousal value of songs, since the level of arousal of music can modulate mood and affect the performance of cognitive tasks (Nadon et al., 2021).

All in all, even though the dimensional VA-model has some deficiencies in terms of validity too, it is clear that the VA-model has the capacity of modeling much more nuanced emotional experiences, with higher resolution, and higher inter-rater consistency than the discrete model. When having a dataset with dimensional ratings it is also possible to transform the ratings to lower resolution bins, if an application favors this. Hence, Russell's dimensional VA-model is the preferred emotional taxonomy for this project. Remember that there exists some proposed models with more than two affective dimensions, like *dominance* (Mehrabian, 1995), but this will not be tested in this work, primarily for data availability concerns due to the lower popularity of this model within MER. A lower popularity in itself of this three-dimensional model should not, as shown, signal that two dimensions are to be favored over three.

## 4.2 Dataset

In order to guide the choice of datasets, some quality-variables for datasets in both fields are first proposed. In general, it is well-known that deep learning (DL) methods usually require a lot of data, so dataset size will naturally be important, in addition to quality-variables. For example, 1500 samples is not enough for training satisfactory DL models in SER (de Lope et al., 2021). Another aspect that applies to both domains is *inter-rater agreement*, which informs us of the accuracy (or validity) of the annotated labels (by comparing multiple annotators' labeling). In the dimensional setting, a low score across the whole dataset

suggests either unsuitable annotators for the job, an unreliable emotional taxonomy used, or the possibility that all samples in the dataset have low emotional strength (for VA-plane, samples belonging around origin can be more difficult to rate accurately, though easier than for low-strength categorical emotions). This last possibility is very unlikely for large datasets. A final aspect relevant to mention here, which applies to both domains, is data-file-encoding. This includes sample rate and encoding format, like WAV or MP3, but this aspect was ignored, since speech emotion has been proven to be reliably conveyed through frequency-limited and compressed telephone speech, and excellent classification results have been demonstrated with as low as 8kHz sampling rate (i.e. how many data points stored per second) (Wu et al., 2011). In comparison, default sampling rate for CD audio is 44.1kHz, so as long as a dataset is not using extremely low values, like close to 8kHz, it will probably not have much impact. Indeed this could still have *some* negative impact on classification, meanwhile the chosen variables below are probably much more impactful.

### 4.2.1 Speech Dataset

For SER, the suggestions for the key quality-aspects are naturalness, number of actors, sentence-purity and inter-rater agreement. There are differences between how emotion is conveyed in speech when being spontaneous (natural) compared to acted. Some physical emotional cues cannot be consciously mimicked (though the emotion may sound very authentic), and prototypical and strong emotions (as in some acted scenes) are very rare in everyday situations (Burkhardt et al., 2005). Results have shown that agreement for arousal perception was seriously higher for natural speech than acted speech ($r = 0.81$ vs. $r = 0.64$), and the opposite for valence-ratings ($r = 0.56$ vs. $r = 0.68$; Weninger et al., 2013). These results support the notion that emotions tend to be portrayed more prototypically in enacted settings, and that valence is more difficult to accurately perceive than arousal (e.g. Grekow, 2021; Griffiths et al., 2021). Moving on with quality-aspects, the *number of actors* should not be too low, as this would limit the generalizability of a model, by overfitting to the actors' unique voice characteristics. Likewise, it is also important to represent both genders well in the set of actors. The aspect of *sentence-purity* (my own term) relates to the degree of which the emotion label of an utterance can be given away solely based on the verbal content itself—to a machine learning model. To exemplify, let us imagine a dataset of 4 emotion classes, and 4 scenarios of how the dataset can be built up:

- In scenario (1), each emotion could be acted with one unique sentence each (even if the semantics of this sentence is emotionally neutral).

- In scenario (2), each emotion could be acted with a set of, say 5 sentences, and these 5 sentences could be unique for each emotion class, such that if one sentence is used for emotion $x$ it is not used for $y$.
- For scenario (3), the dataset has 20 sentences, but the usage of these are not uniformly distributed across the 4 emotion classes, such that some sentences have a significant higher chance of belonging to emotion $a$.
- For the last scenario (4), the dataset contains only one sentence, which is emotionally neutral, and is uttered for all emotion classes (but uttered in a relevant way to each emotion).

The scenarios are sorted in an ascending amount of sentence-purity (1. is least pure). For (1), the model could learn features that simply informed it of which *words* were in the sentence (instead of *prosody*; see definition in chapter 2) in order to make a good prediction of the emotion label.

Within available SER datasets, the following were considered: IEMOCAP (ca. 10k samples, Busso et al., 2008), MSP-improv (8438 samples; Busso et al., 2017), SEWA (2000 minutes and natural; Kossaifi et al., 2021) and MSP-podcast (73042 samples, 6810 minutes in version 1.8 and natural; Lotfian and Busso, 2019).[1] MSP-podcast (version 1.8) was chosen due to it being natural and having the most data. The samples originate from public podcasts (highest possible *sentence-purity*) and are the result of an extensive screening and preprocessing process, to ensure clean speech with low noise etc. The data is annotated with both discrete and static dimensional labels (valence, arousal and dominance). The dimensional annotations are on a Likert-scale from 1 to 7. Each instance received final annotations by 5 people at minimum, through Amazon Mechanical Turk. The Krippendorff's alpha (section 2.4) agreement for the whole dataset at time of publication was 0.426 for arousal and 0.459 for valence (Lotfian & Busso, 2019). Additionally, the VA-space was almost entirely covered, which implies a much more balanced dataset than earlier SER corpora. Finally, the male/female ratio in version 1.8 is 54.84/45.16. Note also that the database is being updated yearly, with a goal of reaching 400 hours.

The MSP-podcast data was preprocessed (in this work) by only including samples with duration 5 seconds or higher, even though the final MER samples are in the 10-15 sec range. The shorter the samples, the more padding will be needed, and their annotations are probably less accurate (see 'initial orientation time' in section 4.2.2), which is feared will negatively impact transfer learning performance. This lower threshold resulted in 37339 samples. This utterance-duration is assumed (more) okay, considering the high amount of data. At the same time, it is, again, acknowledged that such short samples may be linked to less

[1] https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html

| Quadrant | Number of samples |
|---|---|
| 1 | 14767 |
| 2 | 10315 |
| 3 | 7881 |
| 4 | 4376 |

Table 4.1: Label distribution for SER data

accurate annotations. The static VA-ratings were normalized to the range $[-1, 1]$, then translated to their respective quadrant. A 4Q (VA-quadrants) problem is used for the current experiments, in order to create a simpler classification problem. The label distribution is showcased in table 4.1, which shows a strong imbalance between the classes. The strong over-representation of Q1 has a high risk of creating a bias for this class in trained models—which was exactly what occurred, according to section 5.3. It was explored to down-sample the dataset to the minority class, as well as combinations with allowing samples with lengths also below 5 seconds. Yet, none of these two measures lead to improved validation-set performance when trained on SER and tested on MER (which is the key performance to measure here), and were therefore discarded. This was only tested with the less complex of the two models in this study (DCNN).

### 4.2.2 Instrumental Music Dataset

Moving over to the MER quality-criteria, the suggested ones are inter-rater reliability, *instrumentalness* and *emotional content variability*. In relation to the absence of vocals in the songs (instrumentalness), some criteria for this are needed to be defined for this work. A song is considered instrumental (enough) if any of three criteria matches:

- (1) no vocals (talking, singing, humming) at all
- (2) the vocals were barely audible in the sound-mix (imagine someone mumbling to you silently, in a noisy cafe environment)
- (3) the vocals were of only one phrase, lasted for max a couple of seconds and any emotional content perceived from the vocals were incongruent with the overall emotional content of the song.

Concerning us now with the emotional content variability MER criteria, this is relevant since songs can have sections of varying moods. It is therefore important

to try to ensure the variation is low for a chosen song-segment (i.e. clip in the final dataset), now that we are dealing with the static emotion recognition problem. The ratings in such a static problem are effectively the average emotion rating for a given segment (e.g. 15 seconds), or the whole song. These ratings can be the result of an annotator listening to the segment and intuitively giving an average rating, or one can start with continuous ratings (e.g. for every second of the track), then average those. As a final note, playing an instrument or singing can both be done with varying degree of naturalness (spontaneity)—for instance through playing the instrument on the go vs. looking at a note-sheet. The degree of naturalness in the MER data will not be treated in this work, partly due to time and partly due to low chances of finding or creating the necessary data.

Across all MER datasets that could be found, there were not enough instrumental music datasets. It was desired to acquire enough samples so that it was feasible to conduct experiments where music data was used both for training and testing, in order to have something to compare cross-domain results with—i.e. not solely train on SER and test on MER. Some of the datasets that were disqualified and the reasons are shown in table 4.2. The high-level MER dataset strategy was to *compile* a high-quality collection/subset (which is a part of the *final dataset*) of instrumental samples with static VA-ratings, as well as emotional variation within a defined threshold (wherever possible). This data would then be applicable to regression problems, binned VA-ratings classification and the popular quadrant problem (dividing the VA-plane into 4 quadrants). The original datasets from where this VA-rating-level data was derived/compiled are: *Soundtracks* (Eerola & Vuoskoski, 2011) and *DEAM MediaEval* (Aljanaki et al., 2017). In addition to these, the piano dataset EMOPIA, which uses quadrant labels, was also acquired due to its large size—enabling even more training data for the applications where quadrant-level labels are sufficient (e.g. the experiments in this work). The compiled dataset and chosen EMOPIA segments are available on Github.[2] Various preprocessing needed to be applied across the original datasets. This process and descriptive statistics of the final data are documented below.

[2] https://github.com/jonasrodningen/instrumentalMER-data-thesis

| Dataset | Reference | Reason for drop |
|---|---|---|
| CCMED/ WCMED (instrumental) | Fan et al., 2020 | The ratings are derived from rankings, therefore the VA-ratings will not be accurate. |
| NTWICM | footnote[3] | Only 4 raters, and static rating is on song-level. |
| AMG1608 | footnote[4] | Static rating is on song-level. |
| Panda 4Q (SMC2021) | Panda et al., 2020 | Annotation only on VA quadrant granularity. Alternatively, the mood terms could be converted to VA-values, but this would not secure ratings of satisfying precision. If looking to do quadrant-classification problem: only 118 tracks that Spotify indicated to be instrumentals (these would need to be manually checked). Not worth the investment. |
| Grekow 21 | Grekow, 2021 | Only 6-seconds length. Out of the 324 total samples the amount of instrumentals would be low. |

Table 4.2: Some dropped MER dataset candidates

[3] https://github.com/juansgomez87/datasets_emotion#now-thats-what-i-call-music
[4] https://amg1608.blogspot.com/

### Soundtracks

The Soundtracks dataset consists of two sets of tracks, with a non-empty intersection. Some tracks were represented by multiple segments/clips/samples in the dataset. There were also some clips that were duplicates within the sets (originated from the same time-segment of the track). The longest segment of the original samples were about 20 seconds. In terms of the intra-sample variation in the original dataset, music experts had been tasked with looking for segments of 10-30 seconds that represented predefined target emotions within discrete emotions and bins of dimensional VA values (bins like strong-positive and medium-positive valence; Eerola and Vuoskoski, 2011). The annotations were first given by 12 music experts, then later re-tested with 116 university students. The dimensional ratings in the original dataset are in the range $[1, 9]$ and are static.

