# Modelling affective states for transportation systems.

A Dissertation
for the degree of
Doctor of Engineering (Dr.-Ing.)
in the
Faculty of Applied Computer Science
University of Augsburg

presented by
**Vincent Karas**

2025

**First Examiner**
Prof. Dr.-Ing. habil. Björn Schuller

**Second Examiner**
Prof. Dr. rer. nat. Elisabeth André

**Date of submission**
12.09.2024

**Date of oral exam**
20.02.2025

# Declaration of Authorship

I, Vincent Karas, declare that this thesis titled, "Modelling affective states for transportation systems." and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

*"The difference between science and screwing around is writing it down."*

Adam Savage

UNIVERSITY OF AUGSBURG

# *Abstract*

Faculty of Applied Computer Science

Institute of Computer Science

Doctor of Engineering (Dr.-Ing.)

**Modelling affective states for transportation systems.**

by Vincent Karas

This thesis focuses on deep learning for automatic estimation of emotional states. Computational emotion recognition, which is part of the young and rapidly growing discipline of affective computing, has been applied to a large variety of fields in recent years, including education, customer services, and digital health and wellbeing. It has the potential to elevate human-machine interaction, by enabling systems to sense a user's emotions and adjust responses accordingly, for a more natural and intuitive experience. This is especially relevant given the increasing number and complexity of automated systems that we interact with in our daily lives. However, challenges still remain for deploying such technology "in the wild" i. e., under realistic conditions. The nature of emotions is yet not fully understood, and their expression and perception is nuanced, diverse, and strongly dependent on context e. g., cultural setting. Deep-learning based models require large amounts of data for training, but in the absence of an objective emotional ground truth, gathering annotations from humans is a labour-intensive and costly process. Furthermore, data collected in a natural setting is noisy, with varying environmental conditions and recording quality. Accessing multiple complementary signals e. g., vision and audio can help in this situation. Thus, the following thesis examines the multi-modal prediction of emotional states in the wild, from clues in the face and voice. Emotions are modelled as value-continuous variables, in particular valence and arousal. Their fluctuation across time is captured through sequence-based processing. To summarise the thesis, background of emotions theory and signal processing is presented first, followed by methods used for training the models. Experiments on multiple datasets are conducted and discussed, and recommendations for future research given. In terms of core contributions, this work examines different sets of features based on fine-tuned CNNs and Transformers, and their multi-modal fusion on challenging, noisy datasets. It analyses in detail the problem of emotion recognition on non-verbal vocal bursts such as laughter or crying, which are less commonly studied than speech. It addresses the issue of cross-cultural emotion recognition via domain adaptation on a multi-cultural dataset, and demonstrates that unlabelled data can be leveraged to boost recognition performance.

# *Acknowledgements*

I want to thank my family, especially my parents and my brother, for their great support during my time as PhD candidate. I also want to thank my friends for all the fun we had in the last years, as well as for their patience during the times when my work-life balance was more on the work side.

Great thanks is also owed to the current and former members of the Chair of Embedded Intelligence for Healthcare and Wellbeing, in particular to Adrià Mallol-Ragolta, Manuel Milling, Andreas Triantafyllopoulos, Maurice Gerczuk, Meishu Song, Dr. Lukas Stappen, Dr. Alice Baird, Dr. Zhao Ren and Dr. Shahin Amiriparian. Your scientific comments and discussions were invaluable, and it was a pleasure to work with all of you.

I would also like to thank everyone at the BMW Group who helped me during my time as a PhD program member. Special thanks is owed to my department head Dr. Jochen Eckert and to Dr. Christian Knoll, Björn Westphal and Dr. Hans-Jörg Vögel, who were essential for the start of my journey at BMW. I also want to thank my scientific advisors at BMW, Dr. Johannes Feldmaier and Dr. Klaus Ries, as well my team lead Andreas Menath and my mentor Klaus Kapp.

Finally, I want to thank my supervisor, Prof. Dr. Björn Schuller, for his insightful advice and guidance, without which this thesis would not have been possible.

# List of Publications

During the author's time as a PhD candidate, the following works were authored or co-authored:

Karas, V. and B. W. Schuller (2021). "Deep learning for sentiment analysis: an overview and perspectives". In: *Natural Language Processing for Global and Local Business*, pp. 97–132.

– (Aug. 2021). "Recognising Covid-19 from coughing using ensembles of SVMs and LSTMs with handcrafted and deep audio features". In: *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. Brno, Czech Republic: ISCA, pp. 4286–4290. DOI: 10.21437/Interspeech.2021-1267.

Karas, V., M. K. Tellamekala, A. Mallol-Ragolta, M. Valstar, and B. W. Schuller (2022). "Time-Continuous Audiovisual Fusion with Recurrence vs Attention for In-The-Wild Affect Recognition". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2381–2390. DOI: 10.1109/CVPRW56347.2022.00266.

Karas, V., A. Triantafyllopoulos, M. Song, and B. Schuller (July 2022). "Self-Supervised Attention Networks and Uncertainty Loss Weighting for Multi-Task Emotion Recognition on Vocal Bursts". In: *Proceedings of the ACII Affective Vocal Bursts Workshop Competition 2022 (A-VB)*, pp. 1–5. DOI: https://doi.org/10.48550/arXiv.2210.15754.

Karas, V., D. M. Schuller, and B. W. Schuller (2024). "Audiovisual Affect Recognition for Autonomous Vehicles: Applications and Future Agendas". In: *IEEE Transactions on Intelligent Transportation Systems* 25.6, pp. 4918–4932. DOI: 10.1109/TITS.2023.3333749.

Stappen, L., V. Karas, N. Cummins, F. Ringeval, K. Scherer, and B. Schuller (Sept. 2019). "From Speech to Facial Activity: Towards Cross-modal Sequence-to-Sequence Attention Networks". In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6. DOI: 10.1109/MMSP.2019.8901779.

Amiriparian, S., P. Winokurow, V. Karas, S. Ottl, M. Gerczuk, and B. Schuller (2020). "Unsupervised Representation Learning with Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech". In: *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. MuSe'20. Seattle, WA, USA: Association for Computing Machinery, pp. 11–17. DOI: 10.1145/3423327.3423670. URL: https://doi.org/10.1145/3423327.3423670.

Ottl, S., S. Amiriparian, M. Gerczuk, V. Karas, and B. Schuller (2020). "Group-level Speech Emotion Recognition Utilising Deep Spectrum Features". In: *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI '20. Virtual Event, Netherlands: Association for Computing Machinery, pp. 821–826. DOI: 10.1145/3382507.3417964. URL: https://doi.org/10.1145/3382507.3417964.

Amiriparian, S., T. Hübner, V. Karas, M. Gerczuk, S. Ottl, and B. W. Schuller (2022). "DeepSpectrumLite: A Power-Efficient Transfer Learning Framework for Embedded Speech and Audio Processing From Decentralized Data". In: *Frontiers in Artificial Intelligence* 5. DOI: 10.3389/frai.2022.856232. URL: https://www.frontiersin.org/article/10.3389/frai.2022.856232.

Mohamed, M. M., M. A. Nessiem, A. Batliner, C. Bergler, S. Hantke, M. Schmitt, A. Baird, A. Mallol-Ragolta, V. Karas, S. Amiriparian, and B. W. Schuller (2022). "Face mask recognition from audio: The MASC database and an overview on the mask challenge". In: *Pattern Recognition* 122, p. 108361. DOI: 10.1016/j.patcog.2021.108361. URL: https://www.sciencedirect.com/science/article/pii/S0031320321005410.

Song, M., A. Triantafyllopoulos, Z. Yang, H. Takeuchi, T. Nakamura, A. Kishi, T. Ishizawa, K. Yoshiuchi, X. Jing, V. Karas, Z. Zhao, K. Qian, B. Hu, B. W. Schuller, and Y. Yamamoto (2023). "Daily Mental Health Monitoring from Speech: A Real-World Japanese Dataset and Multitask Learning Analysis". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096884.

Schuller, B. W., S. Amiriparian, A. Batliner, A. Gebhard, M. Gerczuk, V. Karas, A. Kathan, L. Seizer, and J. Löchner (2023). "Computational charisma–A brick by brick blueprint for building charismatic artificial intelligence". In: *Frontiers in Computer Science* 5. URL: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2023.1135201.

Triantafyllopoulos, A., A. Kathan, A. Baird, L. Christ, A. Gebhard, M. Gerczuk, V. Karas, T. Hübner, X. Jing, S. Liu, A. Mallol-Ragolta, M. Milling, S. Ottl, A. Semertzidou, S. T. Rajamani, T. Yan, Z. Yang, J. Dineley, S. Amiriparian, K. D. Bartl-Pokorny, A. Batliner, F. B. Pokorny, and B. W. Schuller (2023). "HEAR4Health: a blueprint for making computer audition a staple of modern healthcare". In: *Frontiers in Digital Health* 5. URL: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2023.1196079.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| BN | Batch Normalisation |
| CCC | Concordance Correlation Coefficient |
| CMA | Cross-Modal Attention |
| CNN | Convolutional Neural Network |
| DANN | Domain Adversarial Neural Network |
| ECU | Electronic Control Unit |
| FFN | Feed-Forward Network |
| GELU | Gaussian Exponential Linear Unit |
| GRL | Gradient Reversal Layer |
| GRU | Gated Recurrent Unit |
| HMI | Human Machine Interaction |
| HMM | Hidden Markov Model |
| IAE | Implicit Auto Encoder |
| LN | Layer Normalisation |
| LSTM | Long Short-Term Memory Network |
| MAE | Mean Average Error |
| MHA | Multi-Head Attention |
| MSE | Mean Square Error |
| PE | Positional Encoding |
| RNN | Recurrent Neural Network |
| RELU | Rectified Linear Unit |
| SA | Self Attention |
| SELU | Scaled Exponential Linear Unit |
| TF | Transformer |
| UAR | Unweighted Average Recall |
| UI | User Interface |
| VB | Vocal Burst |

# List of Symbols

| | |
|---|---|
| $x_a$ | audio data or features |
| $x_v$ | visual data or features |
| $y$ | emotion labels |
| $z$ | encoded or latent features |
| $z_\alpha^i$ | query input to cross-modal attention |
| $z_\beta^i$ | key-value pair input to cross-modal attention |
| $d$ | data domain labels |
| $p$ | training progress |
| $lr$ | learning rate |
| $B$ | batch size |
| $L$ | sequence length |
| $D$ | feature dimension |
| $Q$ | query input for multi-head attention |
| $K$ | key input for multi-head attention |
| $V$ | value input for multi-head attention |
| $\mathcal{L}$ | loss function |
| $\mathcal{D}_s$ | source data domain |
| $\mathcal{D}_t$ | target data domain |
| $\mathcal{N}$ | normal distribution |
| $\mu$ | mean value |
| $\sigma$ | standard deviation |
| $\lambda$ | loss weighing parameter |
| $\gamma$ | domain adaptation rate speed parameter |
| $\phi$ | restrained uncertainty weighing loss hyperparameter |
| $\rho$ | Pearson's Correlation Coefficient |
| $\mathcal{T}$ | dynamic weight averaging loss temperature factor |
| $\mathcal{K}$ | dynamic weight averaging loss scaling factor |

# Chapter 1

# Introduction

What if a machine could understand you at least as well as a human observer?

We live in an age of ubiquitous sensors that are constantly generating data, and advances in computational power that let us leverage that data. Smartphones are just one example of a multi-sensor computing platform that we interact with in our daily lives, and they offer a multitude of applications. In the future, the number of highly sophisticated systems we use can only be expected to increase. While this development provides exciting new opportunities, challenges are also to be expected due to the growing number and complexity of features. Users might get overwhelmed or stressed by an overabundance of choices, or get frustrated when an intransparent system is not behaving as expected.

In these situations, an effortless and natural interaction is desirable. This interaction need not be explicit, in some cases, it might even be beneficial for a system to act discreetly and proactively, anticipate the needs of the users, reduce their cognitive load, generate a customised user experience, create a more pleasant environment or otherwise work towards their wellbeing.

For all of this to work, the machine needs to have some form of understanding of the person or persons it is faced with. There is an intuitive approach to solving this problem: Enable the machine to perceive cues that humans rely on for communicating meaning to each other.

Among the key factors in human interaction are emotions. We are quite skilled at detecting signs, from face and body posture to tone of voice and choice of words, and forming from them an impression of how someone is feeling. While this process happens automatically for us, it is quite challenging to replicate *in silico*, and to apply to an actual technical solution. However, as outlined above, the potential benefits of e.g. an empathetic voice assistant are enormous. A recent review on the topic can be found in Raamkumar and Y. Yang, 2023.

Thus, this thesis deals with the question of how machines can be provided with emotional intelligence. It focuses on the detection of emotions and related cognitive states.

The following sections serve as an overview to the thesis by explaining in more detail the motivation of the topic, stating the research questions and summing up the contributions of this work.

## 1.1 Motivation

The main motivation of this thesis stems from the unsolved question of how to design a system that can accurately detect states related to health and wellbeing, in particular emotions, in humans. The detection method should be unobtrusive and based on readily available sensors.

Such a system, as explained above, has the potential to lift human-machine interaction (HMI) to a new level. HMI is already moving away from pushing buttons towards more intuitive forms of communication, like gesture and voice commands, and facial recognition for authentication (Braun, Weber, and Alt, 2021; Tan et al., 2022). Likewise, machines now respond to commands in a more natural way, e. g., by assistants like Alexa or Siri synthesising voices to respond.

These advances in natural user interaction have been made possible by two factors: The availability of large amounts of data, and the increase in processing power, especially in the form of dedicated parallel processing units like GPUs and TPUs. Together, they make deep learning techniques viable, which allow machines to extract hidden meaning from data through multilayered neural networks. Deep learning has revolutionised many fields, with algorithms achieving near-human or even super-human performance on a wide range of tasks.

However, while deep learning has been used in recent years to advance digital health and automatic emotion recognition, those are not solved problems. For one, both health and wellbeing are inherently sensitive, since they rely on humans as subjects of analysis and reveal personal attributes about them. This raises issues of privacy in regards to the acquisition of the data and where it is processed, as well as who has access to the results. There is also the question of explainability. Deep neural networks are generally black boxes to human observers, which makes validating that they do what their designers intended difficult. These models may fail on edge cases or incorporate biases in the data into their decision-making.

Another complicating factor is that emotions and other cognitive states are hard to quantify and highly context dependent, requiring knowledge of psychological models during the data acquisition and annotation process. High-quality data can be expected to be scarce due to the cost of collecting it. This impacts model design since state of the art deep learning models rely on large datasets for training. When it comes to deploying a model into an actual use case, it is required to work well in a realistic setting, not just in a lab environment. The type of data encountered in such a setting provides an additional challenge, both in terms of noisy input signals and due to the fact that people in uncontrolled environments can be expected to show more nuanced and subtle behaviour.

There are many possible applications for systems that can automatically detect health and wellbeing. The research for this thesis was conducted in cooperation with the BMW Group. As vehicles become more sophisticated, autonomous, and connected, attention shifts from the exterior to the interior. Here, the focus is historically on the driver, with applications like driver monitoring for attentiveness, stress and fatigue detection (Koesdwiady et al., 2017; J. Wang, Chai, et al., 2022). These are primarily motivated by driving safety. However, there is a growing trend towards information and entertainment ("infotainment") integration in the car, which places more focus on the passengers (Murali, Kaboli, and Dahiya, 2022). Thus, in-cabin sensing has to expand its scope towards them. This can be done through cameras that monitor the

entire cabin, and microphone arrays capable of identifying speakers. In addition, pressure or capacitance sensors can be integrated into the seats, or smart wearables like Apple Watches can be wirelessly coupled to the car.

Vehicle manufacturers are following the trend towards more intuitive interaction methods (Breitschaft, Pastukhov, and Carbon, 2021; Tan et al., 2022). Gesture control and speech control are established features in premium cars, e. g., the Mercedes EQS and the 7 series BMW. The same sensors can be used to sense passenger states. The car interior can then adapt itself to enhance the wellbeing of its occupants, through different lighting, UI themes, sounds, and even scents. Vehicles are also commonly used by the same persons over extended periods of time. Thus, an intelligent system may even learn preferences of its users and anticipate their intent for seamless interaction and a comfortable and engaging experience.

However, the vehicle interior is also a challenging environment. Lighting conditions can change widely and rapidly, e. g., when driving under trees or into a tunnel. Depending on camera placement and head movements, only parts of faces may be visible. Audio may be noisy, or not present at all, when a solitary driver is travelling in silence. The car may be driven anywhere in the world, with passengers of any culture of ethnicity, speaking different languages. Onboard computing power is limited due to cost factors, and options to shift processing into the backend may be limited or impossibly depending on the use case, due to link latency, bandwidth restrictions or data protection regulations. It is therefore important to develop systems that can operate reliably and efficiently with noisy and diverse data.

## 1.2 Research Questions

This thesis seeks answers to four main research questions:

**RQ–1: How can continuous emotion be recognised on multi-modal data in the wild? In particular, which features, fusion strategies and temporal modelling schemes are beneficial?**

**RQ–2: How can emotions be recognised in a cross-cultural setting?**

**RQ–3: How can data with missing or incomplete labels be used to boost emotion recognition performance?**

**RQ–4: How can emotions be recognised when commonly used cues like facial expressions or speech are unavailable? Specifically, what methods can recognise affect from non-verbal vocalisations?**

Research question RQ–1 is motivated by the challenges of emotion recognition on realistic data, as well as the complexity of emotions themselves. As described in section 1.1, data collected in the real world can be expected to be noisy and diverse, hence it is beneficial to rely on multi-modal cues for complementary information (Poria, Cambria, et al., 2017). In this thesis, the focus will be on the visual and audio modalities, as they are readily available via ubiquitous cameras and microphones and allow for contactless, unobtrusive measurement. Furthermore, emotions are not fixed quantities, but dynamic concepts. They vary across time, hence this thesis will treat emotion recognition primarily as a sequence-to-sequence prediction problem. Emotions are also not treated as distinct categories that are mutually exclusive. Instead, this thesis focuses on modelling emotions as continuous signals that vary in intensity.

RQ–2 is motivated by the assumption that emotional displays and experiences vary by cultural setting, hence a model trained to recognise emotion on one culture will suffer a loss of performance on others. This thesis studies how emotion prediction models can adapt to data from different cultures.

Research question RQ–3 is motivated by the relationship between data and labels. While collecting rich and diverse data in large quantities is possible, e. g., from user-generated content online or from a fleet of vehicles, annotating that data in terms of emotional content would be prohibitively time-consuming and expensive. High-quality emotion labels require multiple human reviewers to examine the data (Kossaifi et al., 2019). Hence, it is desirable to develop models that can use not only the comparatively small amount of samples that will be annotated in terms of emotions, but also the much larger quantity that contains emotional displays or context information, without added labels. This leads to a representation learning problem for features that are discriminative of the emotion recognition task (Bengio, Courville, and Vincent, 2013).

Research question RQ–4 is motivated by the idea of analysing the audio modality as a rich source of emotional clues beyond speech. Affect may be expressed in the form of bursts of laughter, crying, grunts, etc. These non-verbal vocalisations may occur isolated or embedded into speech, their analysis forms a subset of the field of *Computational Paralinguistics (CP)* (Batliner, Hantke, and B. W. Schuller, 2020). For this thesis, multi-task recognition of the type and emotional content of isolated vocal bursts will be investigated.

## 1.3 Contribution

This work makes several main contributions:

**C–1:** A set of experiments aimed at discovering computationally efficient models to predict continuous emotions on a challenging, noisy in-the-wild emotion dataset is presented and the method's advantages and disadvantages discussed.

**C–2:** A Transformer-based approach for predicting emotions on a multi-cultural dataset of short non-verbal vocalisations is presented, and the results of the multi-task experiments are analysed.

**C–3:** An analysis on cross-cultural continuous emotion prediction on a multi-modal dataset of dyadic conversation videos is presented. Methods for leveraging additional unlabelled in- and out-of-domain data via semi-supervised adversarial learning are contrasted with a supervised baseline.

Contribution C–1 is targeted at RQ–1. It uses various uni-modal and multi-modal model architectures with different feature choices for predicting sequences of value-continuous affect. Its findings and parts of the resulting models are re-used in C–3.

Contribution C–2 mainly addresses RQ–4. Unlike the other contributions, it is based on a uni-modal (audio-only) dataset. Emotions are still modelled as value-continuous as in C–1, but here a much richer set of annotations is available, which are leveraged for multi-task learning. The dataset also includes four distinct cultures with specific annotations, hence this analysis also contributes to RQ–2. Furthermore, by leveraging a large model that was pre-trained in an unsupervised manner on tasks unrelated to emotion recognition, it also makes some contribution to RQ–3.

Contribution C–3 addresses RQ–1 by performing continuous emotion prediction on a multi-modal dataset. Its main focus however is on RQ–2, analysing the impact of cross-cultural domain shift on model performance. It also addresses RQ–3, by using data from cultures that do not have emotion labels, and by restricting the amount of labels available to the models for the cultures that are annotated.

The rest of this thesis is structured as follows:

In chapter 2, emotion theory that forms the background of this work is established, the field of affective computing that this thesis belongs to is introduced, and an overview on the state of the art in research and industry is given.

Chapter 3 describes the methods used in this thesis in terms of dataset properties, model architectures, training schemes and evaluation.

For each of the contributions, the corresponding experiments and results are presented in chapter 4.

In chapter 5, those experimental results are then discussed in turn and interpreted regarding the research questions of the thesis.

Opportunities for future work and an outlook on applications focused on next-generation vehicles is given in chapter 6.

Finally, chapter 7 concludes this work.

# Chapter 2

# Background and Theory

This chapter provides an overview on the theoretical foundations of the thesis. It examines psychological concepts of emotions, the theories behind them and their competing viewpoints. Then, it introduces the field of Affective Computing which combines elements of psychology and computer science. Current trends and challenges are summarised, and key aspects regarding data collection and processing are discussed. This is followed by an overview of the state of the art.

## 2.1 Emotion Models and Theories

In order to recognise and quantify emotions in a computational framework, it is first necessary to have a theory on what constitutes an emotion. However, despite extensive research into the topic, the nature of emotions remains a matter of open debate among psychologists. This section presents a number of influential concepts and theories.

### 2.1.1 Categorical Emotion Models

One of the most straightforward approaches to modelling emotions is to divide them into categories, with each category representing an identifiable emotion. This raises the question which emotional experiences should count as a category. A popular suggestion has been the proposal that there are a small number of fundamental emotional states. These are frequently referred to as "basic emotions" or the "common view of emotions" (Barrett, 2016).

There are several theoretical approaches to what makes an emotion a basic one. A popular line of argumentation is that emotions are biological in origin, and that each basic emotion is linked to a distinctive physiological display, explained to be evolutionary vestiges of behaviour once advantageous for survival. From this point of view, basic emotions are hardwired into the body and possess a unique signature from which they can be recognised.

In particular, early research in this direction focused on facial expressions. Tomkins, 1962 spoke of "affect programs" and considered the face as the primary location of affect display. Ekman and Friesen, 1971 investigated facial expressions that are displayed and recognised across different cultures, and identified six categories of basic emotions that have become known as the "Big 6", namely, happiness, sadness, fear, anger, disgust and surprise.

Which emotions are basic and how many of them there are remains undecided. Ekman and Cordaro, 2011 attempts to define the concept of basic by listing 13 criteria,

and proposes several additional emotions that may fulfil (most) of them, including positive emotions like wonder and excitement and negative emotions like guilt and shame. Izard, 2011 names interest, enjoyment/happiness/contentment, sadness, anger, disgust and fear as "first-order emotions", while being ambivalent on contempt. A first-order emotion is defined as requiring no higher cognitive functions and having a specific feeling component that is fixed by evolutionary legacy and cannot be learned. Levenson, 2011 considers enjoyment, anger, disgust, fear, surprise, sadness as basic emotions. An even more diverse picture appears when other theorists are included, as summarised in Ortony, 2021. The discrepancies between the lists of basic emotions proposed by different theorists, or even by the same theorist over the course of their career, point to a wider problem: There is still no unified concept of what constitutes an emotion, and how emotions differ from other cognitive states. In this light, Ortony, 2021 suggests that the attempts to classify emotions into basic and non-basic may be meaningless, and more fundamental issues should be addressed.

There have been extensive criticisms and defences on the subject of basic emotion theory. While some of these arguments concern the definition of being basic and which emotions do or do not qualify for it, the underlying assumptions of emotion categories and the methodologies used for justifying them have also been criticised.

The idea of the universality of basic emotions has been questioned due to the vagueness of the emotion concept and the language that describes it. Each language lexicalises a different set of emotions, so some feelings may not be translatable cross-culturally. One logical approach to choose candidates for basic emotions would be to start with salient concepts whose words appear frequently in the respective culture. Previous studies have usually selected candidates from a list of English words, with the assumption that these words represent universal concepts. However, this modelling approach risks incorporating biases of the researchers (Ortony, 2021).

### 2.1.2 Dimensional Affect Models

An alternative to thinking about emotional states in terms of categories is to model them in terms of affective dimensions, i.e. as a continuously varying vector in a space spanned by one or multiple axes. Each axis represents a constituent property of the affect. Two broadly identifiable dimensions are pleasure and activation, which describe the subjective wellbeing and mobilisation of energy in the experience, respectively (Russell and Barrett, 1999). These dimensions appear under various names in the literature, often as combinations of opposing terms denoting the ends of the axis, e.g. *pleasure-pain* and *activation-deactivation*. In this thesis, pleasure will be called valence and activation will be referred to as arousal.

If two dimensions are used, the resulting model is a circular space. Russell, 1980 introduced a *circumplex* of affect, in which affective experiences fall into different parts of the circle and the neutral state is at the center.

It has been argued that the circumplex model reflects the *core affect*, which is an ever-present affective state tied to the underlying neurophysiological state. Core affect is unspecific, not directed at any particular target. It captures non-emotional states like sleepiness, but it may miss differences between emotions, e.g. negative feelings like fear and anger that share similar core affect (Russell and Barrett, 1999).

In order to extend this model and better differentiate between emotions, additional affect dimensions have been proposed, such as dominance (Russell and Mehrabian, 1977). These dimensions add contextual information related to the causes, consequences and cognitive appraisal of emotions (Russell and Barrett, 1999).

### 2.1.3 Hybrid Models

It is also possible to model emotions as a hybrid form between categorical and dimensional variation.

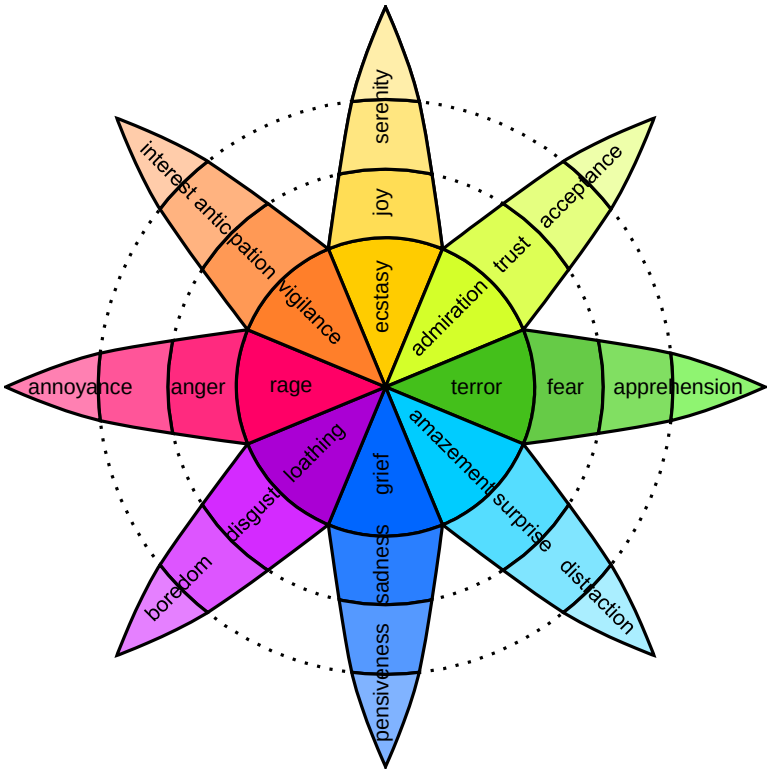



FIGURE 2.1: Plutchik's Wheel of emotions, illustrated as unwrapped cone surface. 8 primary emotions (joy, sadness, anger, fear, disgust, acceptance, anticipation and surprise) are arranged in opposing pairs and vary in intensity, increasing towards the centre (the base of the cone).

Plutchik and Kellerman, 1980 proposed a model that is frequently referred to as "Wheel of Emotions". It is based on 8 primary emotions, which are arranged as 4 pairs of polar opposites: joy-sadness, acceptance-disgust, surprise-anticipation, and fear-anger. Each primary emotion may vary in intensity, e.g joy going from serenity to ecstasy. In addition, Plutchik modelled secondary and tertiary emotions, such as love and guilt, to arise from combinations of the primary emotions. Plutchik's model is usually depicted as a 3D cone or as unwrapped cone surface flattened in 2D. The two-dimensional version is illustrated in fig. 2.1.

### 2.1.4 Semantic modelling of emotions

In order to overcome the limitations of previous approaches in modelling emotions, statistical methods have been proposed. The goal is to discover insights into the semantic space used to describe emotional experiences, and where those experiences fall within it. A. S. Cowen and Keltner, 2017 showed 2185 videos chosen to evoke 34 common emotional categories to test subjects and analysed their self-reported responses, which included both categorical ratings and free discussions. They found evidence of 27 variants of emotional experiences. While categorical labels were more useful in capturing the self-reported states than affective dimensions, the discovered categories were fuzzy, linked to each other by smooth variations in the gradients of affective dimensions.

Although the connection between the internal, subjective emotional state and the self-reported state may depend on additional factors, the computational approach provides interesting evidence that the emotional landscape is more complex than previously thought. A. S. Cowen and Keltner, 2017 note that the smooth gradients between categories cast doubt onto the assumption of prototypical emotion fingerprints proposed by basic emotion theory. However, this finding is more in agreement with the final theory that will be introduced here, the theory of constructed emotions.

### 2.1.5 Theory of constructed emotions

The theory of constructed emotions (Barrett, 2016) attempts to explain emotions based on recent developments in neuroscience. These include *neuron degeneracy* i. e., the finding that varying neuron populations contribute to a task, and the *predictive coding hypothesis*, where the brain creates simulations of external and internal sensations in order to efficiently allocate resources. The theory argues that emotional experiences are constructed from *concepts*, which are representations learned by the brain as it matches simulated and actual sensory input.

This framework has some interesting implications, both for understanding perception and for analysing emotions. Barrett et al., 2019 criticise traditional psychological concepts of basic emotion categories associated with specific brain regions and body cues as reductive and rooted in experiments that replicated biases. Instead, they argue that emotions are diverse and nuanced.

The emotion theories outlined above lead to the following premises for this thesis:

Emotions are highly complex and still not fully understood, they are diverse and context-dependent and should be modelled as continuous variables. Furthermore, models should be based on data that is collected in the wild and preferably multimodal.

The remainder of the thesis will show how theory is translated into computational practice.

## 2.2 Affective Computing

This section introduces the reader to the field of Affective Computing, which forms the technical background of this thesis. It touches upon the origins of this relatively new discipline, and presents applications, ongoing trends and challenges.

### 2.2.1 History

Affective computing is an evolving interdisciplinary field that combines elements of computer science, psychology, but also physiology and paralinguistics, among others. The term was coined in 1995 by Rosalind Picard, who defined it as "computing that relates to, arises from, and deliberately influences emotion" (R. W. Picard, 2010). In the two and a half decades since its inception, the field has developed rapidly, from initial scepticism towards being a mainstay in computer science conferences and having its own IEEE journal. It has been an early adopter of popular machine learning algorithms such as recurrent neural networks, and has now matured to a point where numerous companies are developing affective solutions for mainstream use (B. W. Schuller, 2018; B. W. Schuller, R. Picard, et al., 2021). Several application scenarios for this technology will be presented in the next subsection.

### 2.2.2 Applications

Affective Computing has a wide range of applications. It has the potential to transform HMI, as the machine becomes capable of reacting more adequately to the human's current emotional state. It could then either mirror the feelings it detects, which would be desirable for a cheerful and engaging conversation with a voice assistant, or, in the case of negative emotions, modify its behaviour to guide a person towards a more positive state.

Aside from improving the user experience in HMI, affective computing also has potential for beneficial use in a healthcare and wellbeing setting, by observing people and detecting the onset of depression or early signs of burnout, or assisting in the diagnosis of autism in children (B. W. Schuller, R. Picard, et al., 2021).

Affective computing can also be applied in a business analytics setting, helping companies improve their business models. For instance, it can be used in call centres to rate the emotional state of a caller and offer them better service. It could also be used to monitor conference calls and provide feedback to the parties afterwards. Furthermore, it can be applied to product reviews of customers or professional reviewers, allowing a manufacturer to gather valuable data on the emotional impact of their design choices and identify issues.

From an automotive perspective, affective computing can be applied in numerous ways both inside and outside the vehicle. Driver monitoring is already used to detect sleepiness and distraction from eye movements, and it can be extended towards negative states like stress and aggression (Zepf et al., 2020). Mitigating negative feelings in the driver helps prevent road rage and increases driving safety (Jeon, 2015). In addition, vehicles are moving towards a more natural, intuitive interaction with their passengers. Voice-based control of functions like making calls and entering destinations into a navigation system are already common. Future emotion aware vehicle assistants that have a holistic understanding of the passengers' states and the vehicle cabin could adjust the user experience to improve wellbeing (Vögel et al., 2018). This could be achieved by dynamically changing display UIs, mood-dependent interior lighting, the voice of the assistant sounding more empathetic, and adjustment of the conversation flow to avoid stress and irritation (Karas, D. M. Schuller, and B. W. Schuller, 2024).

### 2.2.3 Challenges

Despite strong recent advances, challenges remain in the field of affective computing. This subsection gives an overview of the major issues.

One significant challenge lies in developing systems for use "in the wild", i. e., systems that generalise well to a wide range of scenarios that can occur in a real-life application. As discussed in section 2.1, naturalistic emotional displays can be nuanced and varied, making an accurate estimation of the underlying emotional state difficult. The context in which these emotions occur can also vary significantly. In addition, the signals captured by sensors in real life may be of lower quality, noisy, or even missing occasionally, compared to recordings from a controlled environment. All these factors make it difficult to design and validate affective systems that can transition from a lab into nature.

A major factor in driving affective computing forward has been the adoption of deep learning methods (B. W. Schuller, 2018), as this has enabled the development of more powerful models. However, deep learning comes with its own set of issues. The first is the need to collect large amounts of data for training, as the model will need to see enough samples to capture the underlying distribution. If training and test data distributions differ, performance may decrease (Pan and Q. Yang, 2010). A related issue of large datasets is the rising cost of annotation, especially if the label is hard to obtain, as is the case with emotions. Furthermore, the larger models used for deep learning may require significant computational resources for both training and inference. Constraints of the end user hardware may require optimisation or choosing models with a smaller footprint.

Furthermore, deep learning presents issues in terms of interpretability and transparency. Algorithms based on deep networks are usually inscrutable "black boxes" whose decisions cannot be understood by humans. This is particularly worrisome for systems that observe human physiology or behaviour, as is the case with affective computing. Thus, there is a drive towards *explainable AI (XAI)*, see section 6.1.4.

On a related note, affective computing also raises issues of privacy, given that it uses personal data. It is essential to protect the data of people analysed by this technology, and prevent it from being shared with other parties without their consent. This can be done by either keeping the algorithms running on a local device and not sharing data with a backend server, or by anonymising data, see section 6.1.1, section 6.1.2.

However, the personal nature of affective computing also raises the opportunity for adaptive, personalised systems that learn the preferences of a user, by running for extended periods of time on a device that the user regularly interacts with, e.g. a smartphone or a car. Given that emotions are inherently subjective, this type of longitudinal learning may help improve the performance of affective computing significantly.

## 2.3 Signal Modalities

This section introduces commonly used signal modalities in Affective Computing. Of these, audio, visual and text account for the majority in the literature (Poria, Cambria, et al., 2017). Audio and visual features form the backbone for the experiments conducted in this thesis. Textual and physiological signals are out of scope, but are also presented here for completeness.

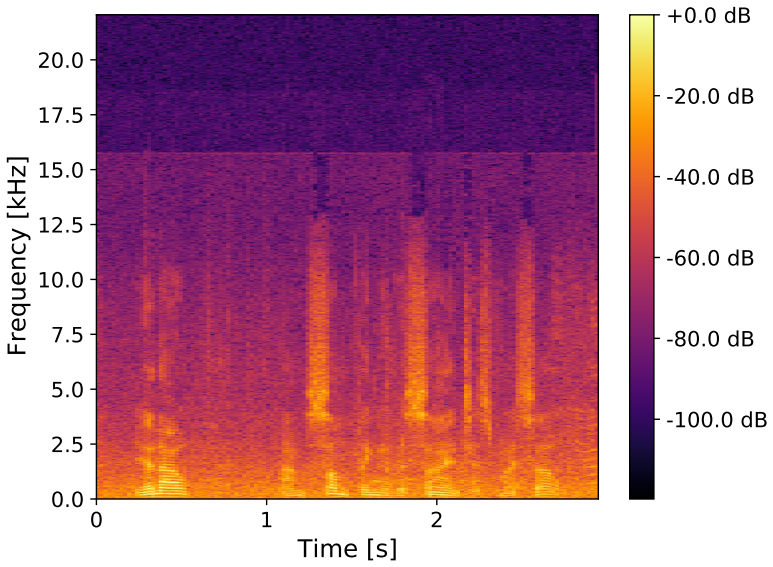



FIGURE 2.2: Example visualisation of a short clip of human speech as spectrogram, with power converted to decibel and linear frequency scaling.

### 2.3.1 Audio

Audio analysis for the purpose of affective computing primarily focuses on the voice as carrier of emotional meaning, although some works also attempt to infer emotional states from music or other sound sources.

The choice of appropriate features to extract from raw audio is an ongoing research topic in the field of *Speech Emotion Recognition (SER)*. Features should be robust against noise and language or cultural differences (B. W. Schuller, 2018). A number of popular features is described here.

The *Mel Spectrogram* is obtained from the audio signal by first computing the power spectrogram (see fig. 2.2) and then multiplying it with a Mel-frequency filter bank. In order to obtain the power spectrum, a *Short-Term Fourier transform (STFT)* is applied to the audio signal and the amplitude of the complex result is squared. In addition, the power may be converted to decibels and lower levels cut off to reduce noise. While the STFT is linearly scaled, the Mel scale is based on human hearing capability, which is more sensitive to changes in pitch at lower frequencies. A Mel filter bank is commonly computed as a series of overlapping triangular band-pass filters spaced linearly across the desired frequency range. As the frequency increases, filters become lower and wider.

*Mel-frequency cepstral coefficients (MFCCs)* are a set of short-term descriptors that encode the vocal timbre (Eyben, Scherer, et al., 2016). They are obtained from the logarithmic Mel spectrogram by applying a discrete cosine transform. MFCCs and other low-level descriptors form the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), a standardised set of 88 parameters selected by Eyben, Scherer, et al., 2016. The much larger Computational Paralinguistics Challenge (ComParE) feature set, named after the long-running challenge of the same name (B. W. Schuller, Batliner, et al., 2021), contains 6373 features. These handcrafted feature sets can be extracted with the openSMILE toolkit [1]. In SER, low-level features like MFCCs are

[1] https://www.audeering.com/research/opensmile/

frequently used, which are then processed with encoder networks such as 1D-CNNs (Chatterjee et al., 2021).

There is a trend towards deep feature extraction from audio. The AUDEEP toolkit [2] (Freitag et al., 2018) uses spectrograms as input for recurrent autoencoders, whole latent representations are then used as features. DEEPSPECTRUM [3] (Amiriparian, Gerczuk, et al., 2017) instead treats spectrograms as a computer vision problem by processing them with pre-trained 2D-CNNs. This approach has been proven effective across a range of audio analysis tasks (Amiriparian, Hübner, et al., 2022).

While hand-crafted features and pre-trained models feature extractors have their merits, some works have instead opted for learning directly from the audio using end-to-end learning. Shallow 1D-CNNs followed by LSTM layers are a popular architecture for end-to-end continuous emotion recognition (Trigeorgis et al., 2016), (Tzirakis, J. Zhang, and B. W. Schuller, 2018), and end-to-end emotion classification (Rizos et al., 2020). Shallow 2D-CNNs are also shown to be useful for learning local features on spectrograms (J. Zhao, Mao, and L. Chen, 2019).

Unsupervised learning has also been employed for SER, e. g., through fuzzy clustering (Rovetta et al., 2020). More recently, the Transformer models popular in NLP have been applied to audio analysis with architectures like WAV2VEC2 (Baevski et al., 2020), including for SER (Wagner et al., 2023).

In terms of dimensional emotion recognition, the voice has been shown in the literature to be particularly effective in determining the arousal dimension, with lower performance for valence (Srinivasan and Martinez, 2018).

### 2.3.2 Visual

In the visual domain, analysis of affect is mostly based on the face, although some works also investigate body posture (K. Wang et al., 2018).

*Facial Action Units (FAUs)* are descriptors of facial muscle activity defined in the *Facial Action Coding System (FACS)* pioneered by Ekman and Friesen, 1978 and revised in 2002 by Ekman, Friesen, and Hager, 2002. The underlying idea is that any facial configuration can be described by a combination of muscle contractions stretching the skin above them, and a trained human can then annotate images accordingly.

The trend towards deep learning is also present in the vision domain, with many works extracting features via 2D-CNNs, including popular computer vision architectures e. g., ResNet (He et al., 2016), Inception (Szegedy et al., 2017) or squeeze-and-excitation networks (SENets) (Hu et al., 2020). In order to compensate for the relatively small dataset sizes in affective computing, transfer learning is used, with the CNNs usually being pre-trained for facial recognition on datasets like VGGFace2 (Cao et al., 2018). Recently, vision Transformers (Dosovitskiy et al., 2020) have also received attention (Roka and Rawat, 2023; Panlima and Sukvichai, 2023).

Information extracted from the face has been shown to be particularly useful for predicting the valence dimension (Poria, Cambria, et al., 2017; Ringeval, B. Schuller, Valstar, Cummins, Cowie, and Pantic, 2019).

[2] https://github.com/auDeep/auDeep
[3] https://github.com/DeepSpectrum/DeepSpectrum

### 2.3.3 Text

Linguistic features capture semantic and syntactic content of language and are obtained from tokenised text. Natural language processing (NLP) mainly relies on deep embeddings generated by Transformer style models. In particular, Bilingual Encoder Representations from Transformers (BERT ) (Devlin et al., 2019) caused major breakthroughs in NLP. It is based on unsupervised pre-training with two tasks, namely, prediction of masked tokens (*Cloze task*) and of the next sentence. The resulting network can be fine-tuned towards the desired downstream task. Recent advances in AI have been driven by large language models (LLMs) e. g., ChatGPT.

For dimensional emotion recognition, text is particularly effective at determining the valence dimension (Wagner et al., 2023).

### 2.3.4 Physiological

Physiological signals are particularly interesting for applications in healthcare and wellbeing, such as stress detection from elevated blood pressure. The need for easy, low-cost sensing has inspired research into contact-less methods. For instance, to record cardiac activity, the passing of the *blood volume pulse (BVP)* across the body can be measured visually. *Photoplethysmography (PPG)* aims to detect changes in the reflective properties of tissue due to the BVP. Visible, green wavelengths (500-600 nm) are particularly useful for this method, due to a combination of skin penetration depth and the optical characteristics of haemoglobin (Allen, 2007). Similarly, *blood oxygen saturation ($SpO_2$)* can be measured by RGB cameras due different light absorption of oxygenated and deoxygenated blood (McDuff, 2023). *Breathing rate* is another physiological signal that can be captured without contact by detecting micro-motions via camera or radar.

However, since specialised sensors are required to capture many physiological signals with high accuracy, data is more difficult to collect compared to audio and video. Several commonly used sensing methods are briefly described below.

#### Muscular Activity

The cardiac cycle can reveal a considerable amount of information, e. g., stress level (Lanatà et al., 2015) or underlying medical conditions. A common measurement method is *Electrocardiography (ECG)*, which measures the electrical activity of the heart muscle via electrodes on the skin. The ECG results in a series of characteristic peaks, from which the *heart rate (HR)* can be derived. Changes in heart rate over time can be expressed as *heart rate variability (HRV)*. Similar to the ECG, the excitation of other muscles in the body can be measured via *Electromyography (EMG)*. Signals are obtained with skin-mounted electrodes or needles (for greater accuracy). EMG has been used to estimate the driver state (Paredes, Ordonez, et al., 2018; Lv et al., 2021).

#### Central Nervous System Activity

Electrical activity inside the brain can be measured via *Electroencephalography (EEG)*, which places multiple electrodes around the subject's head. The electrodes can be wet for improved conductivity, but dry electrode headsets have been proposed for greater comfort while still providing sufficient signal quality (Hinrichs et al., 2020). EEG is a popular signal in research on intelligent vehicles, used for estimating drowsiness (Zhu et al., 2021), attention (Guo et al., 2018), stress (Kim et al., 2022),

personal preferences (Ling et al., 2021) and emotions (C. Park, Shahrdar, and Nojoumian, 2018).