By extracting all the unique clips across Soundtracks set 1 and set 2, the resulting set contained 365 samples (segments), which were further clipped to the maximum duration possible between 10-15 seconds. These extracted segments were always selected to be the middle 10-15 seconds of each clip, in an attempt to filter out possible transitions that could exist in the tails of the clips (to minimize variability). Note that the annotations were already static, so there was no efficient way to identify the actual variations within clips.

The metadata of the original dataset only contained the album name (often the movie title) and track number, but not the artist and song title. This needed to be acquired so that duplicate-checking could later be done across all the preprocessed datasets—which is crucial such that the same song or clip does not appear in both training and test sets. It was first attempted to use an unofficial Shazam API (ShazamIO), without success. Several song-recognition solutions were explored, and ACRCloud[5] was favored. Most songs were identified through their API, and the rest were manually located by searching through Filmmusicsite[6] and Musicbrainz[7]. All matches on these sites were further double-checked by comparing the audio given by a song and artist Spotify search vs. the local sample in the dataset.

### DEAM

The original dataset (consisting of *Mediaeval* datasets years 2013; 2014; 2015) has 1744 45-second clips and 58 full length songs. Each artist were allowed maximum 5 songs in the dataset, and the 45-second clips had been extracted uniformly at random (allows high variability in emotion) (Aljanaki et al., 2017).

[5] https://acrcloud.com/
[6] https://filmmusicsite.com
[7] https://musicbrainz.org

Amazon MTurk[8] was used for annotation, where each song was annotated by either 5 or 10+ workers—if only 5 workers, it was the top performing workers, as determined by earlier tests. Each annotator's ratings were normalized in a smart way to account for subjective trend differences across workers. The labels were dimensional VA in the range $[-1, 1]$, and given with a frequency of 2Hz across the sample. The first 15 seconds of annotations for each clip was cut away, due to "initial orientation time"—based on cited findings that showed that participants required averagely 8.31 seconds of listening to initiate reliable continuous emotional judgments on a 2D plane, for music (Aljanaki et al., 2017). The continuous annotations' reliability in terms of Cronbach's alpha (section 2.4) for arousal (average across all the songs) was $0.28 \pm 0.28$, $0.31 \pm 0.30$ and $0.66 \pm 0.26$ for 2013, 2014, 2015 datasets, respectively. This alpha for valence was $0.28 \pm 0.29$, $0.20 \pm 0.24$ and $0.51 \pm 0.35$. The consistency for the 2015 dataset was significantly higher than the two others (Aljanaki et al., 2017).

Arguably, the low Cronbach scores are problematic (i.e. too low agreement between annotators), perhaps except for the 2015 dataset. However, when extracting 10-15 second segments (more about this soon) and transforming their continuous ratings into static ones, it is assumed that the reliability of these resulting ratings improves drastically and at least enough for binned classification tasks (e.g. granularity of 0.1 or worse). Still, these annotations are probably not reliable enough for high-precision regression tasks—which will not be within the scope of this thesis. The original dataset also contains static ratings for each song, but using these directly would not allow high-accuracy static ratings for custom extracted segments (of low emotion variation)—which is why static ratings later in this section are derived as an average of continuous ratings.

The procedure to identify which of the 1802 original songs were actually within my instrumentalness criteria (section 4.2.2), the Spotify API[9] was utilized to find the Spotify ID of each song based on a query formed by artist and title (included in dataset). Spotify found 961 correct matches. Spotify's instrumentalness score (originally produced by an ML model) was used to filter out all songs with a score below 0.5 (these were completely discarded), which is the limit Spotify intended to represent instrumental-only songs, while the confidence of prediction rises with higher scores. The need was identified to manually inspect all the 283 songs with Spotify instrumentalness score above 0.5 and whether they passed my own instrumentalness criteria, through listening to the *local* audio files. This was due to the presence of too much vocals in some samples with scores as high as 0.92. Additionally, when comparing the lower-cased tokens within {artist +

[8] https://www.mturk.com/

[9] https://developer.spotify.com/documentation/web-api/reference/#/operations/search
(visited Mar 3, 2022)

title} strings of the local database, and the artist + title fields of the returned best match from Spotify, 56 songs had less than 70% of the same tokens (while they should be nearly 100%). To summarize, the Spotify API was used to filter out all candidate songs (those who had a Spotify match) from the original DEAM dataset with worse than 0.5 instrumentalness score. Moreover, Spotify's scoring cannot be fully trusted on its own in this context, possibly due to Spotify trying to give a score that summarizes the instrumentalness of a *full* song. Anyhow this measure was a helpful indicator in the process of finding a total of 243 instrumental songs in the DEAM dataset.

It was chosen to identify segments within each song of 10-15 seconds length (the longest possible segment was prioritized, with 15 sec as max). The 10 seconds lower boundary was to be congruent with the identified initial orientation time, as mentioned earlier. The 15 second upper boundary is a result of needing each sample to not be so long that the emotional variation would always be too high. Looking at how the segmentation was done, a windowing search algorithm across the continuous valence and arousal ratings was developed. Various metrics were explored, with the goal of identifying all feasible subsets of each song, of the predefined lengths of 10-15 seconds, which had emotional variations within a certain max-boundary. The testing was done by manually listening to segments suggested as 'safe' by the algorithm and looking for those where the audio still revealed serious intra-segment variations in either arousal or valence. The metrics explored (comparing between search windows) were standard deviation change, mean rating value change as ratio and mean rating value absolute change, where the latter worked best. Various mean rating absolute change (across two windows) limit-values, window lengths and overlap-amounts were tested; window sizes down to 3 seconds and overlaps from 0 to 75%. The chosen parameters were window size of 5 seconds (to accommodate songs with only 30 seconds of annotations), no overlap between windows and 0.05 as the max-boundary for mean rating absolute change. This max-boundary was chosen to be overly strict, as values up to 0.15 also produced results that was still acceptable. The songs that were tested during the algorithm development were first 5 random songs that were found to contain lots of variation, and later tests during this process were done through manual confirmation of windows the algorithm had flagged as being too different emotionally. The results were that flagged windows that in the end were tested were rather too strict (had an acceptable VA-variability) than not, which adds confidence to the developed algorithm's performance. Still, a more thorough and systematic testing process could be beneficial, for increased confidence, but this was not prioritized, especially due to the lack of high-granularity intra-sample variation ensurance of the other original datasets in this thesis (EMOPIA and Soundtracks; more on EMOPIA later).

When achieving sufficient confidence about the ability to identify quick change

in emotion (between two five-second windows), the algorithm identified candidate segments consisting of 2-3 subsequent windows (prioritizing three windows) that had no problematic change from one window to the next (across both valence and arousal). These candidate segments were then analyzed to filter out segments were the delta between the mean rating of the first window and the last window was beyond the max-value (0.05). This was to target segments were a more progressive change occurred. If a 3-window long segment failed this test, a feasible new segment consisting of two of the windows was picked instead. The result of all this is that no segment can have more than 0.05 in delta of the windows' mean values across all the 2-3 windows in an extracted segment. This implies that the finest granularity for any machine learning problem on the final data should be 0.05 for valence and arousal values, on VA-scales within $[-1, 1]$. Additionally, in some cases segments were also allowed to be between 10-15 seconds when the segment included the end of the original clip. A final constraint was set so each original clip (45s to full-length songs) contributed max 5 segments to the final dataset. The final static VA values for each segment was set to be the average of the continuous ratings for each. The DEAM part of this thesis' final dataset ended up containing 377 audio clips (i.e. segments).

### EMOPIA

The EMOPIA set is annotated with the 4 VA quadrants, the average segment length is 40s, and it originally consists of 1087 segments from 387 piano-songs (Hung et al., 2021). The segments were carefully selected by the four authors to be emotionally consistent according to subjective perception (based on a higher granularity than just quadrants). The segments were also extracted only at "cadential" arrivals (i.e. respecting musical/melodic phrases) (Hung et al., 2021). By personal inspection, some of the original segments in the dataset contained a short intro, and it was therefore decided to extract the middle 10-15 seconds of each original clip (15 seconds if possible). This duration was also so it matches the max length of the Soundtracks and DEAM data. Each original song was further constrained to have maximum 5 segments in the final dataset. The result was 868 clips between 10-15s.

### The Final MER Dataset

Soundtracks and DEAM data were normalized into VA values within $[-1, 1]$. It was necessary to ensure that no duplicate songs were used across the final Soundtracks, DEAM and EMOPIA dataset-parts. To do this, the artist field of each song was lower-cased and word-tokenized with NLTK (i.e. the string is split into separate words). Each song is compared to every other song (cross-dataset), and if 66% of the artist-tokens were found in the compared song's, the two songs

were added to a list for manual inspection of titles, and audio if needed. In the EMOPIA case, only Youtube video-titles were available, which usually contained song artist, title and other information. This comparison process identified some songs with the same artist, but all had different song titles (so no issues). The final MER dataset's quadrant label distribution is shown in table 4.3. We see that unlike SER, this data is well-balanced.

Since multiple audio-segments could originate from the same song, it was crucial to ensure that train and test folds of this data did not both have segments of the same song in them. After this, since quadrant labels was going to be the label type used in this thesis, it was meaningful to stratify the folds such that the relative amount of each class was as balanced as possible. It was not found any such programming implementations already, so a novel one was made, even though, unfortunately, after creating this algorithm a similar existing solution was discovered in Sklearn.[10] Regardless, a short high-level description of the novel algorithm follows. It first utilizes the Sklearn's StratifiedShuffleSplit on the song-level (of the combined dataset), to ensure that no song is in both training and test simultaneously. Then, in a loop it swaps from each of train and test one random song (and all its segments) of the most imbalanced class at that point of time with the most imbalanced class in the other partition. It continues until the mean squared error of the relative class-distributions of train and test is below the value of 0.01. Finally, the algorithm ensures that the ratio of segments in test, compared to train, is within $[19.75\%, 20.25\%]$, by swapping one random song at a time from each class over to the other fold.

| **Quadrant** | **Number of samples** |
|---|---|
| 1 | 437 |
| 2 | 384 |
| 3 | 385 |
| 4 | 404 |

Table 4.3: Label distribution for MER data

[10] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedGroupKFold.html
(visited Mar 3, 2022)

## 4.3 Features

As seen in the SLR, an enormous variety of features can be extracted in both SER and MER domains, and might facilitate good prediction. The low-level features; log mel-spectrogram, stacked together with its delta and double deltas (see definitions in chapter 2), were used successfully in Truong Pham et al. (2021), and contributed to impressive performance of 91.90% accuracy for 7 discrete emotions for SER (see section 3.2.5). One benefit of the mel-spectrogram and spectrograms in general is their abundance of information, due to being low-level features. In essence they can contain nearly all the information stored in the original audio waveform (audio file). These 3 features (referred to collectively as *3D log mel-spectrogram*) are also used in this work with the assumption that the same features and model-architecture also will work for the dimensional VA-plane and its quadrants. Like in general in this thesis, if something is directly reused from other work, it will be specified.

### Sample Loading and Padding

All the features were extracted using the Python *Librosa* library (McFee et al., 2021) (more on the process below). All files were loaded and down-sampled to a sample rate of 16kHz (yields 8kHz as maximum sound-frequency). The lower sample rate, the narrower the frequency spectrum's range will be, and though humans can hear up to around 20kHz, 11.025kHz is argued to be high enough for most use cases.[11] Nevertheless, 16kHz is still chosen, due to the impressive demonstration by Truong Pham et al.. The loaded samples were zero-padded symmetrically on the sides (centered, along time-axis) such that all samples' features were 15 seconds. One extra column was added on the right side when needed. Two other alternatives to center-padding could have been to use trailing padding or positioning the sample at a random location of the 15-second array. Since the average length of SER samples was much shorter than 10-15 seconds, trailing padding would probably lead to a biased model towards the start of the feature, when trained on short SER data. This would lead to the SER-model overlooking important parts of the MER data for the prediction on MER. Random padding is not tested in this work, but could lead to even less bias, as centering likely introduces some bias towards the center.

### Feature-extraction

The mel-spectrogram parameters used were: min-frequency of 300Hz, max-frequency of 8kHz, number of mel-filters was 40 (in short, the filters summarizes the frequency-

[11] https://librosa.org/blog/2019/07/17/resample-on-load/ (visited Mar 30, 2022)

dimension into 40 buckets), window-length was 400 samples (25ms), the hop-length was 240 samples (yields 10ms overlap between windows) and power was set to 1 (which gives us an energy spectrogram and not power spectrogram). There are also other parameters available for Librosa's mel-spectrogram, but these were kept at default. Especially the window-length and number of mel-filters are determining the resolution/dimensions of the image-array (time vs. frequency dimension, respectively). The reason for cutting off frequencies, and lowering the mel-filters and window-lengths, is to reduce the complexity of the feature space and save computation. The lower cutoff frequency of 300Hz is a bit too strict and cuts off much valuable emotional information, according to some random listening checks. Seemingly, 150Hz would have been a better trade-off. 8kHz seemed like an okay upper cutoff. Additionally, 40 mel-filters is a low number of filters (low resolution of frequencies). Despite all of this, these values from Truong Pham et al. (2021) are still used, as maximizing performance is not the main focus of this thesis. Training models with *power* mel-spectrogram (chapter 2) was tested, but decibel-scale amplitude gave up to 20% performance boost when training and testing on MER (validation data). Conversion from decibel was done with Librosa's *amplitude_2_db* function (since mel-spectrogram was with power=1), and the parameter *ref* was set to 1 such that all data-samples were calculated with the same scale (using each sample's max amplitude would be another option, which would normalize locally). Note that the parameter *top_db* for this function did not have any effect on the returned values, perhaps due to a bug in the library. Further, the log mel-spectrogram was used to calculate the deltas and double deltas. The 3D log mel-spectrogram was then constructed by stacking the log mel-spectrogram, deltas and double deltas arrays along the last axis, i.e. the channels axis (axis 2), such that the result imitates how RGB pixel values are stored in an image array. This 3D feature can then easily be processed by a convolutional neural network (CNN) network. However, the extracted 3D feature was rotated to facilitate easier reshaping of the CNN layers' outputs, for potential succeeding recurrent neural network (RNN) layers in a combined architecture (the architectures are described later in this chapter). The rotation was done by a transpose of axis 0 and 1, which effectively flips the 0-axis, then rotates the 3D feature 90 degrees clockwise, while preserving axis 2. The flipping is required since Librosa stores the extracted mel-spectrogram's y-axis flipped, with the origin in the bottom-left, while images are usually stored with origin in top-left.

#### Standardization of Amplitude

An important decision for the feature-processing was whether to normalize the amplitudes or not. Since the mean of the log mel-spectrograms was significantly greater than the deltas and double deltas, and to avoid that neural networks

wrongly overemphasized one feature over the others, it was decided to normalize. This would then facilitate better learning. The next decision was about doing local (i.e. get normalization params per sample) or global (base normalization params on all samples). Global was preferred, since local would hide the amplitude differences across the samples, which has been shown to be one of the top features for arousal prediction and a significant feature for valence prediction (see e.g. *RMS* in Abri et al., 2021). Standardization (zero-mean and unit standard deviation) was chosen over max-min normalization, since the latter proved to be a bit too sensitive to the outliers in this case, while standardization is, as usual, more robust to outliers. Moreover, this method achieves values where most of the data will be in the range of $[-1, 1]$, which usually is preferred for neural networks.[12] Standardization has been shown to yield the best performance when comparing normalization methods for SER, although the tests were on other features than mel-spectrogram (Böck et al., 2017; Sefara, 2019). The standardization was done per feature-type and the parameters for each were retrieved by analyzing features that had no padding applied. More specifically, each loaded sample was used to calculate the mel-spectrogram, without first applying padding, then the mel-spectrogram was converted to decibel, and the delta-types were further based on this. Next, each feature was flattened to a vector, and concatenated to one global vector per feature-type, that stored all the global values, which lastly was used to calculate the mean and standard deviations to be used for standardization.

## 4.4 Architecture

Two architectures are explored, both implemented with Keras (Tensorflow v.2.5). The first is a dilated CNN (DCNN) architecture, with skip-connection (most methods are explained in chapter 2), which is a simplified model of the attention dilated CNN RNN (ADCRNN) from Truong Pham et al. (2021) (which again is based on Meng et al., 2019)—as a reminder, Truong Pham et al. (2021) was one of the articles from the literature review in chapter 3. The second architecture in this thesis is the full ADCRNN from Truong Pham et al. (2021), which adds attention and dilated LSTM (RNN) to the DCNN. The ADCRNN in this thesis uses a softmax on the final layer instead of a combined softmax and *center-loss* function, and has some added batch-normalization. The combined loss function was skipped due to time-limits, and that it only gave some minor improvements in the reviewed paper. Batch-normalization was beneficial during validation-runs, likely due to the high complexity of the network. The reason for doing a sim-

[12] https://stats.stackexchange.com/questions/421927/neural-networks-input-data-normalization-and-centering/422087#422087
(visited Apr 22, 2022)

plified model first (DCNN) was to at minimum get to test a proof-of-concept, in case there was not going to be enough time to implement the complex ADCRNN-model. Some tuning efforts of some parameters is done with the validation data (no cross-validation, as it was assumed a single validation set was generalizable enough), in a manual evolutionary algorithm fashion, with the goal of improving the performance and reducing overfitting, to some degree. Repeatedly, maximizing overall performance is not a necessity in this thesis, and more thorough tuning and exploration of even more architectural variations are recommended.

### 4.4.1 Dilated CNN (DCNN)

The DCNN is illustrated in fig. 4.1. Note that especially the number of units/filters has been tweaked, and regularization is added, compared to what is used by Truong Pham et al. (2021), to reduce overfitting and improve performance. The 5 last layers in this architecture are also different from Truong Pham et al., primarily since the output features of the final convolutional layer needed to be flattened and some extra learning on these features is useful. Other than this, the parameters are taken from Truong Pham et al.. In the network, all layers that are feasible use l1(0.001) as kernel regularization, constant(0.1) as bias-initializer and relu for activation (except the last layer which uses softmax). All the convolutional layers has kernel size 3. The first convolution layer has 128 filters, the padding is set to "valid", and dilation-rate to 1 (no dilation). The max-pooling layer has a pool-size of (4,2), strides (4,2) and "valid" padding. The pool-size and strides for the previous was also attempted with (2,4) and (4,4) during tuning, while the chosen values were slightly better. For all the dilated convolution-layers, we have 64 filters, the dilation-rate is 2, and padding is "same". The dropout is set to 0.3. The model's optimizer is "Adam", with learning-rate 0.0001 and the loss is "categorical cross-entropy". The compiled model had roughly 39m trainable parameters.

### 4.4.2 Attention Dilated CNN RNN (ADCRNN)

This network architecture is shown in fig. 4.2, and reuses the "DCNN Block". The only hyper-parameters taken from Truong Pham et al. (2021) are (except those mentioned in DCNN) bias-initializers, dropout-rate and dilation rates for dilated LSTM. Batch-normalization layers are new. As described in section 4.3, the extracted mel-spectrogram features were rotated and flipped by doing a transpose during pre-processing. This was to speed up training, such that correct transformation can easily be achieved through a simple reshape layer. A Keras "Permute" layer could have been used as part of the architecture instead of doing it during pre-processing, but this would be slower. The result of reshaping the CNN-output is that each channel/feature-image gets stacked on top of each

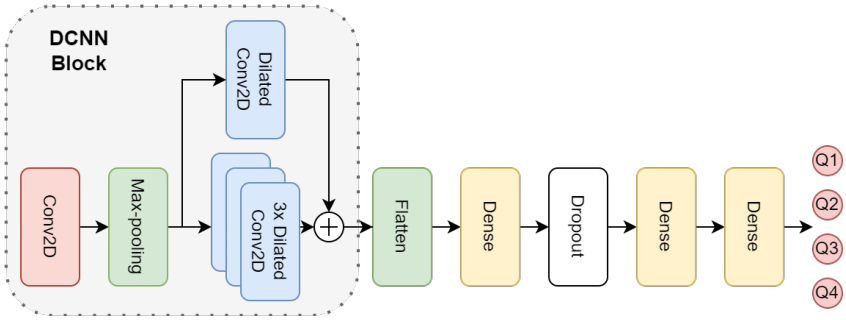



Figure 4.1: Dilated CNN (DCNN) architecture. Conv2D refers to 2D convolutional layer.

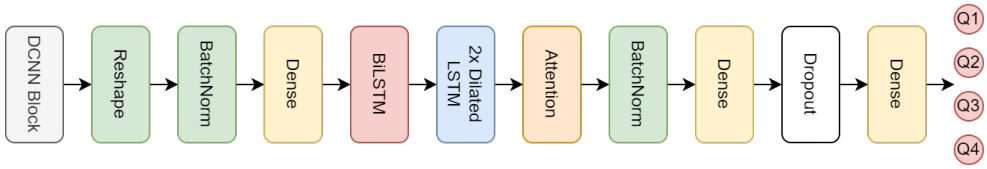



Figure 4.2: Attention Dilated CNN RNN (ADCRNN). It reuses the DCNN block from fig. 4.1.

other, and grouped per time-step which is further input to the RNN. If no transpose was done before reshaping, the feature-vector per time-step would contain information from multiple of the original time-steps. It is suspected that this detail may easily be overlooked in several attempts to create a CRNN in other research, especially for beginners like myself, which means that the potential of the RNN-layers cannot be fully utilized.