#### Electrodermal Activity

Through electrodes that measure changes in skin conductance, the *Galvanic Skin Response (GSR)* can be determined. GSR is a popular signal in affective computing for estimating stress (Stappen, Meßner, et al., 2021).

## 2.4 Multi-modal Fusion

Next, popular schemes to combine the signals described above will be introduced. Multi-modal fusion has been shown to be useful for improving the predictive power of a model, and unsurprisingly, there is a trend in affective computing towards multimodality (Poria, Cambria, et al., 2017).

### 2.4.1 Early Fusion

*Early fusion*, also commonly referred to as *feature level fusion*, combines the features extracted from various signal modalities into a feature vector and trains the model on it. It is advantageous in that correlations between modalities can be used at an early stage, which can improve task performance, but disadvantageous in that all features have to be present in a synchronised format (Poria, Cambria, et al., 2017). In addition, concatenation of features may lead to very high-dimensional vectors, causing curse of dimensionality issues in training (H. Chen, Jiang, and Sahli, 2020).

### 2.4.2 Late Fusion

In *late* or *decision-level fusion*, each feature is used to train a separate classifier, and their decisions are then combined into the final score, e. g., by majority voting, weighted sum or weighted product. Advantages of late fusion are simple implementation, the ability to optimise each modality separately, and robustness against missing signals. However, there are also disadvantages. Learning different classifiers for each modality may become time-consuming (Poria, Cambria, et al., 2017), even more so if multiple feature sets are extracted from a modality. In addition, late fusion only combines the multi-modal information at the decision stage, thus is cannot exploit cross-modal correlations directly like early fusion does.

### 2.4.3 Hybrid Fusion

*Hybrid fusion* attempts to combine the advantages of both feature-level and decision-level fusion. It achieves this by concatenating features at the input level to make use of cross-correlations and also combining the scores of individual classifiers at the decision stage. *Model-level fusion* is a term referring to fusion techniques that make use of correlations between modalities, where information is exchanged at multiple levels within the model (Poria, Cambria, et al., 2017; Han et al., 2021). This definition covers a wide range of algorithms, such probabilistic combination through Bayesian inference and Hidden Markov Models (HMMs), explicitly modelling the correlations as in tensor fusion (Zadeh et al., 2017) and cross-modal attention modules (Tsai et al., 2019).

Now that the signal modalities and methods to combine them have been established, the next section will cover the topic of datasets.

## 2.5 Datasets

High-quality datasets are the prerequisite for model development. The popularity of deep learning has driven the demand for increasingly large corpora. This section introduces strategies for data collection and annotation in affective computing, while listing some datasets as examples.

### 2.5.1 Data Collection

Data collection refers to the process of gathering recordings of subjects, either under controlled conditions in a laboratory or under varying conditions in the wild. This data needs to contain the desired emotional behaviour. Various solutions have been proposed, which can be grouped into acted, induced and spontaneous displays of emotion (Kossaifi et al., 2019).

#### Acted

A straightforward approach to recording emotions is to instruct subjects to act, i. e., to perform a script with different emotions or varying intensities of the same emotion. An example of an acted dataset in a lab setting is *IEMOCAP* (Busso et al., 2008). It contains utterances annotated with arousal, valence and categorical emotions.

Acted data can also be collected from movies or TV shows, e. g., the *MELD* dataset (Poria, Hazarika, et al., 2018) based on *Friends*. MELD is composed of dialogue clips from the show, annotated with categorical emotions.

Yet another option is to let the actors record themselves in the wild for increased data diversity, as in the HUME-VB dataset (Baird, Tzirakis, Gidel, et al., 2022), which contains audio bursts annotated with continuous emotions.

A disadvantage of the acted approach is that emotions expressed in real life situations may be more subtle and context-dependent, see 2.1. Thus, it has been argued that an acted emotion will always be missing something, and a realistic database should capture genuine emotional behaviour (Martin et al., 2006). This leads to the idea of induced datasets.

#### Induced

Induced datasets will usually include some form of elicitation protocol to get the subjects into the desired emotional state. Elicitation can either be done through an external stimulus like a piece of music or a short film, or by asking the study participant to recollect an emotional experience from their past. Once the subjects are thus primed for emotional display, the recording begins. In addition, subjects can be given specific tasks that are designed to elicit emotional responses. Another option is to have an emotional stimulus happen during the recording through a conversation agent or unforeseen events, such as UI malfunctions or bugs.

Examples of induced datasets are *enterface* (Martin et al., 2006), which used short stories for elicitation, and *SEMAINE* (McKeown et al., 2012), which relied on conversational agents controlled by a Wizard of Oz. enterface used categorical annotation

for 6 emotions, while SEMAINE is annotated continuously in 5 affect dimensions (valence, activation, power, expectation and intensity).

#### Spontaneous

Spontaneous human interactions can be considered the most desirable type of data to record for affective computing, since they can be expected to include rich displays of naturalistic emotions (Kossaifi et al., 2019). However, by its nature this type of data is challenging to acquire in the wild.

Examples of spontaneous datasets include *SEWA* (Kossaifi et al., 2019) and *RECOLA* (Ringeval, Sonderegger, et al., 2013), both of which contain free conversations on subjects designed to elicit emotion (advertisements and survival in a hostile environment, respectively). Both are continuously annotated with arousal and valence.

An alternative approach to setting up spontaneous interactions as part of a study is to gather data online. Scraping videos from platforms like YouTube is a viable strategy to gather large datasets, given the proliferation of user-generated content. However, it faces the issue that motivated acted or induced collection, i. e., the need to ensure that the data contains emotional behaviour. A possible mitigation strategy is to filter content based on tags. Spontaneous emotional displays may be found in reaction videos, while reviews will contain sentiment, even if they are mostly scripted. An example of a web dataset containing spontaneous reactions and vlogs is *Aff-Wild* (Kollias, Tzirakis, Nicolaou, et al., 2019) and its successor *Aff-Wild2* (Kollias and Zafeiriou, 2019). It is annotated with both categorical and continuous emotions. *MuSe-Car* (Stappen, Baird, et al., 2021) is a dataset of car reviews taken from YouTube and annotated continuously with sentiment and trustworthiness scores.

For this thesis, datasets that are collected in the wild are highly relevant. Regarding the method of emotion production, spontaneous datasets are preferred. Value-continuous emotion labels are required, and time-continuous annotation is preferred. The following datasets are chosen for experimentation:

- SEWA
- RECOLA
- Aff-Wild2
- Hume-VB

The chosen datasets and their properties will be presented in greater detail in chapter 3. The following section gives an overview of techniques used for deriving the emotion labels of an affective database.

### 2.5.2 Annotation

Annotation is the process of assigning a set of labels to the data for training models and validating their performance. This section describes the challenges of annotating data with affective labels, as well as common methods and tools for addressing those issues. The focus is on continuous annotation of audiovisual data, but most aspects described here can be transferred to other modalities.

Data with a temporal component such as audio, video or physiological traces can be annotated either in a time-continuous manner, assigning a label at fixed intervals, e. g., on a frame-by-frame level, or on a clip level, summing up short segments with a

single label. The labels can also be discrete, or value-continuous e. g., in terms of valence and arousal. Time-continuous annotation has the advantage of capturing the evolution of feelings, while value-continuous annotation can describe mixed emotional states. Thus, continuous annotation is capable of delivering a rich perspective on the affective content of the data.

However, given the subjective and internal nature of emotions, there is no fully objective ground truth (Barrett et al., 2019). Instead, a popular substitute is to define a *gold standard* based on emotions perceived by human raters (B. W. Schuller, 2018).

Since human perception is also subjective, it is common practice to have multiple raters annotate the data independently and then find the gold standard as a consensus between the raters. Having access to more raters can improve quality but comes at increased cost. Furthermore, combining the ratings presents several issues. First, the raters' reaction times may vary. Second, raters have subjective emotional scales, and will therefore vary in terms of absolute values, even if they agree in relative changes (Metallinou and Narayanan, 2013). Cultural biases also impact the labelling process. Ideally, annotators should have the same cultural background as the subjects they are rating, as this will help them to perceive subtle changes in emotion (Kossaifi et al., 2019). Third, there may be inconsistencies between ratings.

Thus, creating high-quality and fine-grained annotations for large datasets is complicated and requires a sophisticated process. Specialised protocols and equipment have been developed for this purpose. For instance, Stappen, Baird, et al., 2021 defined three roles: An administrator who coordinates and assigns tasks, an auditor who checks that protocol is being followed and checks the results delivered by annotators, and the annotators who examine the data. Annotators receive training before working on the dataset.

The task of annotating sequential data over time imposes a higher mental workload on the rater compared to assigning a single value (Metallinou and Narayanan, 2013). Therefore, tools are needed that allow the rater to easily review the data and enter their estimation. Over the years, a number of such tools have been developed, including *Gtrace* (Cowie et al., 2013), *CARMA* (Girard, 2014) and *DARMA* (Girard and C. Wright, 2018). The UI of those tools includes a replay function for the data and a way to continuously record a stimulus from the rater. Joysticks have become a popular sensing solution since they provide an economic handling with proprioceptive feedback and automatic centring when no input is applied (Sharma et al., 2020). In order to further reduce mental workload on the raters, they are often instructed to focus on a single affect dimension, e. g., arousal. This 1D approach has the drawback of costing additional time due to multiple viewings required. In addition, reporting multiple dimensions simultaneously may provide a more complete picture. 2D acquisitions have recently become more popular (Sharma et al., 2020).

While most tools are designed for desktop use, solutions for in-the-wild annotation on end devices are also being developed. This would have the advantage of allowing a potentially large number of raters to easily record their emotional experience of content anywhere. The disadvantage is that the uncontrolled context may bleed into the ratings. One example of such a solution is T. Zhang et al., 2020, who developed a mobile interface for viewing videos and annotating valence and arousal by touching a transparent overlay.

Once the raters' annotations have been collected, they need to be combined. Since a simple averaging of the traces would be ineffective for the reasons outlined above, a

number of more sophisticated methods have been developed. Time shifts between reviewers can be addressed with e. g., Dynamic Time Warping (DTW) or Probabilistic Canonical Correlation Analysis (PCCA). The agreement between raters is well captured by correlation metrics such as CCC, Pearson's correlation coefficient and Cronberg's $\alpha$. Thus, the final labels can be derived from evaluations which show a sufficient amount of inter-rater consistency (Metallinou and Narayanan, 2013). Algorithms like DTW, PCCA or Canonical Time Warping (CTW) can be used to derive a subspace that maximises correlation between the raters (Kossaifi et al., 2019), improving the gold standard.

Annotating the large databases required for state of the art deep learning with emotion labels from multiple trained human annotators incurs considerable expenses. Thus, there have been attempts to develop methods that reduce costs. One option is to use crowd-sourcing, however, this may negatively impact the quality of the labels (Mollahosseini, Hasani, and Mahoor, 2019). Additional quality controls through trustability scores in combination with active learning can help with reaching a consensus between the raters (Hantke et al., 2018). The vast amount of available unlabelled data has also motivated research into semi-automatic annotation techniques (Canales et al., 2022). Part of the data is annotated regularly, and a model is developed to estimate the emotional content of the rest. Those estimations can then be reviewed and corrected by humans, potentially saving time while maintaining high quality.

## 2.6 State of the Art

Having introduced the field of affective computing in section 2.2, signal processing in section 2.3 fusion in section 2.4, and datasets in section 2.5, the purpose of this section is to conclude the chapter by giving the reader an overview of machine learning competitions relevant in the context of this thesis, as well as industrial applications of affective computing.

### 2.6.1 Competitions

In this sub-section, a number machine learning competitions in the field of emotion recognition are listed. The selection is based on the importance the competitions had in the context of the thesis, by virtue of dealing with continuous emotion recognition and using relevant datasets.

#### AVEC

The 9th Audio/Visual Emotion Challenge and Workshop (AVEC 2019) "State-of-Mind, Detecting Depression with AI, and Cross-cultural Affect Recognition" focused on detecting valence, arousal and liking in its Cross-Cultural Emotion Subchallenge (CES) (Ringeval, B. Schuller, Valstar, Cummins, Cowie, Tavabi, et al., 2019). It was based on a subset of the SEWA database. Baselines were computed with hand-crafted (FAU, eGeMAPS), bags-of-words, and deep (DeepSpectrum, VGG-16, ResNet-50) features used to train a recurrent network, and then combined in a late fusion approach. CCC was used as the metric.

7 teams participated in AVEC 2019. The winning paper, (J. Zhao, Li, et al., 2019), used an unsupervised adversarial domain adaptation approach, to account for cultural differences. The runner up, (H. Chen, Y. Deng, Cheng, et al., 2019), used a

combination of pre-trained 2D CNN and a 1D CNN to extract spatial-temporal features from audiovisual data.

AVEC is considered highly relevant due to dealing with emotion recognition across several cultures, on a subset of a dataset that contains spontaneous emotion interactions in the wild.

#### ABAW

The Affective Behavior Analysis in the Wild (ABAW) Competition was first introduced in 2020 as part of FG'20 (Kollias, Schulc, et al., 2020). It is based on the Aff-Wild2 dataset. The organisers presented three sub-challenges: valence-arousal estimation, basic expression classification, and action unit prediction. The evaluation metrics were CCC, weighted average of F1 and accuracy, and unweighted average F1 score per AU, respectively. Baselines were based on PatchGAN (for dimensional affect) and MobileNetV2 (for expression and AUs). The winning team of the valence-arousal challenge D. Deng, Z. Chen, and Shi, 2020 used a student-teacher approach, which allowed training on partially labelled videos.

In the ABAW2 challenge (Kollias and Zafeiriou, 2021), held in conjunction with ICCV'21, the Aff-Wild2 dataset was re-used, with additional annotations. VGG-FACE was used as the backbone for the baselines. 40 teams competed for the Valence-Arousal (VA) challenge, 55 for the Basic Expression (EXPR) Challenge, and 51 for the AU challenge. The winning team of the VA challenge D. Deng, Wu, and Shi, 2021 used a multitask learning approach that leveraged uncertainty estimation on an ensemble of models.

The ABAW3 Challenge (Kollias and Zafeiriou, 2021) held with CVPR'22, added more videos to the dataset and expanded the AU sub-challenge from 8 to 12 AUs. It also introduced a fourth sub-challenge aimed at multi-task prediction on a static subset of Aff-Wild2. The evaluation metric was the sum of the metrics used in the three individual tasks. The organiser baselines employed ResNet50 (for valence-arousal) and VGG16 pre-trained on VGGFace (for the other tasks). The winner of the valence-arousal sub-challenge Meng et al., 2022 used a multi-modal ensemble approach with temporal encoding.

The fourth ABAW competition (Kollias, 2023) was held with ECCV'22 and contained a multi-task learning challenge on Aff-Wild2.

The fifth ABAW challange (Kollias, Tzirakis, Baird, et al., 2023), held with CVPR'23, repeated the affect, expression and AU sub-challenges and introduced at fourth, Emotion Reaction Intensity Estimation, on the HUME-REACTION dataset. The metric for this new task was the average Pearson correlation coefficient across 7 emotional intensities.

The sixth ABAW competition (Kollias, Tzirakis, A. Cowen, et al., 2024) returned to CVPR'24 with the affect, expression and AU sub-challenges, as well as new challenges on compound emotion and emotional mimicry.

The relevance of ABAW for this thesis is due to its focus on the large and challenging Aff-Wild2 dataset.

#### ExVo

The ICML Expressive Vocalizations (ExVo) Workshop & Competition (Baird, Tzirakis, Gidel, et al., 2022), held in 2022, dealt with the recognition, generation and personalisation of vocal bursts, the first challenge of its kind to address this topic. Vocal bursts can convey a wide range of emotions, e. g., gasps, growls or groans may indicate surprise, anger, or sadness. The challenge was based on the HUME-VB dataset. It offered three sub-challenges, which were: EXVO-MULTITASK, for the joint recognition of emotional intensities and speaker demographics (age and native country); EXVO-GENERATE, for the generation of bursts coloured in each of the 10 different emotions; and EXVO-FEWSHOT, for leveraging the speaker identity to recognise emotions from a small number of samples. The baseline for EXVO-MULTITASK was a fully connected network processing acoustic features generated with OPENSMILE, OPENXBOW, and DEEPSPECTRUM. The metric for the sub-challenge was a harmonic mean of MAE for age, UAR for native country, and average CCC for emotions. In EXVO-GENERATE, MSG-GAN (Karnewar and O. Wang, 2020) generated baseline spectrograms for each emotion class, and *Fréchet Inception Distance (FID)* served as metric. Finally, END2YOU was used to train an end-to-end 1DCNN-LSTM for EXVO-FEWSHOT, and mean CCC of the emotions was the metric.

The relevance of ExVo in the context of this thesis was mainly as a stepping stone towards the next competition, A-VB.

#### A-VB

The 2022 ACII Affective Vocal Burst Workshop & Competition (A-VB) (Baird, Tzirakis, Brooks, et al., 2022) introduced the first iteration of the A-VB challenge, which is based on the large-scale, in-the-wild Hume-VB dataset. Similar to EXVO, the data is again comprised of vocal bursts. Participants were invited to any or all of 4 sub-challenges, including: A-VB-HIGH, a multi-label regression task to predict the intensities of 10 emotions; A-VB-TWO, to predict the affect dimensions valence and arousal; A-VB-CULTURE, requiring the prediction of culture-specific emotions; and finally, A-VB-TYPE, to classify the vocal bursts into one of 8 types, e. g., laughter. The metric for A-VB-TYPE was UAR, for all other tasks it was the mean CCC of the respective emotions. Baselines were provided following two approaches: In the feature-based approach, COMPARE and EGEMAPS features were extracted with OPENSMILE and processed with a 3 layer fully connected network. The end-to-end approach used END2YOU to train a shallow 1DCNN-LSTM network on raw audio.

A-VB is highly relevant for this thesis since it deals with a vocal burst dataset annotated with multiple continuous emotions.

### 2.6.2 Industrial Applications of Affective Computing

Affective computing has matured greatly since the field's inception, and is now close to being applicable at scale (B. W. Schuller, R. Picard, et al., 2021). It is therefore unsurprising that many companies are developing commercial solutions. Due to the specialised nature of the problem, many of these companies, e. g., Affectiva, Noldus and iMotions act as suppliers, licensing or selling their products to academia or OEMs. They collect their own, non-public datasets and develop algorithms on them,

which are either provided directly to the customer, e. g., as part of the software running on an automotive ECU, or integrated into a platform that the client can access. The latter is particularly useful for data analytics, e. g., in call centres gauging customer satisfaction, marketing agencies planning advertisement campaigns, game developers wanting to improve user experience, or behavioural researchers designing and evaluating studies.

#### Visual

Given the proliferation of computer vision and facial recognition, there are numerous products for facial expression analysis. Frequently, the affective analysis is one component of a platform solution that also detects additional features, e. g., age and gender, from the face. A selection of companies and their products is given below.

Noldus[4] is a behavioural analytics company. Its *FaceReader* software claims robust face detection and classification of the 6 basic emotions, as well as neutral state and contempt. It also provides estimates for valence and arousal, head pose and action units, age and gender.

RealEyes[5] develops computer vision based solutions for measuring attention and engagement with media content. Its emotion algorithm makes use of facial action units and detects happiness, surprise, disgust, confusion, fear, empathy, and contempt, as well as engagement, negativity and valence.

Kairos[6] offers a computer vision platform with APIs for facial recognition and facial analysis, including emotion recognition.

Affectiva[7] is an affective computing company that spun out of the MIT media lab where the field originated, see section 2.2. It provides solutions for automotive and media analytics, as well as social and behavioural research. Affectiva was acquired by the Swedish Smart Eye group in 2021.

#### Audio

Audio based affective computing products have applications ranging from customer and business analytics, through empathetic voice assistants, to psychological and medical research.

Amazon has integrated different emotions and speaking styles into its Alexa voice assistant[8]. This is intended to make conversation more natural. Available emotions include excitement and disappointment, which can be displayed at different intensities.

Audeering[9] is an audio AI company that focuses on voice analytics. Its products can detect a variety of speaker states and traits, including emotions, sentiment (via text), age and gender. Additionally, health monitoring is possible by detecting the impact of physiological effects of diseases, e. g., Covid-19, on the vocal tract. While some services are offered as a web API, others focus on efficient real-time processing on

[4] https://www.noldus.com/
[5] https://www.realeyesit.com/
[6] https://www.kairos.com/
[7] https://www.affectiva.com/
[8] https://developer.amazon.com/en-US/blogs/alexa/alexa-skills-kit/2019/11/new-alexa-emotions-and-speaking-styles
[9] https://www.audeering.com/

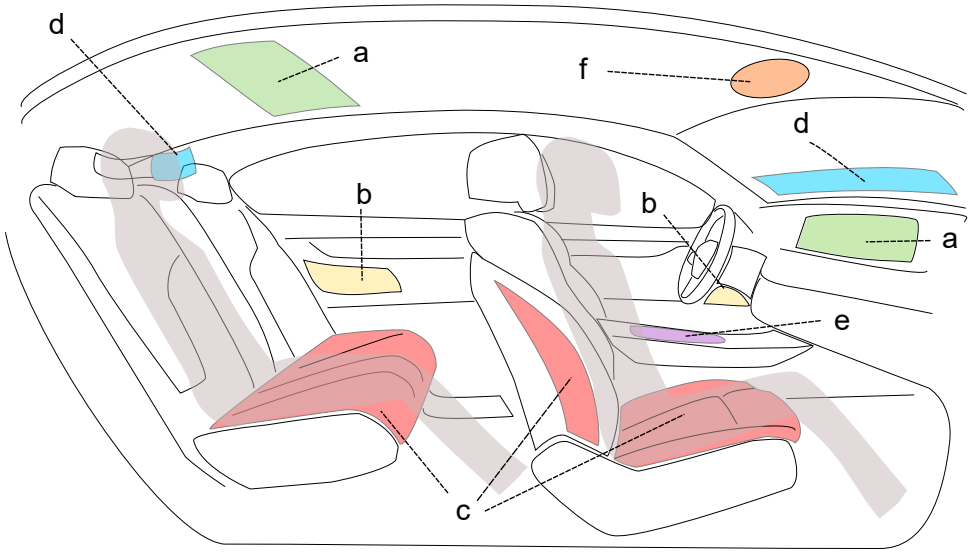



FIGURE 2.3: Illustration of interior sensing and interaction technologies in a vehicle, adapted from Karas, D. M. Schuller, and B. W. Schuller, 2024. a) Display screens, b) Microphone and speaker arrays, c) seat pressure sensors, d) Head Up Display and AR glasses, e) Haptic control surface, f) interior camera or radar system.

embedded devices, which is important for resource-constrained environments and privacy, especially for health-related use cases.