The dilated LSTM layer is a custom implementation created for this project. Effectively it changes the LSTM such that each time-step receives input only from $dilation - rate$ time-steps before it (and the current time-step itself), instead of getting input from the previous one. The implementation is based on the logic provided by the repository of Chang et al. (2017),[13] and it has been ported and optimized for Tensorflow version 2.x (including 2.5.0 which had many bugs that made graph compilation stricter). The layer is implemented as a Keras layer, and is also compatible with Tensorflow's graph-compilation (which gives very good execution-speed improvements). The creation and debugging of this layer required a couple of weeks, but it now successfully gives the same output as

[13] https://github.com/code-terminator/DilatedRNN (visited May 10, 2022)

the original implementation. Even though the ADCRNN architecture is inspired by Truong Pham et al., some arguments for the dilated RNN part is included regardless: LSTMs can assign too much weight on the early time-steps, especially for long sequences. By including bidirectional LSTM and RNN-dilation in the architecture, one solves that problem (Schoene et al., 2020). By using a backward layer in addition to a forward LSTM, on can capture surrounding information at each time-step instead of information solely from the previous one. The dilation part helps retaining long-term connections in the data.

Since the label of a sample, in this context, is the average 'emotion-value' of the sample, and both speech and music is dynamic in emotional content, a simple attention layer is useful to 'summarize' the hidden state of each time-sequence into one global feature-vector, created by weighting the importance of each time-step to the final label of the clip. If the dilated LSTM layer did not output sequences (and did not have a succeeding attention-layer), then its output would only be the hidden state of the last time-step of the LSTM (which is not a weighted sum of the time-steps). The attention mechanism is the same as the approach used in "ACRNN" in Chen et al. (2018), and implemented in Keras for this project. The mechanism is a trainable *self-attention* style attention, and since we deal with a single output classification problem, it is a many-to-one setting (as opposed to sequence-to-sequence). The attention layer outputs a context-vector $c$ of length $LSTM\text{-}output\text{-}units$, which is defined in eq. (4.1), where $\alpha_t$ is the output of eq. (4.2) and $h_t$ is the hidden state output of the previous layer's $t$-th time-step. In eq. (4.2) we have a softmax-function which normalizes the exponent of dot-product between weight-matrix $W$ and $h_t$.

$$c = \sum_{t=1}^{T} \alpha_t h_t. \tag{4.1}$$

$$\alpha_t = \frac{\exp(W \cdot h_t)}{\sum_{t=1}^{T} \exp(W \cdot h_t)} \tag{4.2}$$

Moving on to the important parameter per layer, from left layer to right (what is not mentioned is kept as default): The DCNN Block is similar to the DCNN architecture, except using an l1-regularization of 0.01 instead of 0.001 (for all layers that apply). Reshaping is done by permuting the first two axis (batch-size not included). The first dense has 256 units (no activation was used for this, mistakenly, though relu was intended). BiLSTM has 256 units (returns sequences). There are two dilated LSTM layers (each with 256 units), with dilation rates of 1 (effectively a normal LSTM) and 2, respectively. This layer also returns sequences. The attention layer has no hyper-parameters. The next dense is 64 units, with relu-activation. Dropout layer has a ratio of 0.3, and the final dense of course has 4 units (softmax activation). All the applicable layers

use l1-regularization of 0.01 and constant bias-initializer of 0.1. The model's optimizer is similar to DCNN, except learning-rate of 0.00001. The compiled model had roughly 3m trainable parameters.

# Chapter 5

# Experiments and Results

The current chapter explains the experimental plan, the experimental setup and presents the results. The results are discussed in chapter 6.

## 5.1 Experimental Plan

### Performance Measure

The performance measure for the experiments is accuracy, since this is the predominant metric in the reviewed classification work. For example macro-averaged f1-score could be used in addition, but this is not too useful since the music emotion recognition (MER) test-set is balanced (as shown in section 4.2). For those unfamiliar, macro-averaging means to take the arithmetic mean of each class' f1-score—where poor performance of one class will drag the total score down, which is useful for imbalanced data... However, the confusion matrices will be included, in order to clearly illuminate any potential bias resulting from the overly represented Q1-class of the speech emotion recognition (SER) data.

### Experiment 0: Baselines

The main baseline is a 4-class—as the current ML problem is to distinguish between 4 valence-arousal quadrants—random classifier with probability weights that reflect the label distribution of the MER training-set (which is highly balanced, but not perfectly balanced). The baseline is evaluated on the MER test-set. A higher accuracy is expected when using weights for each class, in contrast to 25% probability for each class. This experiment is designed to give us a reference point for comparison with the other ML models. For example, if we take that

the accuracy of a trained model was 80%, it might not sound equally impressive if we know that a random-choice baseline managed 75% accuracy.

A second baseline is included: SER to SER (SER2SER). This model is validated with the SER *validation*-set. This is useful because one gets a reference score that can be compared with the performance of the transfer learning experiments. The reason for not creating a separate SER *test*-set in this work is to maximize the data available for training in the main experiments. There is one such validation-result per architecture.

### Experiment 1: DCNN Architecture

The first experiment uses the simplified dilated CNN model. The experiment is divided into two parts:

- (1a) SER to MER (train on SER and test on MER; SER2MER)
- (1b) MER to MER (train on MER and test on MER; MER2MER)

Performance of SER2MER will then be compared to MER2MER. If the emotional representation (coding) of the SER and MER datasets have perfect overlap, then SER2MER is expected to perform better than MER2MER, since it has been trained on significantly more data.

### Experiment 2: ADCRNN Architecture

This round is for the complex ADCRNN architecture, with the intention of capturing more useful time-dependent relationships in the input-data. Similarly, it has two parts (2a) and (2b) analogous with experiment 1.

## 5.2 Experimental Setup

Some more information about the experimental setup for this work is included, to facilitate reproducibility. The MER dataset is split into a training (68%), validation (12%) and test-set (20%), with stratification and keeping all clips belonging to the same song in the same partition. These sets were of 1045, 240 and 325 clips, respectively. The SER dataset was split with stratification into training-set (99.4%; 37114 clips) and validation-set (0.6%; 225 clips). The reason for the low validation size here was to maximize data for training and have the size be comparable to the MER validation-set. Note that no control was added as to whether the same speaker could be present in both subsets, as the size of the dataset is very large. For convenience, the label distributions are repeated here, for the datasets before splitting into partitions for the experiment; the MER-set

had 437, 384, 385 and 404, for Q1 to Q4. The SER-dataset had 14767, 10315, 7881, 4376 samples, respectively.

The batch-size of 32 is taken from Truong Pham et al. (2021). The models are trained for 5 runs (test results are averaged, to limit effect of weight-initialization) and for up to 200 epochs for each run. Model training uses two strategies based on validation-accuracy: early-stopping and checkpointing. The latter of these saves the best performing model seen during training, while the former stops the training when no improvement of validation-accuracy is seen during 30 consecutive epochs. There will thus always be two model-versions after each run of training; one from each strategy. Both of these models (e.g. SER-early-stopping and SER-checkpoint) are later evaluated on their respective validation-set (e.g. SER validation-set). The best performing of these (one for SER and one for MER) is saved for testing on the MER test-set. During tuning—i.e. before the final runs on the test-set—the best SER-model was also evaluated on MER validation-set. Slight transfer-learning improvements could perhaps have been gained if the best SER model version out of checkpoint and early-stopping was determined by the performance on MER validation-set, instead of SER validation-set, but this was skipped.

The models were trained using Keras (Tensorflow 2.5.0) on an A100 GPU with 80GB VRAM, on a high-performance cluster (Själander et al., 2019) (which currently only offered this tensorflow version). The data is served to the models through *Tensorflow Dataset* (batched with the defined batch-size). Total training time for all runs was 2 hours for the DCNN (experiment 1) and peak VRAM 5.1GB. For experiment 2 (ADCRNN), the total training time was 27 hours and peak VRAM 4.6GB.

## 5.3 Experimental Results

### Experiment 1: DCNN Architecture

The results are included in table 5.1. We will look at one experiment at a time, starting with experiment 1 (DCNN) for now. In that, we see that SER2MER performs substantially better than the random baseline—meaning that some feature transfer is taking place—and performs about 10 base-points below MER2MER. Moreover, both models failed to generalize the high performance demonstrated on the MER validation-set (meaning that overfitting has happened).

In the confusion matrices in fig. 5.1 (showing the best test-run per experiment), one can observe a strong bias for Q1 in SER2MER (which is expected because of SER-data imbalance). MER2MER seems to have developed a bias for Q2, which is surprising since this is the class with fewest samples in this dataset (though the labels are almost completely balanced). This could be explained by

| Experiment | Accuracy (test) | Accuracy (validation) |
|---|---|---|
| 0 Random | $0.24 \pm .019$ | |
| 0 SER2SER (DCNN) | | $0.415 \pm .024$ |
| 0 SER2SER (ADCRNN) | | $0.534 \pm .019$ |
| 1a SER2MER (DCNN) | $0.3317 \pm .017$ | $0.419 \pm .02$ |
| 1b MER2MER (DCNN) | $0.431 \pm .018$ | $0.48 \pm .008$ |
| 2a SER2MER (ADCRNN) | $0.307 \pm .012$ | $0.347 \pm .009$ |
| 2b MER2MER (ADCRNN) | $0.492 \pm .023$ | $0.516 \pm .011$ |

Table 5.1: Performance results from experiments, including test and validation scores

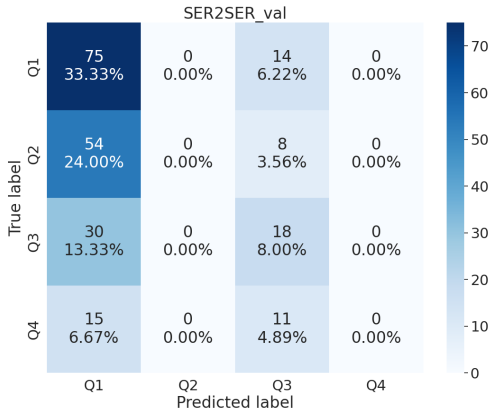
the model getting stuck in a local loss-minima during training (lowering learning rate and/or changing loss function are possible remedies)... Regardless of biases, if we look at the "true label" and grouping Q1+Q2 and Q3+Q4, i.e. high vs. low arousal, one can see evidence of useful arousal features being both learned and transferred between the domains. E.g. the predictions are usually low-arousal for samples that are low-arousal, and vice-versa. The same trend applies to the low-arousal area for SER2MER, though a bit weaker, as it is strongly influenced by the bias for Q1 here. A greater skill within arousal than valence is very prominent for MER2MER. When it comes to valence (Q1+Q4 vs. Q2+Q3), both models were having more trouble.

## Experiment 2: ADCRNN Architecture

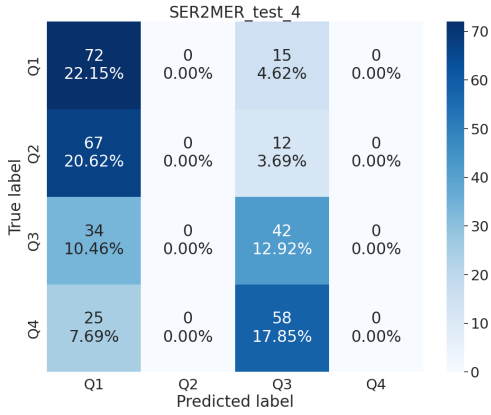
Looking at the experiments in table 5.1, there is a much higher performance for ADCRNN settings than DCNN, except for SER2MER. This means that the addition of the dilated RNN and attention to the architecture was beneficial intra-domain, but has not been shown to be beneficial for cross-domain. The (2a) (cross-domain) result is not much lower than the same setting for DCNN (1a), on test-sets, but falls 7 base-points below DCNN on validation sets (table 5.1). Further, the ADCRNN SER2MER (2a) has a much lower drop from validation to test score—this so far means the ADCRNN performs worse on test than DCNN, but was more generalizable from the validation-set (i.e. it does a 'bad' job very well).

Figure 5.2 shows confusion matrices for SER2SER (validation), SER2MER (test) and MER2MER (test)—for the best run for each. Accordingly, MER2MER

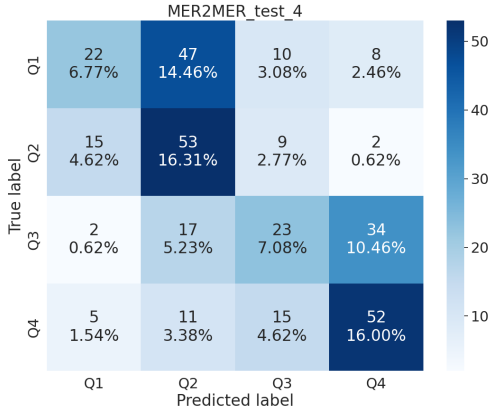
does very well and picks up useful valence- as well as arousal features (see the diagonal trend), yet again the MER2MER setting does better for arousal than valence (clear preferences e.g. for correct Q1+Q2 vs. Q3+Q4). The diagonal trend shows that the architecture is generally well-suited for the MER2MER task and data. SER2SER also shows some suitability, though the result is less generalizable (no test-set). SER2MER is biased towards Q1 in experiment 1 and instead towards Q3 in experiment 2 (more on biases in chapter 6). It is noteworthy that when the model here first ventures away from its bias, it does so with high accuracy (52.1%, according to the confusion matrix) for Q1, instead of being more random and distributed. In the SER2MER and SER2SER settings, both the architectures (see also fig. 5.1) show lack of intelligence for Q4 (nearly no predictions; more about this in chapter 6), and all except ADCRNN SER2SER show the same for Q2. Experiment 2 shows that some useful arousal features were successfully transferred across domains (especially when considering the high Q1 accuracy), though the bias for Q3 distorts much of the significance, especially for the Q3+Q4 part (true labels), which were nearly always predicted as low-arousal. If one compared with an imagined classifier that always predicted Q3 (like the high bias here), the observed lower-arousal 'transfer' would be nearly the same as now. The arousal transfer were regardlessly stronger in experiment 1.




(a) SER2SER (validation)

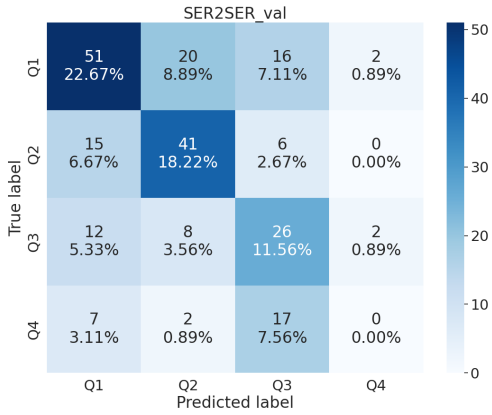



(b) SER2MER (test)

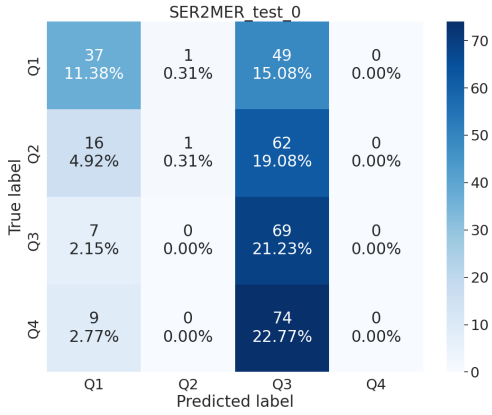



(c) MER2MER (test)

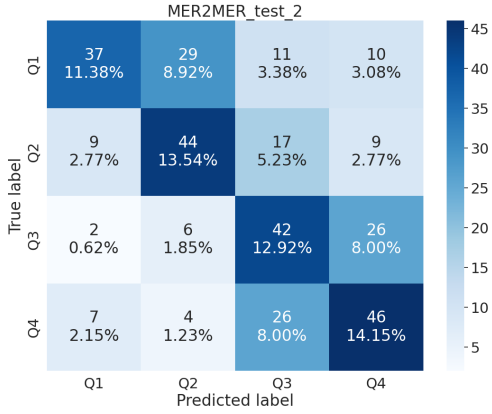
Figure 5.1: DCNN confusion matrices (experiment 0 and 1)




(a) SER2SER (validation)




(b) SER2MER (test)




(c) MER2MER (test)

Figure 5.2: ADCRNN confusion matrices (experiment 0 and 2)

# Chapter 6

# Discussion

In this chapter, the results (from chapter 5) are thoroughly discussed, causes and solutions are suggested and the most important implications of the results are drafted. Comparing the performance-results of SER2MER (speech to music) with MER2MER (music to music) is sufficient for answering research question (RQ) 3 of this thesis. The results are not directly compared to any other work, performance-wise, since the most similar studies Coutinho et al. (2014), Coutinho and Schuller (2017) and Weninger et al. (2013) all use time-continuous ratings, as opposed to static ratings. The former two research questions have already been discussed in chapter 4.

## 6.1 The Results

The overlap for arousal demonstrated in experiment 1 is large, but was not shown to completely overlap. It is unknown whether the true overlap is less than full, or if the experimental conditions were unable to exploit it to its potential. For example, the features learned can be too specialized to SER, even though more transferable features were latent (this idea is expanded later in this chapter). In experiment 1, all the settings showed particular difficulty with valence distinction, this suggests that the architecture and feature combination used here is insufficient for that part of the classification problem. The same observation also supports the notion that valence usually is more difficult to classify than arousal (both for humans and machines) (e.g. Grekow, 2021; Griffiths et al., 2021).

Experiment 1 showed a big performance drop from validation to test data for SER2MER and MER2MER. Since this did not happen for MER2MER in experiment 2, it suggests that the MER test data is representative enough, while overfitting to the validation data occurred in experiment 1 (through over-tuning).

Therefore, it is recommended to use cross-validation during model-tuning (especially considering small validation and test sizes), though wrongly assumed it would be unnecessary. Early-stopping and checkpointing by validation-loss instead of accuracy could also give a more generalizable model, since two different weight-configurations could give the same accuracy on one validation-set, while one of the two weight configurations can be more generalizable. If SER2SER had its own test-set, it could have shown better generalization from its validation-set than what was demonstrated by the MER2MER, due to its large training-set size.

Let us direct our *attention* over to the ADCRNN results. The addition of the attention dilated RNN (ADRNN) to the DCNN architecture improved intra-domain, but not cross-domain performance (table 5.1). Keep in mind that little tuning was done this time, so it is not necessarily the case that a model without these layers is the better choice (though it could be). The drop from the validation to test data was smaller here than for the large drop in experiment 1, which means this architecture at least was less overfitted to the validation data. A weaker transfer of arousal was shown in general, compared to DCNN (fig. 5.2; section 5.3). Nevertheless, since Q1 had significant accuracy, while the rest of the predictions were biased towards Q3, it means that both valence and arousal features were notably transferred, and the valence part being significantly higher than in DCNN (which was close to none; fig. 5.1). If less valence and arousal features were transferred, the predictions that were not Q3 would have been more random. Since the most impressive part of SER2MER confusion matrix in ADCRNN (fig. 5.2) is the surprising accuracy for Q1 (majority class), this again points to the need for more SER data for the other classes, perhaps not due to an imbalance, but simply insufficient training material for a good cross-domain performance for these classes—at least under the lens of the current architecture.

The SER2SER ADCRNN confusion matrix showed a diagonal trend, except for mixing up Q3 vs. Q4 (perhaps due to lack of data, since Q4 is minority), which hints at some degree of compatibility of the architecture to the problem. Remember that the strength of this specific result is weakened without a separate test set. A diagonal trend was shown for MER2MER too, and even stronger than for the SER2SER, which again suggests some degree of architecture vs. problem match. There are room for lots of performance improvements intra-domain, especially if we consider the results of Truong Pham et al. (2021) for SER2SER (91.9% accuracy), who used another dataset and a discrete emotion taxonomy. A diagonal trend was not observed for SER2MER, which implies either that the true overlap of emotional code between the domains is less than full, or the features learned when training on SER was not generalizable enough for prediction on MER. This latter option is possible if the learned features were specialized to features of the vocal tract and the fact that the sound is 'human voice', while

lower-level (common) patterns in the data unexploitedly exists. One of the limitations that could have caused this is the low resolution used for the extracted 3D log mel-spectrograms. To add, MER2MER also shows signs of insufficient training data, as it gets the 'main idea' through its diagonal trend, but still has large room for accuracy improvement.

Due to ending up with high values for the units in these model-layers (ADCRNN), these layers are potentially too complex for optimal learning generalization both intra- and cross-domain. On the other hand, the model might instead be insufficiently complex, failing to see grand-scheme connections in the data. The ADRNN part of the ADCRNN had 2.67 million parameters vs. 225 thousand for the DCNN part. An optimal balance and level of complexity between the DCNN and ADRNN parts should be explored in future work, as well as doing thorough tuning of hyperparameters and exploring strategies that can improve generalization. One such example is to combine softmax and center loss into a combined loss function, similarly to Truong Pham et al. (2021).

There was a lack of Q2 and Q4 predictions for all SER variants (intra- and cross-domain, for both experiments), except for SER2SER ADCRNN. Was there a lack of data here (class-imbalance), which caused this, or are the samples of this class too close to the true class boundaries (i.e. being close to neutral in at least one VA-dimension)? All SER2MER models and one SER2SER model (across both experiments) showed bias towards either Q1 or Q3, and as a matter of fact most predictions for the SER models were Q1 or Q3. This suggests that these classes were easiest/optimal to distinguish, and likely had the largest inter-centroid distances in the learned feature space (Q1 and Q3 are also *diagonal* to each other in the VA-space). There were more samples in Q1+Q3 than Q2+Q4 combined, which can explain why this pair was favored (i.e. minimizes total loss). This supports the future attempt of a combined loss function (section 4.4), which has the potential to separate the classes in a more feasible way. Further, remember that down-sampling was attempted during DCNN validation trials (not during ADCRNN!), but did not improve the DCNN SER2MER case. The ultimate scenario would regardless be to have a balanced SER dataset from the get-go, with total samples per class similar to Q1 (ca. 14k). Perhaps future MSP-podcast dataset releases will be more balanced, or that data augmentation strategies can be explored (see e.g. Truong Pham et al., 2021). Additionally, using f1-macro score—or weighted-average f1, which also respects per-class performance—as the measurement-metric for early-stopping and checkpointing could mitigate biases even more than using accuracy or (the current) loss-function.