### Physiological

iMotions [10] is another subsidiary of Smart Eye, which now integrates Affectiva's systems in their platform. The platform aggregates signals from various biosensors, such as eye trackers and physiological sensing equipment. It is marketed as a study design tool for researchers as well as a business analytics solution.

### Automotive Applications

Modern vehicles include a large number of sensors directed at the environment as well as at the interior. Interior sensing is motivated by stricter driving safety requirements, such as driver monitoring for sleepiness and distraction (Koesdwiady et al., 2017; J. Wang, Warnecke, et al., 2020). Regulators are also moving towards requirements for e. g., advanced occupant crash protection and child presence detection (Euro NCAP, 2017; McStay and Urquhart, 2022), which necessitates more holistic in-cabin sensing. At the same time, comfort and entertainment functions are also becoming more ubiquitous, shifting the focus from driver analytics to all occupants (Tan et al., 2022). This trend is expected continue as driving becomes more autonomous (Vögel et al., 2018) and the distinction between driver and passenger gradually disappears.

Affective computing in the car is already being deployed in concept studies, as well as in series production vehicles. Technical solutions mainly rely on the visual and audio modalities, via in-cabin camera systems and microphone arrays. Physiological

[10] https://imotions.com/platform/

signals from wearables are also used to enhance wellness functions. An illustration of interior sensing and interaction technologies is given in fig. 2.3. Some examples of use cases from concept cars and production vehicles are listed below (Karas, D. M. Schuller, and B. W. Schuller, 2024).

Audi has introduced the concept car *Elaine* (Audi, 2017), which includes a voice-controlled personal assistant. *Elaine* can detect the driver's vital parameters and activate revitalising functions to alleviate stress and fatigue.

Toyota has presented the *LQ*, which contains an emotional AI assistant named *Yui* (Toyota, 2019). *Yui* is supposed to engage in empathetic conversations and offer assistance to stressed drivers. It can adjust various aspects of the interior, including illumination, music, air flow, and fragrance.

KIA has introduced *Real-Time Emotion Adaptive Driving (R.E.A.D)* at CES 2018, a system based on facial expression and ECG signals to adapt their concept car's interior to the passengers' emotions (KIA, 2019).

Electric vehicle startup NIO[11] has integrated an emotional assistant named NOMI into their vehicles. NOMI takes the form of a small sphere on the dashboard, which can display emoji-like expressions and turn towards individual passengers to signal attention. It reacts to voice commands and can access an interior camera (Nio, 2020).

Mercedes Benz [12] cars include a personalised user interface named MBUX that responds to touch, voice and gesture commands (Mercedes-Benz, 2022b). For additional comfort, Mercedes offer a solution called *ENERGIZING*. *ENERGIZING* is a collection of programs that combine features designed to improve passenger wellbeing (Mercedes-Benz, 2022a). These include climate control and scented air, relaxing music, seat position adjustment and massages. There is also an interactive coaching system that can analyse the driver state through driving behaviour, as well as accessing vital parameters through wearables connected to the car.

BMW [13] includes a voice assistant named *IPA* in their vehicles, which can control various interior functions. One of them is the *Caring Car* mode, a comfort function designed to increase the driver's wellbeing. *Caring car* adjusts the audio settings, climate control and interior lighting. There are two options aimed at relaxing or revitalising. Thus, *Caring Car* can be considered a type of mood regulation.

BMW has also integrated a roof-mounted camera into their 2021 electric flagship, the iX. The fish-eye camera can see all occupants and enables a number of interior sensing capabilities. This includes the *Happy Snapshot* use case, which can detect smiling and take a selfie (BMW, 2021).

With this chapter having established the background of the topic and the state of the art, the next chapter will focus on the methodology used for this thesis.

[11] https://www.nio.com/
[12] https://www.mercedes-benz.de/
[13] https://www.bmwgroup.com/

# Chapter 3

# Methodology

In this chapter, the methodology used for the experiments in chapter 4 is established. Aspects of the methodology include:

1. Criteria for dataset selection
2. Feature Extraction Methods
3. Losses and Metrics
4. Model architecture

## 3.1 Datasets

Given that the focus of this thesis is on detecting affective and affect-related states in a realistic setting, datasets are chosen with a preference for naturalistic, in-the-wild data.

For the purpose of continuous valence and arousal recognition, the following datasets are chosen:

- Aff-Wild2
- SEWA
- RECOLA
- Hume-VB

Aff-Wild2 (Kollias and Zafeiriou, 2019) is chosen for being a large in-the-wild database of videos containing spontaneous emotional displays. Thus, it is well suited for RQ–1. It offers highly challenging data due to being sourced from YouTube, which causes large variations in camera setup and recording quality. The topics of the videos are also quite diverse, from public talks to vlogs and reaction videos. Aff-Wild2 encompasses 558 videos with more than 2.7M frames, and 458 subjects. Usually there is a single person per video, although some may include multiple subjects. The data is annotated with time-continuous labels on a frame-by-frame basis, at an average frame rate of 30 fps. The labels include valence, arousal, expression in terms of the six basic emotions and neutral, as well as 12 AUs.

An overview of the subset from Aff-Wild2 annotated for valence-arousal and presented in the 3rd ABAW challenge is given in table 3.1. The number of videos exceeds that in the total dataset since some videos were segmented into multiple clips.

TABLE 3.1: Partitioning, numbers of subjects, and duration of the subset of the Aff-Wild2 dataset used for valence-arousal prediction.

| Partition | Subjects | Videos | Duration |
|---|---|---|---|
| Training | - | 341 | 15:40:44 |
| Validation | - | 71 | 02:53:51 |
| Test | - | 152 | 07:05:51 |
| Σ | 455 (277M/178F) | 564 | 25:40:27 |

Usually only one subject appears per video, although some may contain several persons. The total number of subjects in this set is 455, and the number of frames is approximately 2.8M (Kollias, Tzirakis, Baird, et al., 2023). The distribution of subjects across the partitions is not stated by the challenge organisers.

TABLE 3.2: Partitioning, cultures, number of subjects, and duration of the conversations included in the SEWA dataset. The partitioning of AVEC'19 CES, which is a subset of SEWA containing 200 labelled German, Hungarian and Chinese videos is used. German and Hungarian are present in all partitions, while Chinese is only part of the blind test set. The remaining 194 English, Serbian and Greek videos are unlabelled and are used for self-supervised training.

| Partition | Culture | Subjects | Duration |
|---|---|---|---|
| Training | DE | 34 | 1:33:27 |
| | HU | 34 | 1:08:41 |
| | EN | 66 | 02:33:21 |
| | SR | 72 | 02:35:45 |
| | GR | 56 | 02:41:38 |
| Development | DE | 14 | 37:52 |
| | HU | 14 | 28:50 |
| Test | DE | 16 | 46:47 |
| | HU | 18 | 36:18 |
| | CN | 70 | 3:18:14 |
| Σ | | 394 | 16:19:25 |

SEWA (Kossaifi et al., 2019) is chosen for containing spontaneous emotion displays in the wild, and for being a multi-cultural database, covering the cross-cultural aspects of RQ–1. It is composed of dyadic conversations between subjects of 6 different nationalities. Subjects are shown a number of advertisement videos and then discuss freely via a video-chat system that records their faces and voices. A subset of *SEWA* is the *Audiovisual Emotion Challenge (AVEC) 2019* dataset, which consists of 200 German, Hungarian and Chinese videos that are annotated continuously with arousal, valence and liking at 10Ḣz. More information on the composition of the dataset is listed in table 3.2. The rest of SEWA contains more conversations of English, Serbian and Greek subjects. Thus, three more cultures are available for analysis, which is relevant for RQ–2. However, while the videos contain spontaneous emotion displays, they have not been labelled and can not be used for evaluation. Instead, they will be used for unsupervised training, which also contributes to RQ–3.

The *RECOLA* corpus (Ringeval, Sonderegger, et al., 2013) is a multi-modal French

language database containing dyadic conversations between Swiss students. Its design is a combination of task-based elicitation to set social context with spontaneous interactions. Participants were asked to discuss strategies to survive in a hostile environment. Audiovisual data was captured, as well as EDA and ECG signals. The database includes 27 videos and is continuously annotated with arousal and valence at 25Ḣz. The experimental setup of RECOLA is similar to SEWA, but it has fewer subjects and a smaller amount of total runtime. Hence, this thesis focuses on SEWA, and RECOLA is used indirectly for preliminary experiments and pre-training of feature extractors.

TABLE 3.3: Overview on the HUME-VB dataset in terms of partitions, cultures, speakers, samples and audio durations, adapted from Baird, Tzirakis, Brooks, et al., 2022. Test set statistics are unknown since the data is used in a ML competition.

| **Partition** | **Culture** | **Subjects** | **Samples** | **Duration** |
|---|---|---|---|---|
| Training | USA | 206 | 7142 | 03:48:54 |
| | China | 79 | 5120 | 03:35:30 |
| | South Africa | 244 | 5090 | 03:07:55 |
| | Venezuela | 42 | 2638 | 01:46:46 |
| | Σ | 571 | 19900 | 12:19:06 |
| Validation | USA | 206 | 7020 | 03:58:21 |
| | China | 76 | 4999 | 03:21:29 |
| | South Africa | 244 | 4804 | 03:01:54 |
| | Venezuela | 42 | 2573 | 01:44:00 |
| | Σ | 571 | 19396 | 12:05:45 |
| Test | USA | - | - | - |
| | China | - | - | - |
| | South Africa | - | - | - |
| | Venezuela | - | - | - |
| | Σ | 563 | 19815 | 12:22:12 |
| Σ | - | 1702 | 59201 | 36:47:04 |

The HUME-VB dataset is chosen in order to address RQ–4. It contains vocal bursts, i. e., short, non-verbal vocalisations such as laughter or crying. Subjects were given an example clip and instructed to record themselves mimicking the sound, displaying emotions that would be perceived as similar to the original. The bursts were then ranked in terms of their emotional intensities in 10 categories (Awe, Awkwardness, Amusement, Distress, Excitement, Fear, Horror, Sadness, Surprise, Triumph) by an average of 85.2 raters. The dataset contains 59201 recordings from 1702 speakers, for a total of more than 36 hours of audio. Four cultures are present in the data: US American, Chinese, Venezuelan, South African. Information on the dataset is summarised in table 3.3.

HUME-VB has multiple properties that make it well suited for study in this thesis. As it was recorded by the subjects on their own devices during their daily lives, it can be considered in-the-wild data. While it does not contain spontaneous displays of emotions, the mimicked vocal bursts are assumed to still be fairly realistic. Furthermore, the dataset contains rich annotations on each sample, including continuous affect in terms of valence and arousal. It also offers the intensities of each

of the 10 annotated emotions, leading to an interesting multi-class regression problem, reflecting the theories introduced in section 2.1 that treat emotions as smoothly varying concepts. The dataset is reasonably large, and fairly unique in its focus on non-verbal affective vocalisations. An added bonus for this thesis is that it includes data from four distinctive cultures and explicitly provides annotations for each of them.

## 3.2 Features

This section motivates the choices of signal modalities and features used for experimentation.

Regarding the modalities, the focus of this thesis is on audiovisual data. There are multiple reasons for this:

- The modalities naturally co-occur in video material, which is available in large quantities in diverse settings, allowing for the study of affect in the wild.
- The face and the voice contain information which is key to understanding human interaction, as evidenced by the fact that humans rely heavily on them.
- There are established datasets and a significant body of work related to video analysis, which allows for comparisons of methods and results.

Handcrafted descriptors such as AUs and eGeMAPS have been proven to be effective for affective computing. Therefore, they are frequently used as inputs into neural networks, either directly or as functionals across several seconds (Ringeval, B. Schuller, Valstar, Cummins, Cowie, Tavabi, et al., 2019).

However, given the advances in deep learning outlined in section 2.3, the focus of this work is on using neural networks as feature extractors. This will allow the models built on top to leverage representations which may contain information not easily accessible from hand-crafted descriptors. Pre-trained networks can also be used to transfer knowledge from much larger datasets, helping to overcome the bottleneck of data scarcity in affective computing.

The choice of modality which is used as input to the model leads to different architectures, which are described in the following subsections. The feature extraction architectures are sketched in fig. 3.1.

### 3.2.1 Audio

For the audio modality, two different solutions are used. They have in common that they directly process the audio signal, segmented into short overlapping clips, instead of relying on Mel-spectrograms or MFCCs as intermediate features. Thus the deep model is free to learn relevant information directly from the raw data.

In the first approach, 1D CNNs are used to extract embeddings. Compared to the two-dimensional spectrogram and CNN-based approach, this has the advantage of a lower parameter count (on the order of $10^5$ as opposed to $10^6 - 10^7$). In addition, as shown in previous work, even networks with a small number of layers perform well as feature extractors (Tzirakis, Trigeorgis, et al., 2017; Schmitt, Cummins, and B. Schuller, 2019; J. Zhao, Mao, and L. Chen, 2019). Thus, it is possible to train these models end-to-end with relatively low computational cost, and extract their weights

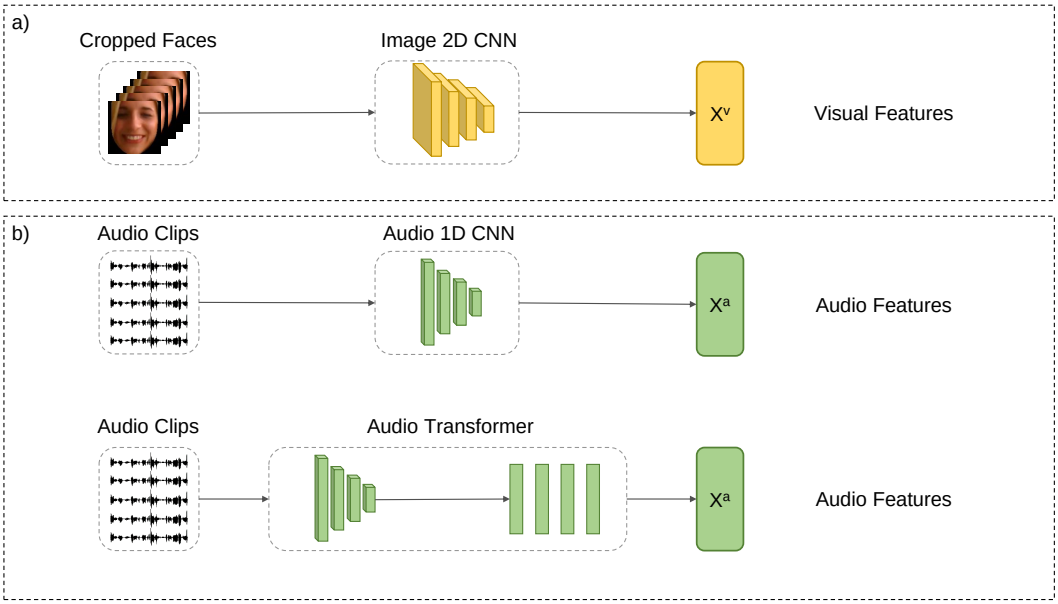



FIGURE 3.1: Schematic of feature extraction networks. Top: Visual 2D-CNN processing face crop images. Middle: Audio 1D-CNN processing raw audio waveforms. Bottom: Audio Transformer, consisting of a CNN and multi-head attention based encoder layers, processing raw audio waveforms.

for later use as feature extractors. In this work, this task will be accomplished primarily using the END2YOU toolkit Tzirakis, 2020.

The second approach makes use of the much larger WAV2VEC2 model (Baevski et al., 2020), which is an audio Transformer with approximately 80M parameters in its base variant. This choice is motivated by the strong performances Transformer-style architectures have recently delivered in many fields (Devlin et al., 2019; Dosovitskiy et al., 2020; Wagner et al., 2023). WAV2VEC2 consists of a 1D CNN which creates embeddings from the audio, followed by a stack of encoder layers. The inner workings of those layers will be explained in section 3.4.2.

### 3.2.2 Visual

As discussed in section 2.3 state of the art in computer vision is based on CNNs and Transformers, which are pre-trained on very large datasets to recognise different classes of objects or distinguish people's identities with high accuracy. These models typically have dozens of layers and millions of parameters, which allows them to learn versatile features. They are commonly adapted to new classification tasks by replacing their final layer and training this layer with the new data while the rest of the network remains frozen. Alternatively, the top layer can be removed to yield a high-dimensional feature extractor.

Given the effectiveness of this feature extraction method, pre-trained CNNs will be used as the first stage of visual models. In order to keep the number of trainable parameters manageable, typically only the last layers of the networks will be trainable, or the entire CNN network (without its classification output layer) will be frozen and used as a static feature extractor.

The visual representations returned by the CNN's last hidden layer will need to be processed with a suitable architecture for the downstream task. This will usually

involve some modelling of sequential data, in the form of an architecture that can process multiple images. 3D-CNNs are theoretically suitable for this task, however, the full 3-dimensional convolution is computationally expensive, and previous work suggests that a 2D+1 approach may be more helpful (Kuhnke, Rumberg, and Ostermann, 2020). Hence, this thesis will not rely on 3D-CNNs, but on a combination of 2D-CNNs with a suitable temporal architecture, such as a *recurrent neural network (RNN)*.

The choice of the data and task for pre-training affects the usefulness of the derived representations for the task of emotion recognition. Since this work is focused on detecting emotions from the face, CNNs pre-trained on facial analysis will be used.

## 3.3 Losses and Metrics

A key part of successfully training a model is choosing the appropriate objective function, or *loss function*. During training, the optimiser will adjust the model weights in an attempt to minimise the loss. More complex architectures may have multiple outputs, with losses on each of them needing to be balanced (*Multi-task Learning (MTL)*.

At the evaluation step after following training, as well as at regular intervals in between, the model's performance is checked by *metrics*, which map the labels and predictions to a score, usually between 0 and 1.

This section will introduce common metrics and losses, as well as methods to balance the latter in multi-task settings.

### 3.3.1 Metrics

Most of the research in this thesis is concerned with regression problems, i. e., predicting targets with continuous value ranges such as the arousal and valence dimensions. The only classification problem addressed herein is predicting the type of vocal bursts in the HUME-VB dataset. The choice of classification metric will be introduced first, before moving on to the regression metric.

#### Classification

The *accuracy* is defined as the percentage of samples that were classified correctly:

$$acc = \frac{1}{N} \sum y = \hat{y} \tag{3.1}$$

For binary classification, the *recall* is the rate of true positives to the sum of true positives and false negatives, given by:

$$recall = \frac{t_p}{t_p + f_n} = \frac{\sum y = 1 | \hat{y} = 1}{\sum y = 1 | \hat{y} = 1 + \sum y = 0 | \hat{y} = 1} \tag{3.2}$$

The *precision* is the rate of true positives to the sum of true positives and false positives:

$$precision = \frac{t_p}{t_p + f_p} = \frac{\sum y = 1 | \hat{y} = 1}{\sum y = 1 | \hat{y} = 1 + \sum y = 1 | \hat{y} = 0} \tag{3.3}$$

In the multi-class classification setting, precision and recall can be computed in a class-wise, one-vs-all manner. However, the classes may be unbalanced, e. g., the neutral class is over-represented. In this case, if a weighted average of the scores of each class were to be taken, the performance of the majority classes may conceal poor classification of the minority classes. Thus, the *unweighted average recall (UAR)* has been proposed to gauge multi-class systems.

Due to the advantages of UAR for gauging unbalanced datasets, it is chosen as the classification metric for this thesis. This matches with the choice of the organisers of the A-VB competition for judging the vocal burst classification sub-challenge.

**Regression**

For regression, a popular metric is the *mean square error (MSE)*, which is defined by:

$$MSE = \frac{1}{N} \sum (y - \hat{y})^2 \tag{3.4}$$

The MSE and the similar *mean absolute error (MAE)* measure the average deviation of each sample from the true value. An alternative, which has become increasingly popular for works on dimensional emotion recognition, are metrics that measure the correlation between two sequences of values. The *Pearson correlation coefficient* is defined as:

$$\begin{aligned} \rho(X, Y) &= \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (x_i - \mu_X)(y_i - \mu_y)}{\sqrt{\sum (x_i - \mu_x)^2} \sqrt{\sum (y_i - \mu_y)^2}} \\ \text{where: } \mu_X &= \frac{\sum x_i}{N}, \mu_Y = \frac{\sum y_i}{N} \end{aligned} \tag{3.5}$$

The Pearson correlation coefficient measures only the linear relationship between the variables, without considering scale and bias (Pandit and B. Schuller, 2020). The *Concordance Correlation Coefficient (CCC)* proposed by Lin, 1989 thus modifies the Pearson coefficient $\rho$ with a factor $C_b$ as follows:

$$\begin{aligned} \rho_c(X, Y) &= \rho * C_b = \rho * \frac{2}{\left(v + \frac{1}{v} + u^2\right)} \\ \text{where: } v &= \frac{\sigma_X}{\sigma_X}, u = \frac{(\mu_X - \mu_Y)}{\sqrt{\sigma_X \sigma_Y}} \\ \rho_c &= \frac{2\sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 + (\mu_X - \mu_Y)^2} \end{aligned} \tag{3.6}$$

In eq. (3.6), $v$ acts as a scale penalty, while $u$ acts as a shift penalty.

Thus, a CCC of 0.0 means there is no correlation between two variables, while 1.0 means they correlate perfectly. CCC has been widely adopted in the field of affective computing for measuring the performance of continuous emotion recognition models, including in the aforementioned competitions AVEC 2019 (Ringeval, B. Schuller, Valstar, Cummins, Cowie, Tavabi, et al., 2019), ABAW (Kollias and Zafeiriou, 2021; Kollias, Tzirakis, Baird, et al., 2023) and A-VB (Baird, Tzirakis, Brooks, et al., 2022).

Given its prominent role in the literature for the type of problem addressed in this thesis, CCC is chosen as the primary metric.

### 3.3.2 Loss functions

Next, the loss functions used for optimising the neural networks will be described.

$$\mathcal{L}_{ccc} = 1 - CCC \tag{3.7}$$

The main loss function for optimising the continuous emotion prediction task is the CCC loss in eq. (3.7), obtained by subtracting the CCC of predictions and targets from 1.0. This loss is chosen as it has been proven effective for sequential data (Ringeval, B. Schuller, Valstar, Cummins, Cowie, and Pantic, 2019). Using MSE as the loss function instead may lead to issues in training when combined with CCC as metric. This due to the relationship between the two, for a detailed discussion see Pandit and B. Schuller, 2020. However, MSE may be used as a supplementary loss term along with the main CCC loss for smoother convergence, making it easier to estimate when to stop training to prevent overfitting.

This thesis tackles multi-regression problems, as the goal is to predict multiple continuous variables (usually valence and arousal, or the emotional intensities in the HUME-VB dataset). Thus, a multi-task approach is chosen to optimise them jointly. This has been shown to be effective due to the dependencies between the affect dimensions (Eyben, Wöllmer, and B. Schuller, 2012). However, multi-task learning raises the question of how to balance the individual tasks. The simplest way is to take the mean across task losses, as in eq. (3.8).

$$\mathcal{L} = \frac{\mathcal{L}_{arousal} + \mathcal{L}_{valence}}{2} \tag{3.8}$$

This loss treats each task identically, irrespective of their difficulty or importance. Alternatively, the training can be adjusted to emphasise a particular task, e. g., arousal prediction, by weighing the loss towards it, as in eq. (3.9). The weights are scaled to sum up to 1. Training multiple models with different weighted losses allows for selective optimisation of each task.

$$\mathcal{L} = w_{arousal}\mathcal{L}_{arousal} + w_{valence}\mathcal{L}_{valence} \tag{3.9}$$

The previously shown loss functions have the disadvantage that their weights are static, i. e., fixed during training. They need to be chosen well for balancing the training process across the tasks, which gets harder the more tasks there are. Static weights may lead to sub-optimal results due to the model essentially focusing on the easier tasks. To combat this, researchers have introduced adaptive weight schemes. For instance, *Dynamic Weight Averaging* scales loss weights over time based on how the loss changed in previous steps (Liu, Johns, and Davison, 2019). The weighting

function contains a temperature softmax, with the temperature $\mathcal{T}$ smoothing the weight distribution and $\mathcal{K}$ scaling the weights to sum up to 1.0.

$$\mathcal{L}_{dwa} = \sum_K \lambda_k(t)\mathcal{L}_k = \sum_K \mathcal{K}\frac{exp\left(\frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)}/\mathcal{T}\right)}{\sum_K exp\left(\frac{\mathcal{L}_k(t-1)}{\mathcal{L}_k(t-2)}/\mathcal{T}\right)}\mathcal{L}_k \tag{3.10}$$

The *Revised Restrained Uncertainty Weighting* (RRUW) proposed by Song et al., 2022 balances tasks via trainable parameters $\alpha_k$. Constraints are imposed, in order to prevent the training from collapsing into trivial solutions e. g., setting loss weights to 0. Logarithmic terms ensure that very small values of $\alpha_k$ have a considerable impact on the loss. In addition, a positive value $\varphi$ is defined and the sum of weights is driven towards it. The complete loss function is given in eq. (3.11)

$$\begin{aligned}\mathcal{L}_{rruw}(w,\alpha) = &\sum_K \frac{1}{\alpha_k^2}\mathcal{L}_k(w) + \sum_K \log\left(1+\log\alpha_k^2\right) + \\ &|\varphi - \sum_K (|\log\alpha_k|)\,|,\end{aligned} \tag{3.11}$$

By combining DWA and RRUW, the *Dynamic Restrained Uncertainty Weighting* (DRUW) (Song et al., 2022) is derived. It contains both dynamic and uncertainty terms, as in eq. (3.12).

$$\begin{aligned}\mathcal{L}_{druw}(w,\alpha) = &\sum_K \left(\frac{1}{\alpha_k^2} + \lambda_k(t)\right)\mathcal{L}_k(w) + \sum_K \log\left(1+\log\alpha_k^2\right) + \\ &|\varphi - \sum_K (|\log\alpha_k|)\,|,\end{aligned} \tag{3.12}$$

In the experiments in the following chapter, both static and dynamic weighing of multiple CCC losses for arousal and valence prediction, as well as for the prediction of other continuous emotion tasks, will be used. Dynamic weights are expected to deliver performance gains compared to the static counterparts.

Another strategy to assist the training of the model is to supplement the regression loss with a classification loss. Humans do not perceive emotions in terms of a vector of continuous values, but categorise them. Thus, a model may benefit from multi-task learning, where one task is to predict values of valence and arousal, and another is to predict a categorical label. Such multi-task models have been used on corpora where both affect dimensions and basic emotions are annotated. If there are no categorical annotations available, the value-continuous labels may be discretised (Toisoul et al., 2021).

In this work, a discretisation scheme of the 2-dimensional valence-arousal space is proposed based on a polar coordinate transformation as in eq. (3.13). This transformation maps a pair of valence-arousal labels $(y_v, y_a)$ to a segment of the 2D affect plane. The number of bins can be chosen freely with the radii and angle limits of the segments.

$$
\begin{aligned}
r &= \sqrt{y_a^2 + y_v^2} \\
\theta &= \tan^{-1}\left(\frac{y_a}{y_v}\right)
\end{aligned}
\tag{3.13}
$$

When a classification task needs to be solved, this thesis will employ the categorical cross-entropy as the loss function.

The train loss has to be computed batch-wise for performance reasons. However, doing the same for the validation step may not give an accurate picture of model performance. A validation loss computed batch-wise can be less stable, leading to issues with algorithms that rely on tracking it during training. Following the methodology of the AVEC challenges, (Ringeval, B. Schuller, Valstar, Cowie, et al., 2018; Ringeval, B. Schuller, Valstar, Cummins, Cowie, Tavabi, et al., 2019), instead the predictions and labels are concatenated across the entire validation set before the CCC is computed.

## 3.4 Models

This section describes the methodology for model architectures used in this thesis. On an overarching level, there are four aspects to the design, which have some overlap. These are knowledge transfer, temporal modelling, cross-modal interactions and multi-modal fusion, and cross-cultural recognition.

### 3.4.1 Transfer Learning

The first aspect, as described in section 3.2, is on the topic of transfer learning, i. e., transferring existing knowledge from a related problem or dataset into the model (Weiss, Khoshgoftaar, and D. Wang, 2016). It is motivated by the trends of the field of deep learning in general and the data scarcity issues related to affective computing in particular. For the purpose of this thesis, the transfer is accomplished in one of two ways. The first, used with the CNN models, is to re-use pre-trained models that have been trained on an identical or sufficiently similar task on a different dataset. This includes audio CNNs trained for emotion recognition, and vision CNNs trained for facial recognition. The second approach is to use models trained on large amounts of unlabelled data like WAV2VEC2, which has let them obtain general knowledge.

The models can be used in a frozen configuration during training, in which case they become simple extractors for the features listed above, or they can be tuned by making their layers trainable. Fine-tuning the weights helps mitigate domain discrepancies and optimises the model for its new task. However, it comes at increased computational expense due to the additional backward pass.

### 3.4.2 Temporal Modelling

It can be assumed that in most cases, the model will benefit from capturing temporal dynamics in the data. This is because emotions are not static but transient states that fluctuate. Therefore, the inputs to the model will consist of sequences of feature vectors corresponding to consecutive time steps. The model can then either

incorporate a *sequence-to-sequence (seq-2-seq)* prediction component, or aggregate the representations from the time steps into one clip-level prediction.

For the temporal component, the simplest solution may be a RNN with LSTM or GRU cells. This can form a baseline model component. Alternatively, 1D-CNNs may be used, as demonstrated in Schmitt, Cummins, and B. Schuller, 2019.

Another, more complex solution is proposed based on the aforementioned Transformer. The Transformer, introduced by Vaswani et al., 2017, is a seq-2-seq model that uses attention instead of convolutions and recurrent layers. It consists of an encoder, which performs self-attention on the input tokens, and a decoder, which combines the encoded embeddings with the output tokens. The model predicts in an autoregressive manner, i. e., one token at a time.

At the heart of the Transformer is Multi-Head Attention (MHA), which lets the model discover relationships between the elements of a sequence. The basic version of the Transformer used scaled dot-product attention, which is defined as in eq. (3.14).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3.14}$$

MHA parallelises the attention process onto multiple sub-networks called heads, by mapping the query, key, and value inputs into different sub-spaces. The results from the heads are then recombined and projected into the output dimension. It is defined as:

$$\begin{aligned} &\text{MHA}(Q, K, V) = \text{Concat}\left(\text{head}_1, ..., \text{head}_n\right) W^O, \\ &\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \end{aligned} \tag{3.15}$$

A Transformer layer is obtained by combining MHA with a (fully connected) feed-forward network, residual connections and layer normalisation. In the encoder, the queries, keys and values for a given layer are identical, which is called *Self-Attention (SA)*.

$$\begin{aligned} P(k, 2i) &= \sin\left(\frac{k}{n^{2i/d}}\right) \\ P(k, 2i+1) &= \cos\left(\frac{k}{n^{2i/d}}\right) \end{aligned} \tag{3.16}$$

Unlike recurrent cells e. g., LSTM, the attention layer processes the elements of a sequence in parallel and has no inherent concept of sequence order. Therefore, Vaswani et al., 2017 propose adding position encodings to the input. eq. (3.16) shows the popular sinusoid encoding. Here $k$ is the index of a feature vector (token) in the input sequence, $d$ is the dimensionality of the features, $i$ is the dimension index running from 0 to $d/2$ and $n$ is an adjustable parameter. In the next chapter, $n = 5000$ will be used based on preliminary experiments.

The self-attention block is chosen as a suitable seq-2-seq component for experimentation. It is illustrated in fig. 3.2.

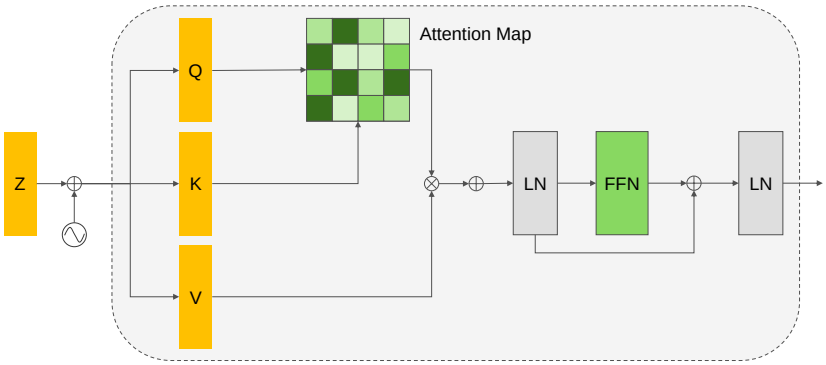



FIGURE 3.2: Self-attention block from the Transformer architecture. A sequence of embeddings (to which position encoding may be added) is passed as inputs to queries Q and key-value pairs K, V. Multi-head attention is performed, followed by a feed-forward network with layer normalisation and addition.

Furthermore, combining Transformer encoders and recurrence can be a viable solution. J. Huang et al., 2020 adapted an audiovisual transformer for continuous emotion recognition on the AVEC 2017 dataset, processing the features with self attention encoders before applying a cross-modal block. Adding a single LSTM layer on top of the network was found to benefit performance.

### 3.4.3 Cross-modal interaction and multi-modal fusion

The third aspect is the focus on cross-modal interactions as a way to build richer representations in a multimodal setting.

Multi-modal fusion can be expected to deliver superior results to uni-modal approaches, assuming that the modalities contain complementary information (Poria, Cambria, et al., 2017). If some modalities are missing for extended periods of time, or are strongly affected by noise, fusion may not give significant benefits. In those cases, a uni-modal system may be preferable, especially if computational resources are limited.

For the purpose of this thesis, most experiments will be conducted with multiple modalities. The exception are ablation experiments for quantifying the effectiveness of fusion, and all experiments conducted for RQ–4, as the vocal burst dataset contains only audio data.

The benefits of early, decision and hybrid level fusion have been introduced in section 2.4. All of them have been used successfully by previous works (W. Wei, Jia, and Feng, 2017; Ghaleb et al., 2017; H. Chen, Y. Deng, Cheng, et al., 2019; Hamieh et al., 2021).

In the context of this thesis, early and hybrid approaches are preferred to late fusion. This is motivated by a focus on cross-modal interactions. As discussed in section 2.4, early and hybrid fusion can fuse information at the feature level, which late fusion can not. The trade-off is that late fusion is more flexible, allowing to combine

decisions of classifiers without the need to have all modalities available simultaneously on the same computational platform. It is also potentially more light-weight by avoiding large feature embeddings.

Early fusion may be implemented with a simple tensor concatenation operation. The combined tensors are then passed into an arbitrary networks described in section 3.4.2 for extracting higher-dimensional representations.

Model-based fusion can help model cross-modal interactions more explicitly than early fusion. Tsai et al., 2019 proposed MulT, a type of transformer which fuses audio, visual and textual information. It does so by combining modalities pairwise in cross-modal blocks, with one modality attending to the other similar to a transformer decoder. Thus, for 3 modalities there are 6 interactions. The resulting embeddings are then concatenated per query modality and fed into transformer encoders, before being combined for the final prediction. Cross-modal fusion was shown to be able to capture relations between two sequences of data, even if those sequences are not perfectly aligned. Therefore, the multi-modal experiments for this thesis will also make use of it.

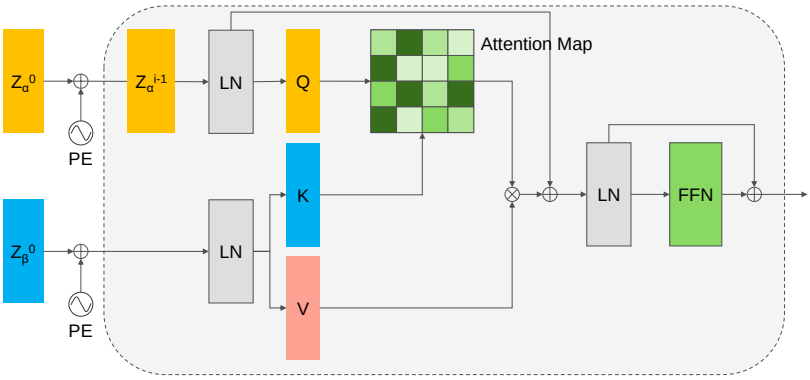



FIGURE 3.3: Cross-modal attention fusion block. Two different sequences of embeddings (with optional position encodings) serve as queries Q and key-value pairs K, V, respectively. Multi-head attention is performed, followed by a feed-forward network. The placement of the layer normalisation may vary by implementation.

The cross-modal fusion block is illustrated in fig. 3.3. It is a modification to the standard self-attention block, which computes attention between the tokens of one sequence. The CMA block can process two input sequences $z_\alpha$ and $z_\beta$, which may be of different lengths but should have the same feature dimensionality for the block's internal matrix multiplications. The block attends one sequence to the other, $\beta \rightarrow \alpha$, by using $z_\alpha$ as queries and $z_\beta$ as key-value pair inputs to multi-head attention. This is then followed by the usual feed-forward network. When multiple cross-modal blocks are stacked, the queries become the output of the previous block $z_\alpha^{i-1}$, but the key-values re-use the original $z_\beta^0$.

In a hybrid fusion approach, the audio and visual modalities may thus be used as either input to the cross-modal attention encoder. It is also possible to run multiple encoders in parallel, with different modalities attending to each other, or to

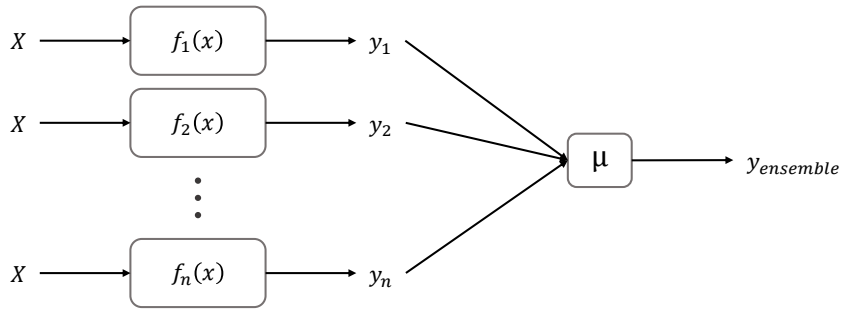



FIGURE 3.4: Ensemble strategy for combining $n$ emotion regression models. Each model is trained independently as a function mapping data $X$ to labels $y$. Then the predictions are fused by computing the average for each sample.

stack cross-modal and self-attention blocks, see Tsai et al., 2019, J. Huang et al., 2020. A further iteration of transformer based fusion for continuous emotion recognition is TEMMA, introduced by H. Chen, Jiang, and Sahli, 2020. It consists of a 1D-CNN temporal embedding network, an attention module which models temporal dependencies through self attention and inter-modality dependencies by constructing multi-modal queries, keys and values, and an inference network that concatenates the feature embeddings and predicts valence and arousal with fully connected layers. TEMMA was applied to the AVEC 2016 and AVEC 2019 datasets.

Finally, while late fusion is not directly used in this thesis, due to the desire to capture lower-level interactions between the modalities, some of its benefits can still be accessed. This is accomplished in the form of model ensembles, i. e., the predictions of multiple trained models are combined before the final performance metric is computed. This can be thought of as a late fusion where all models have access to the same features. Ensemble strategies have been shown to be effective across many tasks, including on emotion estimation in the wild (Meng et al., 2022). They work by pooling the knowledge of multiple models.

To perform an ensemble step, a combination method is needed. The rule used in this thesis is chosen to be simple to implement while being effective: Classifier models are combined via majority voting, while regression models are combined by computing the mean of their predictions. This is illustrated in fig. 3.4.

### 3.4.4 Cross-cultural adaptation

This subsection presents methods needed for investigating RQ–2, i. e., how to improve the performance of emotion recognition models when presented with affective displays of people from cultures other than those the model was trained on.

In general machine learning terms, this constitutes a *Domain Adaptation* problem. A domain $\mathcal{D}$ is defined as a combination of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ from which the samples are drawn, whereas a task $\mathcal{T}$ is a combination of a label space $\mathcal{Y}$ and an objective function $f$ (Pan and Q. Yang, 2010). In this case, the source domain $\mathcal{D}_{src}$ and the target domain $\mathcal{D}_{tgt}$ differ by distribution, while the feature spaces are identical $\mathcal{X}_{src} = \mathcal{X}_{tgt}$. The tasks are also identical i. e., continuous affect prediction. The purpose of domain adaptation strategies is to

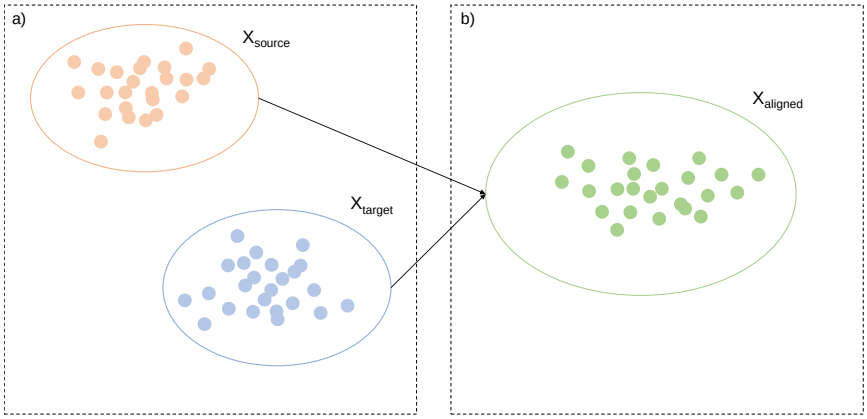



FIGURE 3.5: Visualisation of domain adaptation. a) Representations with separate distributions in source and target domains. b) Alignment into a common, domain-invariant representation

mitigate the shift between source and target domains. Numerous methods for correcting distribution differences exist, e. g., using autoencoders or training classifiers with pseudolabels. See for instance Weiss, Khoshgoftaar, and D. Wang, 2016 for a discussion of domain adaptation in the context of transfer learning (in which the problem presented here would fit into *Homogeneous Transfer Learning*).

The approach used in this thesis will focus on mapping the representations of source and target domains into a common latent space in an unsupervised manner, i. e., without having access to the emotion labels of the target domain. Instead, the data is implicitly labelled by the domain it belongs to. Emotion labels are only needed on the source domain for training the affect task. The adaptation process is illustrated in fig. 3.5.

The domain adaptation task and the emotion prediction task will be learned simultaneously. In order to accomplish this, the models for supervised multimodal affect prediction described above will need to be extended. For this thesis, the modification takes the form of a Domain Adversarial Neural Network (DANN).

The DANN, originally proposed by Ajakan et al., 2014, is designed to mitigate the general domain shift problem that occurs when training and test data come from different domains. To achieve this purpose, the model is guided to learn an internal representation of the data which contains relevant information for the task, but is invariant to the domain.

Ganin et al., 2016 showed that the DANN can be implemented with three components: A feature encoder, a classifier and a domain discriminator. These can then be trained using regular stochastic gradient descent. However, if this model were simply trained with a multi-task objective, it would learn to discriminate between the domains. In order to make the model agnostic of the domain information instead, it is important to add a *Gradient Reversal Layer*, which flips the sign of the gradient flowing from the domain discriminator in the backward pass. Thus, the discriminator still attempts to learn the difference between the domains, but the updates received by the feature encoder act towards obscuring the domain information instead. The classifier learns to predict the desired task on this invariant representation returned by the encoder.

As shown by several existing works, this approach can be transferred to cross-cultural emotion recognition, by treating the culture as the domain to be obscured. Liang et al., 2019 performed utterance-level classification of English and Chinese cultures. J. Zhao, Li, et al., 2019 used a DANN based on a LSTM for the winning entry into the AVEC2019 CES sub-challenge. The approach was further developed by H. Chen, Y. Deng, and Jiang, 2021, who added frame- and sequence level attention to the domain discriminator.

Thus, this thesis employs the DANN framework, considering one culture as the source domain $\mathcal{D}_{src}(x_i, y_i)$ and another as the target domain $\mathcal{D}_{tgt}(x_i)$.

The three components of the network are denoted as follows: The feature encoder $F$ with parameters $\theta_F$ yields embeddings $z$, the emotion regressor $E$ with parameters $\theta_E$ returns time-continuous predictions of arousal and valence $\hat{y}_a, \hat{y}_c$, and the culture discriminator $C$ with parameters $\theta_C$ gives a culture prediction $\hat{c}$.

$$\begin{aligned} z &= F(x; \theta_F) \\ (\hat{y}_a, \hat{y}_c) &= E(z; \theta_E) \\ \hat{c} &= C(z; \theta_C) \end{aligned} \tag{3.17}$$

$F$ can process features $x$ of shape $[B, L, D]$ from any combination of feature extraction networks in a seq-2-seq model. It encodes the features into a latent state of shape $[B, L, d_{emb}]$. For $E$, a simple fully connected network with 2 layers is chosen. It contains dropout regularisation and returns a sequence of arousal and valence label predictions $y$ of shape $[B, L, 2]$.