## 6.2 Implications

The main finding in this research is that *some* learned features were directly transferable from SER to MER, though not all of them. This is reflected by significantly stronger SER2MER performance than the random baseline. If all features were transferable (full emotional code overlap), one should expect higher performance for SER2MER compared to MER2MER, due to more training data. One implication of these results is the reinforced evidence, in general, of common emotional coding between SER and MER, which has already been suggested by earlier research (chapter 3)—and from now on the notion also applies to *instrumental* music. Some overlap between the domains has here been proved for arousal (strongest) and some for valence. The fact that not all learned features in SER were generalizable is directly congruent with neurological research which has identified areas in the brain which are *common*, and some *specialized*, regarding emotion recognition from speech and music (Frühholz et al., 2016). What strengthens the connection to this is that good performance was achieved for the intra-domain settings, yet much lower performance cross-domain. If humans have specialized mechanisms for some aspects of affective perception of sound, then this could indeed mean the underlying emotional coding of the sound domains *are* indeed not fully overlapping. Another possibility is: it could also be that the specialized areas are responsible for features that are not as effective as the others (the analogous in this work would be a local optima during training), and that these areas are older in the brain. As an analogy, imagine a tribe of monkeys that uses one strategy, for one task, because its the only and best one they know of. Suddenly one day they discover a new one, and both these strategies are inherited through evolution, while the less important/effective gets gradually phased out over generations. Similarly, these specialized areas observed in the brain could be areas that are simply phasing out, instead of existing due to different emotional coding.

The results confirm that there exists a common emotional coding for arousal and for valence between the domains of speech and instrumental music. Simultaneously the true degree of overlap is not concluded, though it seems larger for arousal than valence. One could counter the proposed significance of the finding of a (partly) shared arousal coding and suspect that the loudness of a track is a sufficient feature for arousal predicition, but studies looking at feature selection for SER alone shows it is insufficient (e.g. RMS in Abri et al., 2021).

Since MSP-podcast SER data also comes with discrete labels, it is also an idea to compare performance of the SER architecture on experiments that use discrete labels and dimensional/quadrants, to check the impact of emotional taxonomy choice. It is also possible to run transfer learning experiments like in this work, except explore a regression problem or VA-classification with finer granularity

quadrants. In that case, more instrumental music data with VA-ratings will be needed (e.g. EMOPIA had only quadrants).

Another implication of the findings is that SER data indeed should be able to improve a MER classifier by first pre-training on SER, then continuing tuning of the weights with the limited MER data available—both for instrumental music and music with vocals, in fact it should boost performance even more, the more vocals are in the music. This is recommended as future work for those that are interested.

All the findings of this present research are obviously limited to and impacted by the datasets, extracted features, architectures and experimental setting used. Which are all elements that is recommended for further exploration. Much useful discussion and reasoning has already been provided in chapter 4. Some additions, though: The dataset could be a limiting factor, when it comes to generalizability of the results. For instance, EMOPIA which is only piano music, might have lead a MER model to be overly specialized to find emotion in piano sounds. Even worse, it is possible that there are more similarities between piano music and speech than other types of music. Also, shortening the duration of MER samples to be closer to the SER data could be helpful, even though the attention mechanism should already mitigate much weakness from this large duration difference (by learning that padded parts are unimportant. Simultaneously, as the classes are of low granularity, and the MER dataset compilation process has been thorough, the datasets are very likely of high enough quality to give generalizable results.

# Chapter 7

# Conclusion

The goal of this Master's thesis was to use transfer learning to map from speech emotion recognition (SER) to instrumental music emotion recognition (MER)—with a focus on exploring the amount of emotional coding overlap between these two domains of affective sound (see research question (RQ) 3). The work included a structured literature review, concerning both fields separately, as well as reviewing related work focused on transfer learning between them. Further, two emotional taxonomies were compared for suitability in this context (RQ 1). A novel instrumental music emotion-dataset was compiled (static valence-arousal ratings). This novel dataset was combined with some samples from an online piano dataset which uses quadrant-ratings (which means lower granularity, than valence-arousal ratings). The novel dataset was then converted to quadrants in order to create an easier machine learning problem and enable a larger amount of data for training for the large-scale experiments of this study (RQ 3). A custom *dilated LSTM* (dilated Long Short-term Memory) Keras-layer was implemented. Discussions and implications of the experimental results were included in chapter 6. The rest of this chapter concludes each research question and summarizes all the suggested future work.

## 7.1 Research Questions

RQ 1 asked whether categorical or dimensional (valence-arousal plane in this case; VA) emotion taxonomies were superior for emotion recognition from sound. The answer is concluded to be very application dependent (section 4.1). While both types have validity-deficiencies, the dimensional VA-plane can model more nuanced emotional experiences and usually gives higher agreement among raters. A dataset labeled with VA-ratings can be transformed into various machine learn-

ing problems of varying granularity, e.g. regression problems and the popular four-quadrant (4Q) problem. Thus, a custom instrumental music dataset was compiled with the 2D VA-plane (Russell's Circumplex). 3-dimensional emotion taxonomies were not considered in this research.

RQ 2 sought to acquire feasible datasets (see section 4.2) for this thesis' goal; to do transfer learning between speech and instrumental music emotion recognition. A large natural dataset for SER (MSP-podcast) was acquired (70k samples originally and about 40k used here), which had categorical and dimensional emotion labeling. In order to have enough instrumental MER data, a new instrumental collection was compiled through thorough selection of subsets from two existing dimensional music datasets (e.g. tracks including vocals were filtered out). This compiled instrumental MER dataset had 742 samples between 10-15 seconds, and is labeled with one label per sample (*static* emotion recognition). This compiled dataset was further combined with the 4Q piano dataset EMOPIA (868 samples from the original dataset; 10-15 seconds; max 5 samples from each composition). A feasible split for training, validation and testing sets were made, which ensured that no composition was represented in more than one partition.

The main RQ, number 3, was: "How does training on emotional speech affect recognition performance for instrumental music emotion recognition?". This was explored through extracting 3D log mel-spectrograms (section 4.3), converting labels into a four-quadrant (of the VA-plane; 4Q) problem, then training on SER and testing directly on MER test-set (SER2MER) (chapter 5). Two neural network architectures (section 4.4) were tested: *dilated convolutional neural network* (DCNN) and *attention dilated convolutional recurrent neural network* (ADCRNN)—which were both custom for this study, though heavily inspired by Truong Pham et al. (2021). A custom Keras-layer implementation was created for a dilated LSTM (based on Chang et al., 2017). The results of SER2MER were compared to the MER2MER experimental setting, and higher SER2MER than MER2MER performance was expected if there was complete overlap of the underlying emotional codes (chapter 5). Both architectures showed higher intra-domain performance (SER2SER and MER2MER) than cross-domain. More specifically, DCNN SER2MER and MER2MER accuracy was 33.2% and 43.1%, respectively. ADCRNN SER2MER vs. MER2MER was 30.7% and 49.2%. The results implies either that the true overlap of emotional code between the domains is less than full, or the features learned for SER was not generalizable enough for the MER test data—compared to MER2MER setting (chapter 6). The ADCRNN demonstrated a better fit for the problem than DCNN, but had also undergone longer training. Additionally, confusion matrices showed that strong, although not perfect, arousal feature transfer took place (DCNN more than ADCRNN). The ADCRNN experimental results proved that notable valence feature transfer had happened (more than DCNN).

Neurological research has confirmed shared *and* uncommon brain structures for emotion recognition from music and speech (*instrumental* music not explored; Frühholz et al., 2016). This is fitting with this research's findings. As in chapter 6: it is also possible that the specialized brain-areas are responsible for emotion features that are not as effective as the others, and that these areas are older in the brain and are being phased out (humans continue to evolve).

## 7.2 Future Work

Significant improvements are likely possible both for intra- and cross-domain performance. This was a proof-of-concept study, when it comes to *instrumental* music, and fairly little tuning was done, mostly due to time-constraints. Future work is recommended to explore:

- There is a need for more SER-data for the minority classes, which could also help mitigate the learned biases (chapter 6). MER2MER could also benefit from more data overall. Using data-augmentation strategies are one possible way to get more data. Additionally, it is recommended to analyze the distribution in VA-space (not quadrants) for MER and SER data, which was not done now. As discussed in chapter 6, the reason for low number of predictions for Q4-class could be that the Q4 SER training-samples were close to neutral in at least one of valence and arousal dimensions.

- More similar lengths of each sample, like shortening MER samples to be closer to SER durations. Do keep in mind also that criteria (3) for instrumentalness (section 4.2.2; some short incongruent phrases could be included) could have lead to confusion for the SER model, for some samples in this work—and that it could be even more problematic if samples were shortened more, and that they unluckily end up containing more of the 'meaningless' vocal portions. There were few samples like this now, though.

- Tuning of the log mel-spectrogram parameters. It is suspected that the resolution was too low (chapter 6). It is also recommended to explore 'random padding' during feature-extraction (section 4.2.2).

- Further explorations of model complexity (especially balancing the complexity of CNN and RNN parts of ADCRNN), testing a combined loss-function, thorough hyper-parameter tuning, and explore more dilated RNN layers and/or different dilation rates for these. See chapter 6 for more details. The first dense-layer in DCNN-architecture is probably using way too few units, considering the long vector from the flatten-layer.

- Using cross-validation for tuning and/or f1-macro for early-stopping and checkpointing. Using loss instead of accuracy for these strategies can also help generalization (chapter 6).
- It was forgotten to add activation to one dense layer in ADCRNN during these experiments (section 4.4). Relu was intended. Another complete run on the test-set was done at the end, just to check (with relu added and ADCRNN), which surprisingly gave lower performance (a drop of 2 base-points) on SER2MER (28% accuracy) and SER2SER (51.9% accuracy), while maintaining MER2MER performance (49.2%). This change to the architecture should be further investigated, also with new tuning.
- Pre-training on SER and continue training training on MER, if higher MER performance is the goal. In this work, it was only explored to train on SER and test directly on MER, for the cross-domain experiments (chapter 5).
- To test discrete labels, or finer granularity of bins from the VA-plane than 4 quadrants (see section 4.2.2 for limitations to level of granularity that is possible). Regression is also possible (section 4.1). This way one can see how the emotion taxonomy, or label type, interacts with one fixed model architecture.
- Better feature transfer might be possible in a dynamic as opposed to static emotion classification problem, by learning VA-ratings for multiple parts of a sample, like for every half second (static means one value for the whole sample). In this way, the model could potentially learn more specific features and become more accurate in its predictions. Note that efforts were already made, for the compiled MER dataset, to create clips (segments from a song) by selecting windows from each original song that kept intra-sample emotional content variation within a certain threshold (section 4.2).