Based on previous works, a novel attention-based culture discriminator is proposed. Unlike J. Zhao, Li, et al., 2019, it does not use average pooling across the time steps to predict the culture. The work of H. Chen, Y. Deng, and Jiang, 2021 uses two losses in the discriminator, one for frame-level predictions, and one for sequence-level prediction with attention weights constructed from the former. Instead, herein a discriminator based on multi-head attention is used. The approach is inspired by Y. Zhao et al., 2021, who used a combination of LSTM with MHA for depression detection from speech.

The embedding for the final time step returned by the seq-2-seq model is used as the query $Q$, and the whole sequence serves as the key-value pairs $< K, V >$. Thus, the output is reduced to a single token which carries the attention-weighted contributions of all the time steps.

An example of the proposed DANN architecture, using a multi-layer RNN for encoding the representations and processing an early fusion of audiovisual features as described in section 3.2, is visualised in fig. 3.6. Now, the training method for this architecture will be presented.

$$\mathcal{L}_{culture} = \frac{1}{N} \sum_{n=1}^{N} c \log(\hat{c}) + (1 - c) \log(1 - \hat{c}). \tag{3.18}$$

The loss function for the culture discriminator is the binary cross-entropy loss, where the label $c$ is defined as 0 for the source culture and 1 for the target culture. Thus, the culture loss becomes as in eq. (3.18).

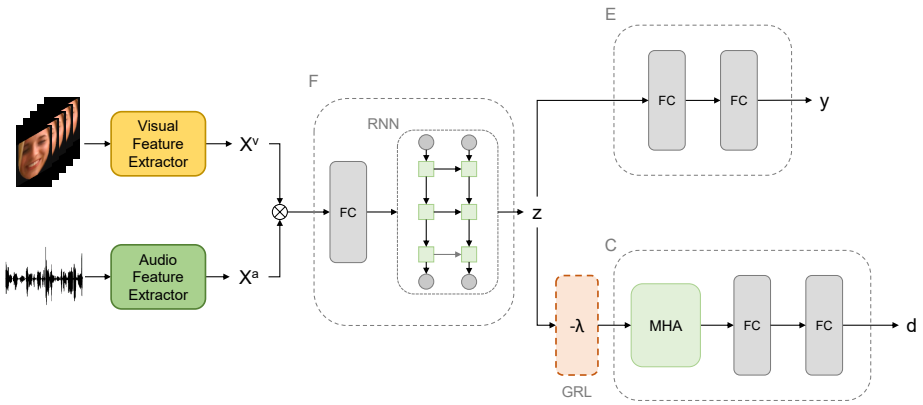



FIGURE 3.6: Domain Adversarial Neural Network (DANN) architecture used for cross-cultural emotion recognition. It receives input from feature extraction networks for audio and visual data. The DANN itself has 3 components: The feature encoder F, the emotion regressor E and the domain (culture) classifier c. A gradient reversal layer between F and C forces the representations z to become domain-agnostic.

$$\mathcal{L}_{emotion} = w_a \left(1 - CCC_a\right) + w_v \left(1 - CCC_v\right). \tag{3.19}$$

For the emotion regressor, the loss function is based on a weighted sum of arousal and valence CCC eq. (3.19), as described in section 3.3. Finally, the loss of the encoder $F$ is the weighted contribution of emotion and culture losses, with hyperparameters $\lambda_E$ and $\lambda_C$.

$$\mathcal{L}_{encoder} = \lambda_E \mathcal{L}_{emotion} - \lambda_C \mathcal{L}_{culture}. \tag{3.20}$$

In this approach, the regressor only has access to labelled samples from the source domain during training. It would, however, be easy to incorporate samples from the target domain if they are available. The culture discriminator is trained with the self-labelled samples from the source and target domains, although it could also be restricted to a single culture (H. Chen, Y. Deng, and Jiang, 2021).

Training adversarial networks is challenging due to the competing aspects of the loss. In the original paper on Generative Adversarial Networks (GANs) by Goodfellow et al., 2014, it was shown that the min-max game between the generator and discriminator parts theoretically leads to a stable end point where the discriminator is in a state of maximum confusion. However, in practice, training is often unstable, leading to sub-optimal results and issues like *mode collapse*, where the model fails to learn the data distribution and produces only one result (Athanasiadis, Hortal, and Asteriadis, 2019). For the cross-cultural DANN, this means that the contributions of the two objectives need to be balanced carefully. In the early stages, the culture prediction is expected to be quite inaccurate, thus, the backpropagated signal may confuse the network and lead to unstable training.

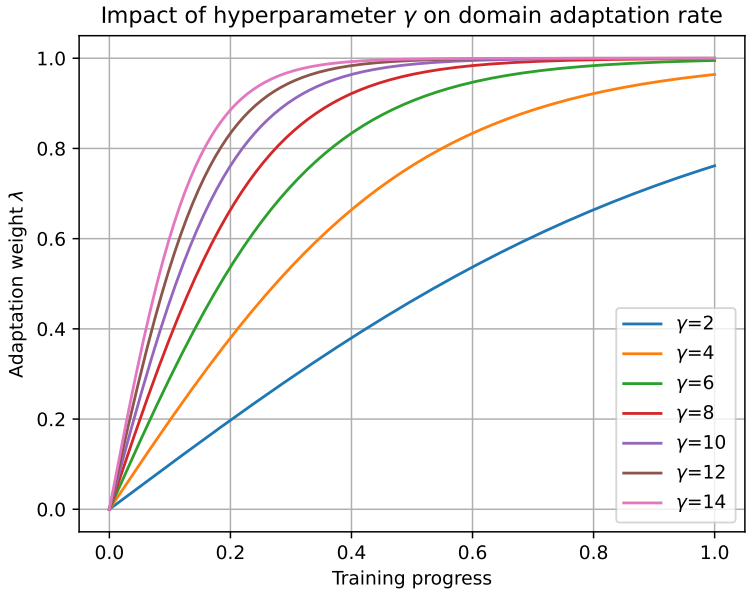



FIGURE 3.7: Visualisation of the domain adaptation rate $\tilde{\lambda}_C$ as a function of training progress $p$, for different values of hyperparameter $\gamma$
. The adaptation rate serves as scaling factor to the feature encoder updates and rises from 0 to 1 during training. Increasing values of $\gamma$ causes a faster rise.

Several modifications are made to the training process for the experiments in C-3 to address this issue.

First, in order to reduce the impact of initial instability, a dynamic weighting of $\lambda_c$ is used, based on Ganin et al., 2016. This increases the impact of domain adaptation over the course of the training $p$, from 0.0 to 1.0, with swiftness $\gamma$, as in eq. (3.21).

$$\tilde{\lambda}_C = \left( \frac{2}{1 + \exp(-p\gamma)} - 1 \right) \lambda_C. \tag{3.21}$$

Larger values of $\gamma$ lead to a faster rise of the domain adaptation weight, as illustrated in fig. 3.7.

Second, the network components are not optimised jointly but in different stages during each epoch, following J. Zhao, Li, et al., 2019, H. Chen, Y. Deng, and Jiang, 2021. $C$ is trained for $S_C$ iterations, then $E$ is trained for $S_E$ iterations. In the first stage, the encoder $F$ is updated with the adversarial information flowing back from $C$, in the second stage, it receives updates from both $E$ and $C$.

### 3.4.5 Chaining outputs for multi-task learning

For multi-label datasets, e. g., the HUME-VB dataset, the different tasks can be addressed by training individual classifiers or with multi-task learning on shared backbones. A modification of the regular multi-task network with parallel classification or regression heads is the *chaining* or *stacking* of outputs. Here, the model is structured such that the result of one task serves as input to one or more others. Thus, in a simple directed chain, the jth prediction is described by eq. (3.22).

$$\hat{y}^j = f\left(x, \hat{y}^1, ..., \hat{y}^{j-1}\right) \tag{3.22}$$

Classifier chains have shown promising results across a wide range of datasets and applications (Read et al., 2021). By adding predictions to the feature space, they can exploit inter-dependencies between labels. An example of such an approach used on continuous emotion recognition is Xin, Takamichi, and Saruwatari, 2022. An inherent issue of stacking predictors atop one another is that the number of possible combinations grows exponentially with the number of labels. The performance of the model depends on the order and interconnections, but a complete search is often computationally infeasible. One possible method is to train a regular multi-task classifier as baseline and chain the tasks in the order of their performance, starting with the strongest, but this may not be optimal (Read et al., 2021). While various other methods have been proposed, their comparison is not the focus of this thesis. Thus, when chaining or stacking is employed herein, greedy ordering by performance and/or heuristics based on assumptions on the label relations are used.

To summarise the methodology before continuing with the experimental part: In this thesis, deep features extracted directly from audio or images via CNNs or Transformers are used. They are combined via early or hybrid attention fusion, and processed via sequence-based models that make use of either recurrent neural networks or transformer encoders. The models are optimised in a multi-task learning manner, i. e., each affect dimension is predicted as its own task, with dynamic loss weighing to balance the tasks during training. The losses for continuous emotion prediction are primarily CCC based, and CCC is used as evaluation metric. Cross-cultural emotion recognition is solved via domain adaptation, using an adversarial network component that forces the internal representations to be invariant between cultures.

# Chapter 4

# Experiments and Results

Herein, the experiments conducted following the methodology of chapter 3 are described. Experiments are grouped into sections following different areas of research. A discussion of the experiments is given in the following chapter.

## 4.1 Multi-modal and Cross-Modal Emotion Recognition

This section describes experiments in valence and arousal prediction on the Aff-Wild2 Dataset, for contribution C–1, see section 1.3. The goal is to answer RQ–1 by comparing the impact of several key design choices on time- and value-continuous emotion recognition:

1. The complexity of the CNN backbones used as feature extractors
2. Using frozen pre-trained networks or end-to-end optimisation.
3. Using recurrence or self-attention for sequence modelling.
4. Using multi-modal versus uni-modal models and early versus hybrid fusion.

### 4.1.1 Dataset Preprocessing

Since Aff-Wild2's subjects are recorded in very noisy and diverse settings, preprocessing is applied to the data. For the visual modality, the face extraction method provided by the dataset's creators is used. This detects and crops out the faces in every frame, as well as aligning their orientation. Finally, the cropped and aligned faces are scaled to a size of 112x112 pixels. The audio is extracted via ffmpeg[1] and converted to 16 kHz mono, 16bit PCM encoding. From the audio feed, overlapping clips of 0.5s length, centred at the frame timestamps, are extracted.

$$t = \frac{N}{30} * T \tag{4.1}$$

As the Aff-Wild2 video frame rate is 30fps, consecutive frames will be very similar to each other, not containing much new information for the model. In order to cover a larger temporal context while avoiding high computational costs from long sequences, a dilated sampling method is used, i. e., creating a sequence of length $T$ by picking 1 in every N frames. This gives a context according to eq. (4.1).

[1] https://git.ffmpeg.org/ffmpeg.git

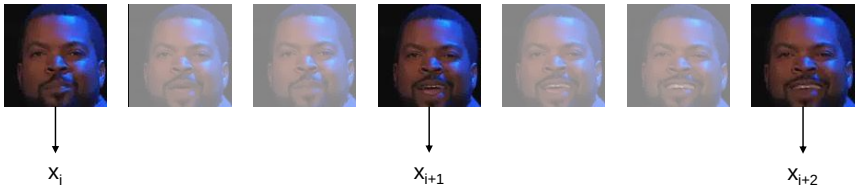



FIGURE 4.1: Dilated sampling example on frames of the Aff-Wild2 dataset. Every N-th frame is taken, so the sequence covers more context of the video.

The sampling strategy is only applied to the training set, while for the validation and test set, sequences from consecutive frames are used. In order to avoid discarding $(N-1)/N$ frames from the training data, an interleaved sampling method is also used. Dilated sampling is illustrated in fig. 4.1.

For data augmentation, the images are manipulated with random affine transformations, including rotation and translation. Saturation, brightness and contrast are modified with a jitter value of 0.2 to further challenge the model. The audio clips are augmented with Gaussian noise.

### 4.1.2 Models

The sequences of faces and audio clips are processed with deep models consisting of CNN feature extractors, followed by fusion and sequence processing modules, and finally prediction heads.

For the visual features, 2D CNNs are used. Two architectures, pre-trained on facial recognition tasks, are used. *FaceNet* (Schroff, Kalenichenko, and Philbin, 2015) is based on InceptionResNetv1 and trained on VGGFace2. It has 27M parameters and returns 512-dimensional embeddings. It is referred to simply as *Inception* below *MobileFaceNet* (S. Chen et al., 2018) is based on the residual bottlenecks of MobileNetv2 and trained on MS-Celeb-1M. It is designed as a lightweight architecture for embedded processing. Herein, a version of MobileFaceNet with 0.99M parameters producing 512-dimensional embeddings is used.

In the audio modality, a 1D-CNN network is deployed, based on an architecture proposed by J. Zhao, Mao, and L. Chen, 2019. It contains 4 local feature learning blocks composed of 1D convolutions and maxpooling, yielding a very lightweight (88K parameters) model. The model is pre-trained as a CNN-LSTM on RECOLA using the END2YOU toolkit, before discarding the recurrent layers and adding a final pooling layer to produce 128-dimensional embeddings. It will be referred to as *1D-AudioNet* here.

When multi-modal input is used, two different fusion strategies are used, as described in section 2.4. The first is a simple concatenation step for early fusion. In the second, modalities are combined via CMA blocks, as shown in fig. 3.3. Two complementary blocks, each attending one modality to the other, are used in parallel, and the end results are concatenated. This setup is inspired by (Tsai et al., 2019).

Fully connected or 1D convolutional layers help reduce the dimensionality of the features before they are passed to the sequence modules. The sequential modelling

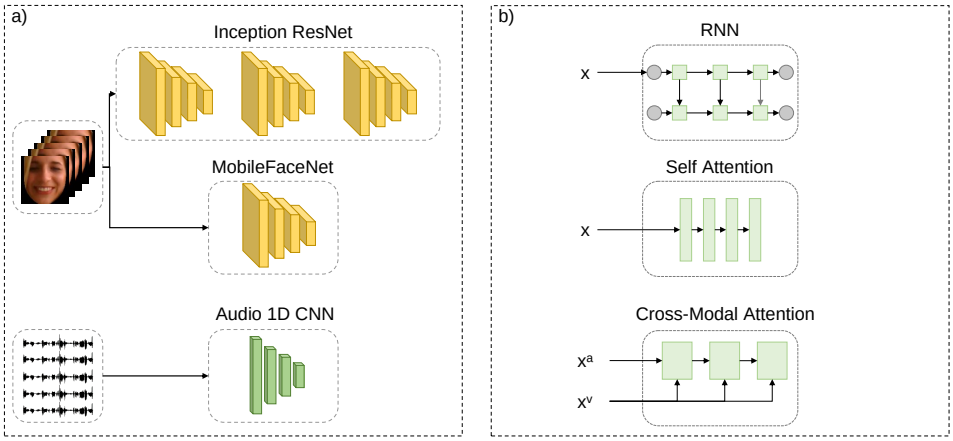



FIGURE 4.2: Overview of the model variants used for predicting valence and arousal on ABAW. a) Feature extractors, b) temporal components.

is done with three different blocks, as described in section 3.4.2: The first uses multi-layer RNNs with LSTM cells in both uni- and bidirectional configurations. The second employs self-attention stacks based on the Transformer encoder. The third uses the cross-modal blocks described above. Attention models include additional sinusoidal position embeddings, to help the model encode the sequence order. Feature extractors and temporal blocks are visualised in fig. 4.2.

### 4.1.3 Training

The models are implemented with PyTorch[2] and trained on Nvidia RTX3090 and A40 GPUs. The loss function is composed of a weighted sum of three terms. These include CCC loss, MSE loss, and a cross-entropy loss based on discretisations of the valence-arousal plane as described in section 3.3. AdamW is chosen as the optimiser. Cosine annealing with warm restarts is used as the learning rate scheduler, with its restart period set to 200 steps.

Data is batched with 64 samples each, and a sequence length of 16 frames.

An extensive hyperparameter optimisation is performed to find the best model configurations. The search space is listed in In order to speed up the process and avoid wasting computational resources, the Ray Tune framework[3] is used to run trials across multiple GPUs. The ASHA scheduler is employed to discover promising configurations, while those that do not perform well are stopped early.

### 4.1.4 Results

In this section, the results of the model training are presented. First, a set of experiments with frozen feature extractor CNNs is described. Second, another set of experiments that was run with end-to-end learning is summarised.

The results are discussed in section 5.1.

[2]https://pytorch.org/
[3]https://docs.ray.io/en/latest/tune/index.html

TABLE 4.1: Search space of the hyperparameters used for training. Since the potential number of combinations is quite large, trial scheduling with early stopping is used via the Ray Tune framework.

| **Hyperparameter** | **Value Range** |
|---|---|
| *General parameters* | |
| $n_{layers}$ | [1, 5] |
| $d_{model}$ | 64, 128, 256 |
| activation | GELU, SELU |
| dropout | [0.1, 0.6] |
| learning rate | $[10^{-5}, 10^{-2}]$ |
| weight decay | $[10^{-3}, 10^{-1}]$ |
| $\lambda_{mse}$ | [0.0, 1.0] |
| $\lambda_{ce}$ | [0.0, 1.0] |
| *Attention Models* | |
| $d_{feedforward}$ | 64, 128, 256 |
| $n_{heads}$ | 2, 4, 8 |
| *Cross-Modal Attention* | |
| $n_{layers}^{V \longrightarrow A}$ | [1,5] |
| $n_{layers}^{A \longrightarrow V}$ | [1,5] |
| *Recurrent Models* | |
| context | unidirectional, bidirectional |
| $n_{layers}$ | [1,5] |
| $d_{hidden}$ | 64, 128, 256 |

**Experiments with frozen feature extraction networks.**

Three sets of experiments are conducted: audio-only, visual-only, and audiovisual (abbreviated as AV). Experiments with a visual component use the large Inception net and the smaller MobileFaceNet.

The validation set results of the best-performing models with frozen CNNs are shown in table 4.2.

The best scores were obtained by the multi-modal models. For valence, the top CCC score was 0.393, achieved by the model using cross-modal fusion and Inception as visual feature extractor. The top arousal CCC was obtained using early fusion of MobileFaceNet and 1D-AudioNet features, followed by self-attention. Finally, the best averaged score by a single model was from early fusion of Inception and 1D-AudioNet followed by RNN, with $CCC = 0.413$.

The audio-only models performed much worse on valence, with top scores of 0.094 for the recurrent models and 0.076 for the attention-based architecture.

On arousal, the audio models performed better, the RNN model achieved $CCC = 0.233$, while the self-attention architecture achieved $CCC = 0.317$.

The uni-modal vision models achieved top CCC scores of 0.285 and 0.357 on valence and arousal using MobileFaceNet and RNN, as well as 0.324 and 0.414 using MobileFaceNet and self-attention respectively. The results from models based on Inception features were slightly lower than those with MobileFaceNet on valence (0.277 with RNNs, 0.318 with attention).

TABLE 4.2: Validation set results (CCC ↑) on the Aff-wild2 corpus from the 2022 ABAW challenge, adapted from (Karas, Tellamekala, et al., 2022). Shown are the scores of models using frozen feature extraction CNNs, with uni-modal (RNN and self-attention) and multi-modal (early fusion and cross-modal attention) architectures. Also shown is the baseline model from ABAW3 (Kollias, Tzirakis, Baird, et al., 2023)

| **Method** | **Visual CNN** | **Audio CNN** | **Valence** CCC | **Arousal** CCC | **Avg.** CCC |
|---|---|---|---|---|---|
| Baseline (ABAW3) | ResNet50 | - | .310 | .170 | .24 |
| *Recurrent Models (RNNs)* | | | | | |
| Audio-RNN | - | 1D-AudioNet | .094 | .233 | .163 |
| Visual-RNN | Inception | - | .277 | .188 | .233 |
| Visual-RNN | MobileFaceNet | - | .285 | .357 | .321 |
| AV-RNN | Inception | 1D-AudioNet | **.339** | .486 | **.413** |
| AV-RNN | MobileFaceNet | 1D-AudioNet | .319 | .436 | .378 |
| *Self-Attention (SA) Models* | | | | | |
| Audio-SA | - | 1D-AudioNet | .076 | .317 | .197 |
| Visual-SA | Inception | - | .318 | .203 | .261 |
| Visual-SA | MobileFaceNet | - | .324 | .414 | .369 |
| AV-SA | Inception | 1D-AudioNet | **.344** | .404 | .374 |
| AV-SA | MobileFaceNet | 1D-AudioNet | .248 | **.529** | **.389** |
| *Cross-Modal Attention (CMA) Models* | | | | | |
| AV-CMA | Inception | 1D-AudioNet | **.393** | .363 | .378 |
| AV-CMA | MobileFaceNet | 1D-AudioNet | .324 | **.460** | **.392** |

Inception performed much worse than MobileFaceNet on arousal ($CCC = 0.188$ with recurrence, $CCC = 0.203$ with attention).

**Experiments using end-to-end training**

Following the experiments with frozen feature extractors, the CNN layers were unlocked, allowing the fine-tuning on the Aff-Wild2 data. Based on the results of the ablation experiments above, it was decided to abandon Inception due to its worse performance compared to MobileFaceNet. This also had the advantage of saving computational resources and allowed for faster iteration.

The results of the best-performing models trained end-to-end with different sequence-to-sequence architectures are listed in table 4.3. The best valence score of $CCC =$

TABLE 4.3: Validation results in CCC ↑, evaluated on the validation set of Aff-Wild2 in ABAW 2022 and adapted from Karas, Tellamekala, et al., 2022. Reported results are for the best multi-modal models trained end-to-end with MobileFaceNet as visual encoder and 1D CNN pretrained on RECOLA as audio encoder, and using RNN, self-attention and cross-modal attention for sequence modelling.

| **Method** | **Valence** | **Arousal** | **Avg.** |
|---|---|---|---|
| E2E-AV-RNN | .361 | **.551** | **.456** |
| E2E-AV-SA | .380 | .520 | .450 |
| E2E-AV-CMA | **.388** | .492 | .440 |

0.388 was obtained using cross-modal attention fusion. The best score for arousal, as well as the highest overall score for a model, were delivered by the early fusion with RNN model, with 0.551 and 0.456, respectively.

TABLE 4.4: Hyperparameter configurations for the best performing models on the validation set of the Aff-Wild2 dataset, obtained using randomly sampled grid search with ASHA scheduler. Models are trained end-to-end with recurrent neural network, self-attention, and cross-modal attention networks, respectively. Search results adapted from Karas, Tellamekala, et al., 2022.

| **Hyper-Parameter** | **E2E Models** | | |
|---|---|---|---|
| | AV-RNN | AV-SA | AV-CMA |
| $n_{layers}$ | 1 | 3 | 4 |
| $d_{model}$ | 64 | 64 | 256 |
| activation | SELU | GELU | GELU |
| dropout | 0.5 | 0.5 | 0.6 |
| learning rate | 0.0002 | 0.002 | 0.0001 |
| weight decay | 0.023 | 0.008 | 0.06 |
| $\lambda_{mse}$ | 0.84 | 0.78 | 0.18 |
| $\lambda_{ce}$ | 0.88 | 0.27 | 0.76 |
| $d_{feedforward}$ | - | 256 | 256 |
| $n_{heads}$ | - | 8 | 4 |
| $n_{layers}^{V \longrightarrow A}$ | - | - | 3 |
| $n_{layers}^{A \longrightarrow V}$ | - | - | 1 |
| Context aggregation | uni | - | - |
| $d_{hidden}$ | 64 | - | - |

As the best-performing end-to-end models were obtained from extensive hyperparameter tuning, their configurations are listed in table 4.4.

**Test set results**

The test set labels of Aff-Wild2 are hidden due to the data being used in a competition setting. Therefore, extensive evaluations were not possible. Instead, test set results were obtained by entering the third ABAW challenge (Kollias, 2022). Due to the limited number of submissions, only the best-performing models were chosen, which were all multi-modal and trained end-to-end. An ensemble of those models was also constructed. The results are listed in table 4.5.

The recurrent, self-attention and cross-modal attention models achieved averaged CCC scores of 0.378, 0.386 and 0.343, respectively. The best overall test set performance with the approach presented here was achieved with an ensemble of the three best end-to-end models. Top CCC scores for valence and arousal were 0.418 and 0.407 respectively, leading to an aggregate score of 0.413.

TABLE 4.5: Test results in CCC↑, evaluated on the test set of the Aff-wild2 corpus from the ABAW2022 challenge. Reported here are the results from Karas, Tellamekala, et al., 2022, based on the three best-performing models on the validation set, as well as their ensemble obtained by averaging the models' predictions.

| **Model** | **Valence** CCC | **Arousal** CCC | **Avg.** CCC |
|---|---|---|---|
| E2E-AV-RNN | .376 | .380 | .378 |
| E2E-AV-SA | .396 | .376 | .386 |
| E2E-AV-CMA | .327 | .359 | .343 |
| E2E-AV-Ensemble | **.418** | **.407** | **.413** |

## 4.2 Cross-Cultural Audiovisual Emotion Recognition

This section describes experiments for continuous arousal and valence prediction on the SEWA dataset. Data from multiple cultures is used to study how domain transfer between them affects the models' performance. Various architectures and training procedures are employed, including ones aiming for domain adaptation to improve generalisation to new cultures. Together with the discussion in section 5.3, these experiments form contribution C–3, aimed at research questions RQ–2 and RQ–3.

### 4.2.1 Dataset preprocessing

Videos from all six cultures present in SEWA (German, Hungarian, Chinese, English, Serbian, Greek) are preprocessed. For German, Hungarian, and Chinese, the partitioning and annotations of the AVEC2019 corpus are re-used. For English, Serbian, and Greek, no such annotations were available. Those cultures are used exclusively as unlabelled training data for domain adaptation.

Audio is extracted from the video clips and converted to 16kHz mono using ffmpeg. Since SEWA consists of dyadic conversations, the timestamps corresponding to speaker and interlocutor turns are also extracted (for the AVEC2019 videos they are already provided with the annotations). The faces of the subjects are extracted from the videos using OpenFace toolkit[4] (Baltrusaitis et al., 2018), the crops are similarity aligned and stored as jpg files with resolution 112x112. From the raw data features are generated using two pre-trained transformer models. Audio is processed with a model based on WAV2VEC2 and fine-tuned on MSP-Podcast [5], yielding sequences of 1024 dimensional vectors. The faces are passed through a vision transformer (ViT-Base) fine-tuned on the FER-2013 dataset for facial emotion recognition, and the CLS token embeddings are used as 768 dimensional features.

### 4.2.2 Models

All models used herein are designed with early fusion and a sequence-to-sequence encoder for prediction of arousal and valence.

The group of models trained for domain adaptation is DANN based, see fig. 3.6. They consist of feature extraction networks for end-to-end training (unless pre-extracted

[4]https://github.com/TadasBaltrusaitis/OpenFace
[5]https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

features a used), a feature encoder that can be based on RNNs or multi-head attention and returns an embedded sequence, an emotion regressor to predict arousal and valence and a culture domain classifier connected through a gradient reversal layer.

### 4.2.3 Training

All models were trained on a mixture of RTX 3090 and A40 GPUs using PyTorch framework. Due to the comparatively small size of the datasets and using compact models (and pre-extracting features for the much larger transformers), training proceeded quickly, with an epoch typically lasting less than 3 minutes, and models taking approximately 35 epochs to train.

TABLE 4.6: Hyperparameter search space for preliminary experiments on cross-cultural affect recognition on the SEWA dataset

| **Hyperparameter** | **Value Range** |
|---|---|
| random seed | [1-5] |
| batch size | [4-16] |
| sequence length | [10-150] |
| Model architecture | |
| activation | [GELU, SELU] |
| RNN cell | [GRU, LSTM] |
| RNN direction | [uni, bi] |
| $d_{embedding}$ | [32-256] |
| $d_{feedforward}$ | [32-512] |
| $n_{layers}$ | [1-4] |
| Optimisation | |
| optimiser | [Adam, AdamW, SGD] |
| weight decay | [$10^{-5}$-$10^{-3}$] |
| learning rate | [$10^{-4}$-$10^{-3}$] |
| loss weighting | [mean, dwa, rruw, druw] |
| learning rate schedule | [cosine annealing, linear, none] |

Preliminary experiments were conducted to find favourable hyperparameter configurations. Randomly sampled grid searches were performed, with parameters drawn from the search space described in table 4.6.

Based on the preliminary experiments, AdamW was chosen as the optimiser, with a weight decay of $1e-4$. The learning rate was set to $1e-4$.

Each model was trained with early stopping set to abort the training run if the validation performance (measured by the average CCC of arousal and valence on the validation set) did not improve by at least 0.01 for 8 epochs. At the end of training, whether by early stopping or by reaching the maximum number of epochs, the best model checkpoint was restored prior to evaluation. The batch size in all experiments was chosen as 8, and the sequence length was set to 100 (equivalent to 10s).

As expected, bidirectional RNNs were found to be superior to unidirectional ones. The results of LSTMS and GRUs in initial experiments were similar, thus GRU was chosen over LSTM due to it being more computationally efficient with its lower parameter count. The hidden size was set to 128, which results in 256-dimensional embeddings. For the depth of the seq-2-seq model, having 2-3 layers in the encoder was found to be optimal in combination with CNN-based features.

Interestingly, early experiments also showed that the activation function had considerable influence on the results. Networks trained with *Gaussion Exponential Linear Unit (GELU)* would consistently outperform their counterparts with *Scaled Exponential Linear Unit (SELU)* by a small margin (usually up to 0.03 CCC). Therefore, further experiments were restricted to using GELU activations.

While the audio transformer features are known to deliver strong results for valence-arousal prediction, the features extracted from ViT showed less promising performance in preliminary experiments when used on the SEWA data. Hence, another fine-tuning was implemented, this time on a subset of the Aff-Wild2 dataset. A regression head was added to the ViT base model and trained to predict valence and arousal. The learning rate for the transformer was set to $5e-5$. AdamW was used as the optimiser, and a learning rate schedule of cosine annealing with warmup and multiple restarts was set. CCC loss i. e., $1-CCC$ was chosen as the objective function, and Dynamic Weight Averaging or Dynamic Restrained Uncertainty Weighing were used to balance the training for valence and arousal prediction. Experiments with these features also failed to show an improvement over the MobileFaceNet activations, thus the following analysis focused on the 2D-CNN features.

Each model configuration was trained 5 times with varying random seeds to account for the effect of different weights initialisation. Afterwards, the predictions of the runs were averaged in an emsemble and the CCC computed again.

#### Baseline training process

In order to have a comparison for the adversarial models, a baseline set of experiments is needed. Thus, a group of models is trained fully supervised on a single source culture, and evaluated on the validation and test cultures. These baseline models are simply combinations of feature extractor CNNS and recurrent or multi-head attention based sequence-to-sequence models (effectively they are an ablation of the DANN models, with the gradient reversal layer and culture classifier head removed).

#### DANN training process

For the DANN models, the training procedure described in section 3.4.4 is used. In terms of training speed, $\lambda_E$ is set to 1.0, and $\lambda_C$ varies from $1.0-2.0$, while the impact of culture classification on the feature encoder is set by an adjustable parameter of the gradient reversal layer, according to eq. (3.21). The parameter $\gamma$ controlling the speed of the ramp-up is set to 8.0.

Two datasets are constructed. The emotion dataset $D_E$ is used for training the affect recognition task, analogous to the baseline models. It contains all labelled training samples from the source culture. The culture dataset $D_C$ is used for the adversarial domain adaptation task. It contains all training samples from the source and target cultures with the binary domain label $d$.

A training epoch consists of a loop which draws $S_C$ batches from the culture dataset $D_C$ and updates the classifier $C$ and feature encoder $F$, followed by drawing $S_E$ batches from the emotion dataset and updating the emotion regression network $E$ and feature encoder $F$. This interleaved training is designed to stabilise training, based on the findings of the related works (J. Zhao, Li, et al., 2019; H. Chen, Y. Deng, and Jiang, 2021). The epoch ends when the culture dataset has been iterated once.

If the emotion dataset, which is by definition smaller, runs out of samples during the epoch, its iterator can be restarted. This guarantees that the model receives the same ratio of updates for the two tasks, irrespective of the size of the training sets for different cultures.

An optional warmup method is also implemented, which if enabled restricts training to the culture task for the first few epochs. While the warmup is running, since the emotion task is not yet being optimised, early stopping is of course disabled.

### 4.2.4 Results

Here the results of the cross-cultural affect recognition experiments are described. First, the results of the baseline experiments is are presented. Then the results of DANNs trained on various culture combinations are shown in relation to the corresponding baselines.

Each set of experiments produces a large number of results (a total of 10 arousal and valence CCC scores on the validation and test set cultures). These are presented in the following order:

- validation results before test results
- arousal CCC before valence CCC
- German before Hungarian before Chinese results

Experiments that only differ in their source culture are grouped together, with German-trained models being described before Hungarian-trained ones.

#### Multi-modal baseline results

The results for audiovisual models trained on the German and Hungarian data respectively is shown in table 4.7. The best achieved CCC scores for models using early fusion of CNN features followed by either GRUs or Transformer encoder with GRU for sequence modelling are depicted.

Beginning with the models trained on German data, the best scores are 0.767 and 0.55 for arousal, as well as 0.714 and 0.536 for valence, on German and Hungarian validation set, respectively. On the test set, consisting of German, Hungarian, and Chinese data, the baseline models achieve top arousal scores of 0.621, 0.516, 0.391 and top valence scores of 0.642, 0.443, 0.452 respectively.

The models trained on Hungarian data achieve validation scores of 0.556 and 0.437 for arousal, and 0.531 and 0.488 for valence, on German and Hungarian respectively. On the test set, top arousal scores are 0.493, 0.521, and 0.477, while top valence scores are 0.48, 0.518 and 0.34, on German, Hungarian and Chinese data respectively.

#### Uni-modal ablation results

In order to gauge the impact that multi-modal fusion has on the baseline models' performance, an ablation study is performed. First, the audio information is removed, forcing the model to learn arousal and valence from face only. The results are depicted in table 4.8.

The vision models trained on German data with MobileFaceNet CNN features achieve top arousal CCC scores of 0.684 and 0.54 on German and Hungarian data, as well

TABLE 4.7: Validation and test set results of audio-visual baseline models trained on German and Hungarian cultures of the SEWA dataset as source domains, respectively. Shown are the CCC scores for arousal and valence, obtained by the best performing aggregated model runs using CNN feature extractors and either RNN or self-attention transformer stack and RNN as seq-2-seq encoder.

| **Source** | **Features** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|
| | | | DE | HU | DE | HU | CN |
| | | *Arousal* | | | | | |
| DE | MFN + 1D-CNN | RNN | **.767** | **.55** | .6196 | .516 | .391 |
| DE | MFN + 1D-CNN | SA + RNN | .741 | .534 | .621 | .507 | .345 |
| DE | MFN + WAV2VEC2 | RNN | 0.699 | 0.522 | **0.641** | **0.572** | 0.352 |
| HU | MFN + 1D-CNN | RNN | .556 | .421 | .493 | .521 | **.477** |
| HU | MFN + 1D-CNN | SA + RNN | .544 | .437 | .47 | .498 | .411 |
| HU | MFN + WAV2VEC2 | RNN | .633 | .476 | .552 | .559 | .411 |
| | | *Valence* | | | | | |
| DE | MFN + 1D-CNN | RNN | .689 | .376 | .576 | .423 | **.46** |
| DE | MFN + 1D-CNN | SA + RNN | .714 | **.526** | .642 | .443 | .452 |
| DE | MFN + WAV2VEC2 | RNN | **.74** | .381 | **.712** | .462 | 0.457 |
| HU | MFN + 1D-CNN | RNN | .531 | .454 | .477 | .431 | .256 |
| HU | MFN + 1D-CNN | SA + RNN | .5 | .488 | .48 | .518 | .34 |
| HU | MFN + WAV2VEC2 | RNN | .675 | .436 | .648 | **.541** | .38 |

TABLE 4.8: Validation and test set results of visual baseline models trained on German and Hungarian cultures of the SEWA dataset as source domains. Shown are the CCC scores for arousal and valence, obtained by the best performing aggregated model runs using CNN as feature extractors and either RNN or self-attention transformer stack and RNN as seq-2-seq encoder.

| **Source** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|
| | | DE | HU | DE | HU | CN |
| | *Arousal* | | | | | |
| DE | CNN + RNN | **.684** | **.54** | .577 | .496 | .379 |
| DE | CNN + SA + RNN | .672 | .484 | **.584** | .502 | .454 |
| HU | CNN + RNN | .575 | .46 | .52 | .506 | .495 |
| HU | CNN + SA + RNN | .583 | .5 | .521 | **.508** | **.517** |
| | *Valence* | | | | | |
| DE | CNN + RNN | **.699** | .441 | **.654** | .435 | **.429** |
| DE | CNN + SA + RNN | .697 | .41 | .649 | .443 | .418 |
| HU | CNN + RNN | .639 | .434 | .619 | .522 | .334 |
| HU | CNN + SA + RNN | .581 | **.442** | .586 | **.529** | .328 |

as 0.584, 0.502 and 0.454 on German, Hungarian and Chinese test sets respectively. For valence, the top scores are 0.699 and 0.441 on German and Hungarian. The test scores are 0.654, 0.443 and 0.429 CCC.

The vision models trained with Hungarian data achieve top arousal CCC scores of 0.583 and 0.5 on the validation data, and 0.521, 0.508 and 0.517 on the test data. The

best valence results are CCC scores of 0.639 and 0.442 for validation and 0.619, 0.529 and 0.334 on test.

TABLE 4.9: Validation and test set results of audio baseline models trained on the German culture of the SEWA dataset as source domain. Shown are the CCC scores for arousal and valence, obtained by the best performing aggregated model runs using either a pre-trained 1D-CNN or a transformer as feature extractors and either RNN or self-attention transformer stack and RNN as seq-2-seq encoder.

| **Source** | **Feature** | **Encoder** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|
| | | | DE | HU | DE | HU | CN |
| | | *Arousal* | | | | | |
| DE | 1D-CNN | RNN | .441 | .27 | .359 | .25 | .121 |
| DE | 1D-CNN | SA + RNN | .288 | .155 | .19 | .14 | .06 |
| DE | WAV2VEC2 | RNN | .58 | .394 | .519 | .45 | .243 |
| DE | WAV2VEC2 | SA + RNN | **.61** | .395 | **.527** | **.454** | **.325** |
| HU | WAV2VEC2 | RNN | .4 | **.41** | .43 | .428 | .139 |
| HU | WAV2VEC2 | SA + RNN | .35 | .407 | .423 | .418 | .145 |
| | | *Valence* | | | | | |
| DE | CNN | RNN | .383 | .0 | .332 | .15 | .21 |
| DE | CNN | SA + RNN | .237 | .0 | .178 | .02 | .164 |
| DE | WAV2VEC2 | RNN | .58 | **.329** | .563 | .34 | .332 |
| DE | WAV2VEC2 | SA + RNN | **.617** | .24 | **.581** | .348 | **.405** |
| HU | WAV2VEC2 | RNN | .397 | .17 | .383 | .357 | .181 |
| HU | WAV2VEC2 | SA + RNN | .328 | .148 | .355 | **.403** | .141 |

Second, the models are denied the visual domain, and forced to learn only from the subjects' voices. The results are listed in table 4.9.

The models trained only on German audio data with the 1D CNN feature extractor achieve top arousal scores of 0.441 and 0.27 on the German and Hungarian validation sets, respectively. The tests scores are 0.359, 0.25 and 0.121 on German, Hungarian and Chinese videos respectively. For valence, the top validation scores are 0.383 and 0.0 CCC, and the top test scores are 0.332, 0.15 and 0.21 CCC, in the culture order above.

When WAV2VEC2 is used as feature extractor instead, a considerable improvement in performance becomes apparent.

The best models trained on German data achieve validation scores of $CCC = 0.61$ and $CCC = 0.617$ on German as well as 0.395 and 0.329 on Hungarian for arousal and valence, respectively. On the test set, this set of models achieves top arousal and valence scores of $CCC = 0.527$ and $CCC = 0.581$ on German, 0.454 and 0.348 on Hungarian, and 0.325 and 0.405 for Chinese, respectively.

Models trained on Hungarian yield arousal scores of 0.4 and 0.41 on German and Hungarian validation data, as well as CCCs of 0.43, 0.428 and 0.145 on the test sets, respectively. For valence, the top scores are 0.397 and 0.17 on German and Hungarian validation, and 0.383, 0.403 and 0.181 on the German, Hungarian, and Chinese test sets.

**Reducing the labels available for training**

In order to answer RQ–3, experiments need to be conducted with different amounts of labelled vs unlabelled data available to the models. Besides adding more target domain data, one simple way to achieve this is to restrict the amount of source domain data.

TABLE 4.10: Validation and test set results of the multi-modal baseline models, trained respectively on German and Hungarian cultures of the SEWA dataset. Shown are the aggregated results of arousal and valence CCC, from the best-performing models, when the amount of training data is decreased to 75%, 50% and 25%.

| **Source** | **Labels** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|
| | | | DE | HU | DE | HU | CN |
| | | *Arousal* | | | | | |
| DE | 0.75 | CNN + RNN | **.729** | **.53** | .615 | .473 | .337 |
| DE | 0.50 | CNN + RNN | .719 | .497 | .621 | .465 | .376 |
| DE | 0.25 | CNN + RNN | .717 | .519 | **.619** | .474 | .438 |
| DE | 0.75 | CNN + SA + RNN | .716 | .52 | .597 | .472 | .501 |
| DE | 0.50 | CNN + SA + RNN | .717 | .495 | .601 | .486 | .449 |
| DE | 0.25 | CNN + SA + RNN | .687 | .482 | .556 | .416 | .44 |
| HU | 0.75 | CNN + RNN | .642 | .45 | .539 | .531 | **.527** |
| HU | 0.50 | CNN + RNN | .618 | .431 | .528 | **.534** | .443 |
| HU | 0.25 | CNN RNN | .597 | .388 | .469 | .529 | .47 |
| HU | 0.50 | CNN + SA + RNN | .482 | .321 | .412 | .474 | .48 |
| HU | 0.25 | CNN + SA + RNN | .418 | .283 | .297 | .386 | .47 |
| | | *Valence* | | | | | |
| DE | 0.75 | CNN + RNN | **.735** | .371 | .654 | .386 | **.493** |
| DE | 0.50 | CNN + RNN | .714 | .4 | **.671** | .387 | .44 |
| DE | 0.25 | CNN + RNN | .692 | .4 | .66 | .371 | .439 |
| DE | 0.75 | CNN + SA + RNN | .719 | **.441** | .662 | .401 | .483 |
| DE | 0.50 | CNN + SA + RNN | .709 | .372 | .632 | .361 | .47 |
| DE | 0.25 | CNN + SA + RNN | .66 | .346 | .569 | .306 | .439 |
| HU | 0.75 | CNN + RNN | .613 | 0.414 | 0.603 | **0.499** | 0.285 |
| HU | 0.50 | CNN + RNN | .623 | .419 | .612 | .498 | .296 |
| HU | 0.25 | CNN + RNN | .608 | .336 | .576 | .484 | .3 |
| HU | 0.50 | CNN + SA + RNN | .437 | .368 | .404 | .374 | .198 |
| HU | 0.25 | CNN + SA + RNN | .38 | .299 | .371 | .323 | .2 |

Therefore, in order to gauge the impact the amount of labelled data available has on the model performance, the source training datasets were successively reduced to 75%, 50%, 25% of the original data. Sampling was performed randomly across the dataset, but the random seed was fixed to maintain comparability between the different training runs. For the audio-visual models using CNN feature extractors, the results are depicted in table 4.10.

These results will be used below, for comparison with DANNs trained on an identical amount of source labels.

**Domain Adversarial Neural Network results**

The results obtained with the domain adversarial neural networks are now presented. The experiments are organised by the combinations of labelled source culture (German or Hungarian) and unlabelled target culture (German, Hungarian, English, Serbian and Greek). For reference, the results of the baseline model sharing the same source culture will be repeated in the following tables. In order to maintain a fair comparison, the same feature extractors and splits of labelled data are used.