# References


Abri, F., Gutiérrez, L. F., Datta, P., Sears, D. R., Siami Namin, A., & Jones, K. S. (2021). A comparative analysis of modeling and predicting perceived and induced emotions in sonification. *Electronics*, *10*(20), 2519.

Agrawal, Y., Shanker, R. G. R., & Alluri, V. (2021). Transformer-based approach towards music emotion recognition from lyrics. *Advances in Information Retrieval*, 167–175. Retrieved from http://dx.doi.org/10.1007/978-3-030-72240-1_12 doi: 10.1007/978-3-030-72240-1_12

Ahmed, N., Natarajan, T., & Rao, K. (1974). Discrete cosine transform. *IEEE Transactions on Computers*, *C-23*(1), 90-93. doi: 10.1109/T-C.1974.223784

Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2017, March). Developing a benchmark for emotional analysis of music. *PLOS ONE*, *12*(3), e0173392. doi: 10.1371/journal.pone.0173392

Anderson, I., Gil, S., Gibson, C., Wolf, S., Shapiro, W., Semerci, O., & Greenberg, D. M. (2021, May). "Just the way you are": Linking music listening on spotify and personality. *Social Psychological and Personality Science*, *12*(4), 561–572. doi: 10.1177/1948550620923228

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Böck, R., Egorow, O., Siegert, I., & Wendemuth, A. (2017). Comparative study on normalisation in emotion recognition from speech. In *International conference on intelligent human computer interaction* (pp. 189–201).

Brewer, M., & Rahman, J. S. (2020). Pruning long short term memory networks and convolutional neural networks for music emotion recognition. In *Neural information processing* (pp. 343–352). Cham: Springer International Publishing.

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

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., ...

Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335–359.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., & Mower Provost, E. (2017, jan). MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, *8*(1), 67-80. doi: 10.1109/TAFFC.2016.2515617

Cañón, J. S. G., Cano, E., Herrera, P., & Gómez, E. (2020). Transfer learning from speech to music: towards language-sensitive emotion recognition models. In *2020 28th european signal processing conference (EUSIPCO)* (pp. 136–140).

Cañón, J. S. G., Cano, E., Pandrea, A. G., Herrera, P., & Gómez, E. (2021, June). Language-sensitive music emotion recognition models: Are we really there yet? In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 576–580). doi: 10.1109/ICASSP39728.2021.9413721

Chang, S., Zhang, Y., Han, W., Yu, M., Guo, X., Tan, W., ... Huang, T. S. (2017). *Dilated recurrent neural networks.* arXiv preprint arXiv:1710.02224.

Chen, M., He, X., Yang, J., & Zhang, H. (2018). 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, *25*(10), 1440–1444.

Chien, J.-T., & Bao, Y.-T. (2017). Tensor-factorized neural networks. *IEEE transactions on neural networks and learning systems*, *29*(5), 1998–2011.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789–8797).

Collier, G. L. (2007). Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, *35*(1), 110–131.

Coutinho, E., Deng, J., & Schuller, B. (2014, July). Transfer learning emotion manifestation across music and speech. In *2014 International Joint Conference on Neural Networks (IJCNN)* (pp. 3592–3598). doi: 10.1109/IJCNN.2014.6889814

Coutinho, E., & Schuller, B. (2017, June). Shared acoustic codes underlie emotional communication in music and speech—Evidence from deep transfer learning. *PLOS ONE*, *12*(6), e0179289. doi: 10.1371/journal.pone.0179289

Cowen, A. S., Laukka, P., Elfenbein, H. A., Liu, R., & Keltner, D. (2019, April). The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, *3*(4), 369–382. doi: 10.1038/s41562-019-0533-6

Cross-corpus speech emotion recognition based on sparse subspace transfer learning. (n.d.). In *Biometric Recognition.*

de Lope, J., Hernández, E., Vargas, V., & Graña, M. (2021). Speech emotion recognition by conventional machine learning and deep learning. In H. Sanjurjo González, I. Pastor López, P. García Bringas, H. Quintián, & E. Corchado (Eds.), *Hybrid artificial intelligent systems* (pp. 319–330). Cham: Springer International Publishing. doi: 10.1007/978-3-030-86271-8_27

De Lathauwer, L. (2008). Decompositions of a higher-order tensor in block terms—part II: Definitions and uniqueness. *SIAM Journal on Matrix Analysis and Applications*, *30*(3), 1033–1066.

Dhondge, S., Shewale, R., Satao, M., & Jagdale, J. (2022). Impact of lightweight machine learning models for speech emotion recognition. In *International conference on innovative computing and communications* (pp. 249–261). Singapore: Springer Singapore.

Djupvik, A. (2020). *Music that feels just right* (Master's thesis). NTNU.

Dossou, B. F. P., & Gbenou, Y. K. S. (2021). *FSER: Deep convolutional neural networks for speech emotion recognition.* arXiv preprint arXiv:2109.07916.

Eerola, T. (2018, March). Music and emotion. In R. Bader (Ed.), *Bader, Rolf (Eds.). (2018). Springer handbook of systematic musicology. Berlin, Heidelberg: Springer, pp. 539-554, Springer handbooks* (pp. 539–554). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-662-55004-5_29

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49.

Ekman, P. (2016). What scientists who study emotion agree about. *Perspectives on psychological science*, *11*(1), 31–34.

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., ... Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, *53*(4), 712.

Fan, J., Yang, Y.-H., Dong, K., & Pasquier, P. (2020). A comparative study of western and chinese classical music based on soundscape models. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 521–525).

Farris, N., Model, B., Savery, R., & Weinberg, G. (2021, June). Musical prosody-driven emotion classification: Interpreting vocalists portrayal of emotions through machine learning. *arXiv:2106.02556 [cs, eess]*.

Feng, K., & Chaspari, T. (2021). Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing*, 1-1. doi: 10.1109/TAFFC.2021.3109485

Frühholz, S., Trost, W., & Kotz, S. A. (2016, September). The sound of emo-

tions—Towards a unifying neural network perspective of affective sound processing. *Neuroscience & Biobehavioral Reviews*, *68*, 96–110. doi: 10.1016/j.neubiorev.2016.05.002

Gideon, J., McInnis, M. G., & Provost, E. M. (2021, October). Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). *IEEE Transactions on Affective Computing*, *12*(4), 1055–1068. doi: 10.1109/TAFFC.2019.2916092

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Grekow, J. (2021, August). Music emotion recognition using recurrent neural networks and pretrained models. *Journal of Intelligent Information Systems*. doi: 10.1007/s10844-021-00658-5

Griffiths, D., Cunningham, S., Weinel, J., & Picking, R. (2021, September). A multi-genre model for music emotion recognition using linear regressors. *Journal of New Music Research*, 1–18. doi: 10.1080/09298215.2021.1977336

Haq, S., & Jackson, P. (2010, aug). Machine audition: Principles, algorithms and systems. *IGI Global*, 398-423.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hu, X., & Yang, Y.-H. (2017). Cross-dataset and cross-cultural music mood prediction: A case on western and Chinese pop songs. *IEEE Transactions on Affective Computing*, *8*(2), 228–240.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Hung, H.-T., Ching, J., Doh, S., Kim, N., Nam, J., & Yang, Y.-H. (2021). *EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation.* arXiv preprint arXiv:2108.01374.

Jahangir, R., Teh, Y. W., Nweke, H. F., Mujtaba, G., Al-Garadi, M. A., & Ali, I. (2021). Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges. *Expert Systems with Applications*, *171*, 114591. doi: https://doi.org/10.1016/j.eswa.2021.114591

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770–814. doi: 10.1037/0033-2909.129.5.770

Kofod-Petersen, A. (2018). *How to do a Structured Literature Review in computer science* (Tech. Rep.). Department of Computer Science, Norwegian University of Science and Technology. Retrieved from https://research.idi.ntnu.no/aimasters/files/SLR_HowTo2018.pdf

Kossaifi, J., Lipton, Z. C., Kolbeinsson, A., Khanna, A., Furlanello, T., & Anandkumar, A. (2020). Tensor regression networks. *Journal of Machine Learning Research*, *21*, 1–21.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., ... Pantic, M. (2021). SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(3), 1022-1040. doi: 10.1109/TPAMI.2019.2944808

Krishnaiah, A., & Divakarachari, P. (2021, October). Automatic music mood classification using multi-class support vector machine based on hybrid spectral features. *International Journal of Intelligent Engineering and Systems*, *14*(5), 102–111. doi: 10.22266/ijies2021.1031.10

Kumar, Y., & Gupta, S. (2021). Development of music player application using emotion recognition. *Journal for Modern Trends in Science and Technology*, *7*(01), 54–57.

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Li, L.-Q., Xie, K., Guo, X.-L., Wen, C., & He, J.-B. (2021). Emotion recognition from speech with StarGAN and Dense-DCNN. *IET Signal Processing*, *n/a*(n/a). doi: 10.1049/sil2.12078

Liu, Z.-T., Xie, Q., Wu, M., Cao, W.-H., Mei, Y., & Mao, J.-W. (2018). Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing*, *309*, 145-156. doi: https://doi.org/10.1016/j.neucom.2018.05.005

Livingstone, S. R., & Russo, F. A. (2018). The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

Lotfian, R., & Busso, C. (2019, oct). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, *10*(4), 471-483. doi: 10.1109/TAFFC.2017.2736999

McFee, B., Metsai, A., McVicar, M., Balke, S., Thomé, C., Raffel, C., ... Thassilo (2021, May). *Librosa/librosa: 0.8.1rc2.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.4792298 doi: 10.5281/zenodo.4792298

Mehrabian, A. (1995). Framework for a comprehensive description and measurement of emotional states. *Genetic, social, and general psychology monographs*.

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology.* Cambridge, MA, US: The MIT Press. (Pages: xii, 266)

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE access*, *7*, 125868–125881.

Nadon, É., Tillmann, B., Saj, A., & Gosselin, N. (2021). The emotional effect of background music on selective attention of adults. *Frontiers in Psychology*, 4472.

Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Panda, R., Malheiro, R., & Paiva, R. P. (2018). Musical texture and expressivity features for music emotion recognition. In *19th international society for music information retrieval conference (ISMIR 2018* (pp. 383–391).

Panda, R., Malheiro, R., & Paiva, R. P. (2020). Novel audio features for music emotion recognition. *IEEE Transactions on Affective Computing*, *11*(4), 614-626. doi: 10.1109/TAFFC.2018.2820691

Pandey, S. K., Shekhawat, H. S., & Prasanna, S. R. M. (2022, January). Attention gated tensor neural network architectures for speech emotion recognition. *Biomedical Signal Processing and Control*, *71*, 103173. doi: 10.1016/j.bspc.2021.103173

Praseetha, V., & Joby, P. (2021). Speech emotion recognition using data augmentation. *International Journal of Speech Technology*, 1–10.

Preeti GUPTA, S. D. (2021, aug). Results of a novel music player using speech and text emotion recognition for mood uplift. *Design Engineering*, 6222-6232. Retrieved from http://www.thedesignengineering.com/index.php/DE/article/view/3122

Qi, C., & Su, F. (2017). Contrastive-center loss for deep neural networks. , 2851–2855.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. (Place: US Publisher: American Psychological Association) doi: 10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. doi: 10.1037/0033-295X.110.1.145

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of research in Personality*, *11*(3), 273–294.

Scherer, K. R. (2005, December). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. Retrieved 2021-11-26, from https://doi.org/10.1177/0539018405058216 doi: 10.1177/0539018405058216

Schoene, A. M., Turner, A. P., & Dethlefs, N. (2020). Bidirectional dilated LSTM with attention for fine-grained emotion classification in tweets. In *Affcon@ aaai* (pp. 100–117).

Sefara, T. J. (2019). The effects of normalisation methods on speech emotion

recognition. In *2019 international multidisciplinary information technology and engineering conference (IMITEC)* (pp. 1–8).

Shuman, V., Schlegel, K., & Scherer, K. (2015, August). *Geneva Emotion Wheel Rating Study* (Tech. Rep.).

Själander, M., Jahre, M., Tufte, G., & Reissmann, N. (2019). EPIC: An energy-efficient, high-performance GPU computing research infrastructure. *arXiv preprint arXiv:1912.05848*.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Tang, G., Liang, R., Xie, Y., Bao, Y., & Wang, S. (2019). Improved convolutional neural networks for acoustic event classification. *Multimedia Tools and Applications*, *78*(12), 15801–15816.

Truong Pham, N., Dang, D. N. M., & Dzung Nguyen, S. (2021, September). *Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition* (1. ed.). arXiv e-prints. Retrieved from https://arxiv.org/abs/2109.09026v1

Tzanetakis, G., Essl, G., & Cook, P. (2001). *Automatic musical genre classification of audio signals.* The International Society for Music Information Retrieval. Retrieved from http://ismir2001.ismir.net/pdf/tzanetakis.pdf

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on machine learning* (p. 1096–1103). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1390156.1390294 doi: 10.1145/1390156.1390294

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE transactions on multimedia*, *10*(5), 936–946.

Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., & Scherer, K. (2013). On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology*, *4*.

Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, *53*(5), 768-785. (Perceptual and Statistical Audition) doi: https://doi.org/10.1016/j.specom.2010.08.013

Xu, L., Xu, W., & Zhang, W. (2021). Multi-dimensional music emotion recognition incorporating convolutional neural networks and Plutchik's emotion wheel. *AMCIS 2021 Proceedings. 9*.

Yu, F., & Koltun, V. (2015). *Multi-scale context aggregation by dilated convolutions.* arXiv preprint arXiv:1511.07122.

# Appendix A

# Structured Literature Review Protocol

## A.1 Introduction

This protocol describes each step of the project's structured literature review (SLR), for reproducibility. This SLR-protocol is used to explore the fields of automatic music and speech emotion recognition (MER and SER), and to increase the chances of discovering the newest, most relevant and highest quality publications. The structure of this protocol is based on the guide in Kofod-Petersen (2018).

## A.2 Research questions

In alignment with the preparation project's (PP) research goal: *Give an overview of the fields of automatic music emotion recognition and speech emotion recognition*—some research questions were created. The short version is given here (see section 3.1 for more elaboration on these)

**PP.Research question 1** *What is state-of-the-art for music emotion recognition and speech emotion recognition?*

**PP.Research question 2** *Do the findings to research question 1 support the use of transfer learning between the two domains?*

**PP.Research question 3** *Do the findings motivate other future work that is suitable for a Master's thesis?*

## A.3 Search strategy

The source for this SLR is Google Scholar, which is a search engine for references, and that searches through multiple academic resources. Google Scholar includes features like citations-search (i.e. find citations of an article) and other handy advanced search tools. Key search terms have been identified through exploring related work from Djupvik (2020) and some *snowballing* (explained below). The terms (Table A.1) are grouped by similar semantic meaning, except *music* and *speech*.

| | **Group 1** | **Group 2** | **Group 3** |
|---|---|---|---|
| **Term 1** | music | mood | classification |
| **Term 2** | musical | emotion | recognition |
| **Term 3** | speech | affect | detection |
| **Term 3** | | ambiance | |

Table A.1: Search terms and groups in SLR

In the final search query, the command *allintitle* is used to reduce the presence of less relevant results, by only searching in article titles. Additionally the concatenation is done implicitly in Google Scholar, so $AND$ is skipped between the *groups* (marked by parentheses). For convenience, the SLR is split up in two parts; one for each domain (SER and MER). The resulting query is therefore split up, and is defined:

```
   allintitle:(Speech)
(Mood OR Emotion OR Affect OR Ambiance)
(Classification OR Recognition OR Detection)

   allintitle:(Music OR Musical)
(Mood OR Emotion OR Affect OR Ambiance)
(Classification OR Recognition OR Detection)
```

## A.4 Inclusion criteria

A set of inclusion criteria (IC) defines which articles from the search result to select and which to filter out. These ensure included studies are relevant (enough) to the SLR research questions. Primary IC are assessed solely on the abstract of the article. Secondary IC requires screening the full text. This results in a sequential process where only the most worthy articles deserve a full-text screening.

Due to time constraints: if there are many results, only the first 25 studies (sorted by "date") from each of MER and SER domains, which passes all primary IC, will be promoted and added to my article database. Similarly, only the first 12, sorted *ascending by date added to library*, from each domain which passes the secondary IC filters are promoted further. These limits clearly lead to excluding work that may have been useful, which is a significant limitation of this work. Some of this weakness can be mitigated to some degree by the use of snowballing—i.e. exploring the graph of references from a set of starting articles—to gather additional articles for review, from any publication year. All articles in my database (which pass primary IC) will be archived such that they can easily be found if my future efforts require additional knowledge which resonate with their titles (even if they are not considered in detail in this report).

### A.4.1 Primary inclusion criteria

In addition to the below primary criteria, some general removal criteria is defined: (1) for duplicate results, the highest ranking source is kept. (2) articles published before 2017 (arbitrary date) are removed. The reasoning for including (2) is that newer work tend to seek to improve upon earlier work (and tend to achieve this).

**IC 1** *The study's main concern is automatic classification of emotion in either music or speech—based on audio and not the lyrics*

**IC 2** *The study is a primary study (not solely a review) and presents empirical results*

**IC 3** *The study is written in English*

### A.4.2 Secondary inclusion criteria

The following secondary inclusion criteria is applied:

**IC 4** *The study describes the implementation of a system for the task. For neural networks, each layer-type in the architecture has to be specified.*

**IC 4** *The study seek to distinguish between 3 or more emotions, if it concerns discrete emotion classification. If a dimensional emotion taxonomy is used, it should have 2 or more dimensions.*

## A.5 Quality assessment

The resulting accepted set of articles from the SLR (referred to as the SLR-set) not only needs to be relevant, but also of high quality (including high strength of

the evidence presented). Therefore the following quality criteria (QC) are copied directly from Kofod-Petersen (2018). The amount of articles that will be brought through the quality check from each domain is dependent on what is feasible with regards to project progress and time until delivery deadline.

**QC 1** *Is there a clear statement of the aim of the research?*

**QC 2** *Is the study put into context of other studies and research?*

**QC 3** *Are system or algorithmic design decisions justified?*

**QC 4** *Is the test data set reproducible?*

**QC 5** *Is the study algorithm reproducible?*

**QC 6** *Is the experimental procedure thoroughly explained and reproducible?*

**QC 7** *Is it clearly stated in the study which other algorithms the study's algorithm(s) have been compared with?*

**QC 8** *Are the performance metrics used in the study explained and justified?*

**QC 9** *Are the test results thoroughly analyzed?*

**QC 10** *Does the test evidence support the findings presented?*

Each study is scored for each QC between yes (1 point), partly ($\frac{1}{2}$ point) or not at all (0 points). If a study scores 0 for any QC, except QC 8, it will be rejected. Further, those articles with a total score of less than 7 are deemed as insufficient quality.

## A.6 Data extraction

For each article in the SLR result set, the following data points are extracted and summarized. I argue that all of these aspects make up the building blocks for the findings of any affective sound study.

- Unique ID
- Author(s)
- Publication year
- Title

- Emotion taxonomy used
- Dataset (and whether acted/natural)
- Machine learning method(s)
- Features used
- Findings and conclusions (and whether cross-validated)

# Appendix B

# Review Quality Criteria Ratings

## B.1 Quality criteria check ratings

Table B.1 shows the ratings from the review process. Due to space, the quality criteria (QC) are referenced by their ID, which are repeated below.

- **QC1:** Is there a clear statement of the aim of the research?
- **QC2:** Is the study put into context of other studies and research?
- **QC3:** Are system or algorithmic design decisions justified?
- **QC4:** Is the test data set reproducible?
- **QC5:** Is the study algorithm reproducible?
- **QC6:** Is the experimental procedure thoroughly explained and reproducible?
- **QC7:** Is it clearly stated in the study which other algorithms the study's algo-rithm(s) have been compared with?
- **QC8:** Are the performance metrics used in the study explained and justified?
- **QC9:** Are the test results thoroughly analyzed?
- **QC10:** Does the test evidence support the findings presented?

Table B.1: Quality criteria (QC) ratings for the reviewed papers. EA means "et al.". Zhao EA first got 1 on method reproducibility, because the method is well explained, but the missing description of features used was overlooked, which was later discovered and lead to a 0-rating. This means it was also partially excluded from the related work part of the report.

| **Article ID** | **Article** | **QC1** | **QC2** | **QC3** | **QC4** | **QC5** | **QC6** | **QC7** | **QC8** | **QC9** | **QC10** | **SCORE** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SLR. SER1 | Pandey EA | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | 1 | 1 | **9,5** |
| SLR. SER2 | Throung Pam EA | 1 | 1 | 0,5 | 1 | 1 | 0,5 | 1 | 0,5 | 1 | 1 | **8,5** |
| SLR. SER3 | de Lope EA | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | 0,5 | 1 | 0,5 | **8,5** |
| SLR. SER4 | Li EA | 1 | 1 | 1 | 1 | 1 | 0,5 | 1 | 0,5 | 1 | 0,5 | **8,5** |
| SLR. SER5 | Zhao EA | 1 | 1 | 1 | 1 | 0 | 0,5 | 1 | 0,5 | 1 | 1 | **8** |
| SLR. MER1 | Krishnaiah and Divaka.. | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | 1 | 1 | **9,5** |
| SLR. MER2 | Cañon EA | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | 1 | 1 | **9,5** |
| SLR. MER3 | Farris EA | 1 | 1 | 1 | 0,5 | 1 | 1 | 1 | 0,5 | 1 | 1 | **9** |
| SLR. MER4 | Griffiths EA | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | **9,5** |
| SLR. MER5 | Grekow | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0,5 | 1 | 1 | **9,5** |

Jonas Klepper Rødningen

Emotion Recognition from Speech and Instrumental Music

NTNU
Kunnskap for en bedre verden