TABLE 4.11: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with German as source domain. CNN feature extractors are used and the full labelled data of the source culture is processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| DE | - | 1.00 | CNN-BASE | **767** | **.55** | .627 | .516 | **.471** |
| DE | HU | 1.00 | CNN-DANN | .758 | .526 | .651 | **.546** | .422 |
| DE | EN | 1.00 | CNN-DANN | .76 | .53 | .651 | .532 | .415 |
| DE | SR | 1.00 | CNN-DANN | .75 | .534 | .659 | .539 | .395 |
| DE | GR | 1.00 | CNN-DANN | .762 | .526 | **.66** | .532 | .411 |
| | | | *Valence* | | | | | |
| DE | - | 1.00 | CNN-BASE | .714 | **.526** | .642 | .443 | .46 |
| DE | HU | 1.00 | CNN-DANN | .756 | .414 | .688 | .443 | .452 |
| DE | EN | 1.00 | CNN-DANN | **.769** | .389 | .672 | .434 | .5 |
| DE | SR | 1.00 | CNN-DANN | .741 | .395 | .679 | .454 | **.484** |
| DE | GR | 1.00 | CNN-DANN | .767 | . 384 | .675 | **.455** | .475 |
| | | | *Average Arousal-Valence* | | | | | |
| DE | - | 1.00 | CNN-BASE | .741 | **.538** | .635 | .479 | **.466** |
| DE | HU | 1.00 | CNN-DANN | .757 | .47 | **.67** | .495 | .437 |
| DE | EN | 1.00 | CNN-DANN | **.765** | .46 | .662 | .483 | .458 |
| DE | SR | 1.00 | CNN-DANN | .746 | .465 | .669 | **.497** | .44 |
| DE | GR | 1.00 | CNN-DANN | .765 | .455 | .668 | .494 | .443 |

The top results of DANNs trained with the full set of German videos are summarised in table 4.11. For arousal, the top scores of DANNs are 0.76 CCC and 0.534 CCC on German and Hungarian respectively, which is below the baseline score. On the test set, the top DANN arousal score for German is 0.66 and the top score for Hungarian is 0.546, both surpassing the baseline. On Chinese, the top score is 0.422, which is below the baseline score of 0.471.

For valence, the best DANN achieves a top CC of 0.767 on German, surpassing the baseline. With Hungarian however, the top result is 0.414, which is worse than the baseline of 0.526. On the test set, the top scores are 0.688, 0.455 and 0.484 for German, Hungarian and Chinese respectively, ass surpassing the baseline.

Finally, looking at the average of arousal and valence scores per culture, the top score for German validation is 0.765 CCC, and the top score for Hungarian is 0.47 CCC,

which is above and below the baseline, respectively. On the test set, the top scores on German, Hungarian and Chinese are 0.669, 0.495 and 0.458 CCC respectively, with only Chinese falling below the baseline result of 0.466.

TABLE 4.12: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with German as source domain. CNN feature extractors are used and the full labelled data of the source culture is processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| HU | - | 1.00 | CNN-BASE | .556 | .437 | .493 | .521 | .512 |
| HU | DE | 1.00 | CNN-DANN | .644 | **.512** | .585 | .547 | **.561** |
| HU | EN | 1.00 | CNN-DANN | .623 | .464 | .566 | .554 | .469 |
| HU | SR | 1.00 | CNN-DANN | **.654** | .447 | .594 | **.563** | .461 |
| HU | GR | 1.00 | CNN-DANN | .646 | .464 | **.595** | .553 | .55 |
| | | | *Valence* | | | | | |
| HU | - | 1.00 | CNN-BASE | .531 | **.488** | .48 | .518 | .34 |
| HU | DE | 1.00 | CNN-DANN | .616 | .452 | .589 | .542 | **.358** |
| HU | EN | 1.00 | CNN-DANN | .609 | .338 | .554 | .534 | .354 |
| HU | SR | 1.00 | CNN-DANN | **.636** | .446 | .63 | .537 | .33 |
| HU | GR | 1.00 | CNN-DANN | .612 | .422 | **.638** | **.542** | .327 |
| | | | *Average Arousal-Valence* | | | | | |
| HU | - | 1.00 | CNN-BASE | .544 | .463 | .486 | .52 | .426 |
| HU | DE | 1.00 | CNN-DANN | .63 | .482 | .587 | .545 | **.456** |
| HU | EN | 1.00 | CNN-DANN | .616 | **.542** | .56 | .544 | .412 |
| HU | SR | 1.00 | CNN-DANN | **.645** | .447 | .612 | **.55** | .396 |
| HU | GR | 1.00 | CNN-DANN | .634 | .443 | **.617** | .548 | .439 |

In table 4.12, the results of DANNs trained with the full Hungarian data as source culture and using CNN feature extractors are summarised.

For arousal the top validation scores are 0.654 CCC and 0.512 CCC for German and Hungarian respectively, both improving on the baseline. On the test set, the top scores are 0.595, 0.563 and 0.561 on German, Hungarian and Chinese, all surpassing the baseline.

Regarding valence, the top validation score for German is 0.636, and 0.452 for Hungarian, which is above and below the baseline, respectively. On the test set, the DANNs achieve top CCC scores of 0.638, 0.542, and 0.358, all surpassing the baseline.

Looking at the average of the arousal and valence results, the DANNs achieve top validation scores of 0.645 and 0.542 on German and Hungarian respectively, both surpassing the baseline. The top scores on the test set are 0.617, 0.55 and 0.456, for German, Hungarian, and Chinese, again all surpassing the baseline.

TABLE 4.13: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with German as source domain. CNN feature extractors are used and 75% of the labelled samples in the source domain are processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| DE | - | 0.75 | CNN-BASE | .729 | .53 | .615 | .473 | .501 |
| DE | HU | 0.75 | CNN-DANN | .733 | **.53** | .621 | **.53** | .44 |
| DE | EN | 0.75 | CNN-DANN | **.754** | .523 | **.625** | .53 | .475 |
| DE | SR | 0.75 | CNN-DANN | .732 | .52 | .616 | .518 | **.513** |
| DE | GR | 0.75 | CNN-DANN | .742 | .527 | .622 | .52 | .493 |
| | | | *Valence* | | | | | |
| DE | - | 0.75 | CNN-BASE | .735 | **.441** | .662 | .401 | **.493** |
| DE | HU | 0.75 | CNN-DANN | .715 | .384 | .669 | **.43** | .387 |
| DE | EN | 0.75 | CNN-DANN | .751 | .391 | .662 | .42 | .468 |
| DE | SR | 0.75 | CNN-DANN | .748 | .396 | **.677** | .424 | .442 |
| DE | GR | 0.75 | CNN-DANN | **.76** | .383 | .655 | .421 | .476 |
| | | | *Average Arousal-Valence* | | | | | |
| DE | - | 0.75 | CNN-BASE | .732 | **.486** | .639 | .437 | **.497** |
| DE | HU | 0.75 | CNN-DANN | .724 | .457 | .645 | **.48** | .414 |
| DE | EN | 0.75 | CNN-DANN | **.753** | .457 | .644 | .475 | .472 |
| DE | SR | 0.75 | CNN-DANN | .74 | .458 | **.647** | .471 | .478 |
| DE | GR | 0.75 | CNN-DANN | .751 | .455 | .639 | .471 | .485 |

Now, the amount of source labels available to the DANN training process is reduced to 75%, using a sampling method identical to that in the baseline experiments. The results when using German as source culture are summarised in table 4.13.

For arousal, the top score of DANNs is 0.754 for German, and 0.53 for Hungarian on the validation data, which is above and on par with the baseline, respectively. The top test set results for arousal are 0.625, 0.53 and 0.513, all surpassing the baseline result for German, Hungarian, and Chinese respectively.

For valence, top validation scores of 0.76 and 0.396 are achieved on German and Hungarian, which is above and below than the baseline, respectively. The top test results are 0.677, 0.43 and 0.476, surpassing the baseline except on Chinese.

Looking at the average scores of valence and arousal gives top test set results of CCC 0.647 on German and 0.48 on Hungarian, which surpass baseline, and 0.485 on Chinese, which falls below.

In the next set of experiments, the same reduction to 75% of emotion labels is repeated with Hungarian as source culture. The results are shown in table 4.14.

Again beginning with arousal, the top validation scores are 0.656 and 0.466 on German and Hungarian respectively, both surpassing the baseline. On the test set, the

TABLE 4.14: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with Hungarian as source domain. CNN feature extractors are used and 75% of the labelled samples in the source domain are processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| HU | - | 0.75 | CNN-BASE | .642 | .45 | .539 | .531 | **.527** |
| HU | DE | 0.75 | CNN-DANN | .631 | .456 | **.577** | .554 | .517 |
| HU | EN | 0.75 | CNN-DANN | .644 | .449 | .531 | .56 | .433 |
| HU | SR | 0.75 | CNN-DANN | .656 | .403 | .553 | **.561** | .519 |
| HU | GR | 0.75 | CNN-DANN | **.656** | **.466** | .556 | .558 | .473 |
| | | | *Valence* | | | | | |
| HU | - | 0.75 | CNN-BASE | **.613** | .438 | .603 | .499 | .285 |
| HU | DE | 0.75 | CNN-DANN | .607 | **.438** | .607 | **.546** | .337 |
| HU | EN | 0.75 | CNN-DANN | .593 | .387 | .601 | .523 | **.348** |
| HU | SR | 0.75 | CNN-DANN | .604 | .371 | .568 | .525 | .342 |
| HU | GR | 0.75 | CNN-DANN | .604 | .386 | **.623** | .54 | .309 |
| | | | *Average Arousal-Valence* | | | | | |
| HU | - | 0.75 | CNN-BASE | .628 | .444 | .571 | .515 | .406 |
| HU | DE | 0.75 | CNN-DANN | .619 | **.447** | **.592** | .55 | .427 |
| HU | EN | 0.75 | CNN-DANN | .619 | .418 | .592 | .55 | .427 |
| HU | SR | 0.75 | CNN-DANN | **.63** | .387 | .561 | .543 | **.431** |
| HU | GR | 0.75 | CNN-DANN | .63 | .426 | .59 | **.549** | .391 |

top results are 0.577, 0.561 on German and Hungarian, above the baseline, while the top score for Chinese is $CCC = 0.519$, below the baseline.

Continuing with valence, the DANNs achieve top validation scores of 0.607 and 0.438, which is slightly below and on par with the German and Hungarian baselines, respectively. On the test set, the top scores in order are 0.623, 0.546 and 0.348, all surpassing the baselines.

Finally, looking at the averaged scoring of arousal and valence, the top validation scores are 0.63 and 0.447 CCC, narrowly surpassing the baseline. On the test set, top scores of 0.592 on German, 0.55 on Hungarian, and 0.431 on Chinese are achieved, all surpassing the respective baselines.

Once the labels are further reduced to 50% of the data, the results in table 4.15 are achieved with German as the source culture.

For arousal this yields top scores of $CCC = 0.741$ and $CCC = 0.508$ on German and Hungarian, and on the test set CCCs of 0.649, 0.545 and 0.436 are scored on German, Hungarian and Chinese respectively. Both the German and Hungarian test scores surpass the baseline, while performance on Chinese falls below.

On valence, the top validation scores are $CCC = 0.738$ and 0.392 respectively. The top test scores are 0.672, 0.425 and 0.471 respectively. Of the test cultures, only the

TABLE 4.15: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with German as source domain. CNN feature extractors are used and 50% of the labelled samples in the source domain are processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation CCC** | | **Test CCC** | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| DE | - | 0.50 | CNN-BASE | .719 | .516 | .621 | .486 | **.449** |
| DE | HU | 0.50 | CNN-DANN | **.741** | .497 | **.649** | **.545** | .419 |
| DE | EN | 0.50 | CNN-DANN | .727 | .493 | .627 | .531 | .436 |
| DE | SR | 0.50 | CNN-DANN | .727 | .481 | .596 | .502 | .413 |
| DE | GR | 0.50 | CNN-DANN | .735 | .508 | .62 | .51 | .347 |
| | | | *Valence* | | | | | |
| DE | - | 0.50 | CNN-BASE | .714 | .4 | .671 | .387 | .47 |
| DE | HU | 0.50 | CNN-DANN | .733 | .384 | **.672** | **.425** | .416 |
| DE | EN | 0.50 | CNN-DANN | **.738** | **.392** | .635 | .399 | .47 |
| DE | SR | 0.50 | CNN-DANN | .715 | .382 | .621 | .371 | **.471** |
| DE | GR | 0.50 | CNN-DANN | .734 | .376 | .613 | .397 | .432 |
| | | | *Average Arousal-Valence* | | | | | |
| DE | - | 0.50 | CNN-BASE | .717 | **.458** | .646 | .437 | **.46** |
| DE | HU | 0.50 | CNN-DANN | **.737** | .441 | **.661** | **.485** | .418 |
| DE | EN | 0.50 | CNN-DANN | .733 | .443 | .631 | .465 | .453 |
| DE | SR | 0.50 | CNN-DANN | .721 | .432 | .609 | .437 | .442 |
| DE | GR | 0.50 | CNN-DANN | .735 | .442 | .617 | .454 | .39 |

scores of German and Chinese are on par with the baseline, while Hungarian is considerably better.

Looking at the averaged arousal-valence scores shows top validation values of 0.737 for German and 0.443 for Hungarian, above and below baseline respectively. For the test sets, the results in order are 0.661, 0.485, 0.453. Only German and Hungarian surpass the baseline here.

Repeating the 50% reduction, this time with Hungarian as the source culture, leads to the results summarised in table 4.16.

Beginning once more with arousal, the top German and Hungarian validation scores are $CCC = 0.659$ and $CCC = 0.454$ respectively, both above the supervised baseline. On the test set, the top German, Hungarian and Chinese CCC scores are 0.597, 0.561 and 0.502, all surpassing the baseline.

For valence, the top validation scores are 0.618 and 0.393, neither surpassing the corresponding baseline. On the test set, German achieves 0.609, Hungarian achieves 0.553 and Chinese gives 0.378, with only Hungarian surpassing the baseline.

Averaging the two affect dimensions gives top validation scores of 0.637 and 0.414 for German and Hungarian respectively. This represents an improvement over baseline on German. On the test set, the top scores are 0.601, 0.545 and 0.445 for German,

TABLE 4.16: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with Hungarian as source domain. CNN feature extractors are used and 50% of the labelled samples in the source domain are processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| HU | - | 0.50 | CNN-BASE | .618 | .431 | .528 | .534 | .498 |
| HU | DE | 0.50 | CNN-DANN | .656 | .45 | .554 | .546 | **.502** |
| HU | EN | 0.50 | CNN-DANN | .636 | .424 | .575 | .545 | .425 |
| HU | SR | 0.50 | CNN-DANN | .59 | .397 | **.597** | **.561** | .445 |
| HU | GR | 0.50 | CNN-DANN | **.659** | **.454** | .593 | .537 | .487 |
| | | | *Valence* | | | | | |
| HU | - | 0.50 | CNN-BASE | **.623** | **.419** | **.612** | .534 | **.48** |
| HU | DE | 0.50 | CNN-DANN | .618 | .378 | .604 | .524 | .313 |
| HU | EN | 0.50 | CNN-DANN | .618 | .375 | .589 | .547 | .378 |
| HU | SR | 0.50 | CNN-DANN | .61 | .393 | .586 | .529 | .333 |
| HU | GR | 0.50 | CNN-DANN | .605 | .365 | .609 | **.553** | .334 |
| | | | *Average Arousal-Valence* | | | | | |
| HU | - | 0.50 | CNN-BASE | .621 | **.425** | .57 | .516 | **.489** |
| HU | DE | 0.50 | CNN-DANN | **.637** | .414 | .579 | .535 | .408 |
| HU | EN | 0.50 | CNN-DANN | .627 | .4 | .582 | **.546** | .402 |
| HU | SR | 0.50 | CNN-DANN | .6 | .395 | .592 | .545 | .445 |
| HU | GR | 0.50 | CNN-DANN | .632 | .41 | **.601** | .545 | .411 |

Hungarian and Chinese, the first two outperforming the baseline while the latter falls below.

Access to the labels is then restricted even further in the last set of experiments on label reduction. Making just 25% of training annotations available to the DANN results in the scores from table 4.17 when German is used as the source culture.

Beginning again with the results for arousal, the top validation scores are $CCC = 0.723$ and $CCC = 0.509$ for German and Hungarian, respectively, the latter falling below the supervised baseline. On the test set, both German and Hungarian surpass the baseline with CCC scores of 0.648 and 0.505 respectively, while Chinese falls below with 0.413.

Moving on to valence, the top validation scores are 0.721 for German and 0.408 for Hungarian, both rising above the baseline. On the test set, the top scores of the DANNs are 0.677 for German, 0.397 for Hungarian and 0.471 for Chinese, all surpassing the baseline.

Finally, for the averaged arousal-valence scores, the top validation results are 0.722 and 0.457 for German and Hungarian respectively, with only German surpassing the baseline. On test, the scores are 0.663, 0.449 and 0.442, all above the baseline.

TABLE 4.17: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with German as source domain. CNN feature extractors are used and 25% of the labelled samples in the source domain are processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| DE | - | 0.25 | CNN-BASE | .717 | **.519** | .619 | .474 | **.44** |
| DE | HU | 0.25 | CNN-DANN | **.723** | .485 | **.648** | .5 | .406 |
| DE | EN | 0.25 | CNN-DANN | .69 | .509 | .602 | **.505** | .346 |
| DE | SR | 0.25 | CNN-DANN | .711 | .503 | .596 | .502 | .413 |
| DE | GR | 0.25 | CNN-DANN | .702 | .506 | .633 | .502 | .392 |
| | | | *Valence* | | | | | |
| DE | - | 0.25 | CNN-BASE | .692 | .4 | .66 | .374 | .439 |
| DE | HU | 0.25 | CNN-DANN | **.721** | .396 | **.677** | **.397** | .398 |
| DE | EN | 0.25 | CNN-DANN | .692 | .402 | .619 | .389 | .445 |
| DE | SR | 0.25 | CNN-DANN | .715 | .4 | .621 | .371 | **.471** |
| DE | GR | 0.25 | CNN-DANN | .708 | **.408** | .616 | .392 | .37 |
| | | | *Average Arousal-Valence* | | | | | |
| DE | - | 0.25 | CNN-BASE | .705 | .46 | .64 | .424 | .44 |
| DE | HU | 0.25 | CNN-DANN | **.722** | .441 | **.663** | **.449** | .402 |
| DE | EN | 0.25 | CNN-DANN | .691 | .456 | .611 | .447 | .396 |
| DE | SR | 0.25 | CNN-DANN | .713 | .452 | .609 | .437 | **.442** |
| DE | GR | 0.25 | CNN-DANN | .705 | **.457** | .625 | .447 | .381 |

The previous experiment is repeated using Hungarian as the source culture instead, and restricting again to 25% of training labels. The findings are reported in table 4.18, and the best performing models manifest the following scores:

On arousal, evaluating the German and Hungarian validation data gives 0.631 and 0.459 respectively, both surpassing the baseline. On test, the best results are 0.55 on German, 0.569 on Hungarian, and 0.541 on Chinese, all representing an improvement on the baseline.

Continuing with valence, here the top validation scores are 0.636 and 0.426 on German and Hungarian respectively, again surpassing the baseline. For the test data, the respective scores on German, Hungarian and Chinese are 0.649, 0.522 and 0.381, all surpassing the baseline.

Looking at the averages of arousal and valence per model, the DANNs achieve 0.636 and 0.432 on the validation culture splits, both surpassing the baseline. Likewise, the top test scores of 0.6, 0.546 and 0.454 for German, Hungarian and Chinese are all well above those of the supervised model.

TABLE 4.18: Validation and test set results of the best DANN models trained on the cultures of the SEWA dataset, with Hungarian as source domain. CNN feature extractors are used and 25% of the labelled samples in the source domain are processed. Top scores for arousal and valence are shown, obtained by aggregating the best performing model runs. The feature encoder of the DANNs consists of either GRUs or self-attention transformer stack followed by GRUs.

| **Source** | **Target** | **Split** | **Model** | **Validation** CCC | | **Test** CCC | | |
|---|---|---|---|---|---|---|---|---|
| | | | | DE | HU | DE | HU | CN |
| | | | *Arousal* | | | | | |
| HU | - | 0.25 | CNN-BASE | .597 | .388 | .469 | .529 | .47 |
| HU | DE | 0.25 | CNN-DANN | **.631** | **.459** | **.55** | .534 | .532 |
| HU | EN | 0.25 | CNN-DANN | .627 | .426 | .547 | .534 | .512 |
| HU | SR | 0.25 | CNN-DANN | .616 | .385 | .499 | .549 | **.541** |
| HU | GR | 0.25 | CNN-DANN | .616 | .428 | .513 | **.569** | .443 |
| | | | *Valence* | | | | | |
| HU | - | 0.25 | CNN-BASE | .608 | .336 | .576 | .484 | .3 |
| HU | DE | 0.25 | CNN-DANN | **.636** | .405 | **.649** | .487 | .339 |
| HU | EN | 0.25 | CNN-DANN | .612 | **.426** | .606 | .487 | **.381** |
| HU | SR | 0.25 | CNN-DANN | .614 | .345 | .59 | .501 | .367 |
| HU | GR | 0.25 | CNN-DANN | .607 | .353 | .612 | **.522** | .362 |
| | | | *Average Arousal-Valence* | | | | | |
| HU | - | 0.25 | CNN-BASE | .603 | .362 | .523 | .507 | .385 |
| HU | DE | 0.25 | CNN-DANN | **.636** | **.432** | **.6** | .511 | .436 |
| HU | EN | 0.25 | CNN-DANN | .62 | .426 | .577 | .511 | .447 |
| HU | SR | 0.25 | CNN-DANN | .615 | .365 | .545 | .525 | **.454** |
| HU | GR | 0.25 | CNN-DANN | .612 | .391 | .563 | **.546** | .403 |

## 4.3 Vocal Burst Affect Detection

This section describes the experiments on the Hume-VB dataset for multi-task prediction of continuous emotion annotations and vocal burst type. The task definitions are identical with those of the four tracks of the A-VB competition (A-VB-TYPE, A-VB-TWO, A-VB-HIGH and A-VB-CULTURE), see section 2.6.1. Various architectures built on top of the audio Transformer WAV2VEC2 are investigated, as well as dynamic loss balancing functions introduced in section 3.3.1. Together with the discussion in section 5.2, these results form contribution C–2, which is focused on answering research question RQ–4, i. e., emotion recognition from non-verbal vocalisations.

### 4.3.1 Dataset Preprocessing

All audio clips are cut to 2.5s length, shorter recordings are zero-padded at the end. No additional preprocessing is performed and the audio is processed directly in the models.

### 4.3.2 Models

The approach for these experiments uses a large model pre-trained via self-supervised learning to extract representations from the vocal bursts. In this case, WAV2VEC2 (Baevski et al., 2020) is chosen, specifically, the wav2vec2-base architecture [6]. It has approximately 78M parameters and consists of two sub-networks: A CNN which returns features of dimension 512 and 12 transformer layers which produce embeddings of dimension 768. There are many fine-tuned variants of this model, but in this work, only the base variant without tuning on a speech corpus, e. g., Librispeech, is used. The motivation for this choice is the particular case of the Hume-VB dataset not having any verbal content, so a transformer fine-tuned towards speech is not assumed to provide a significant benefit over the base variant.

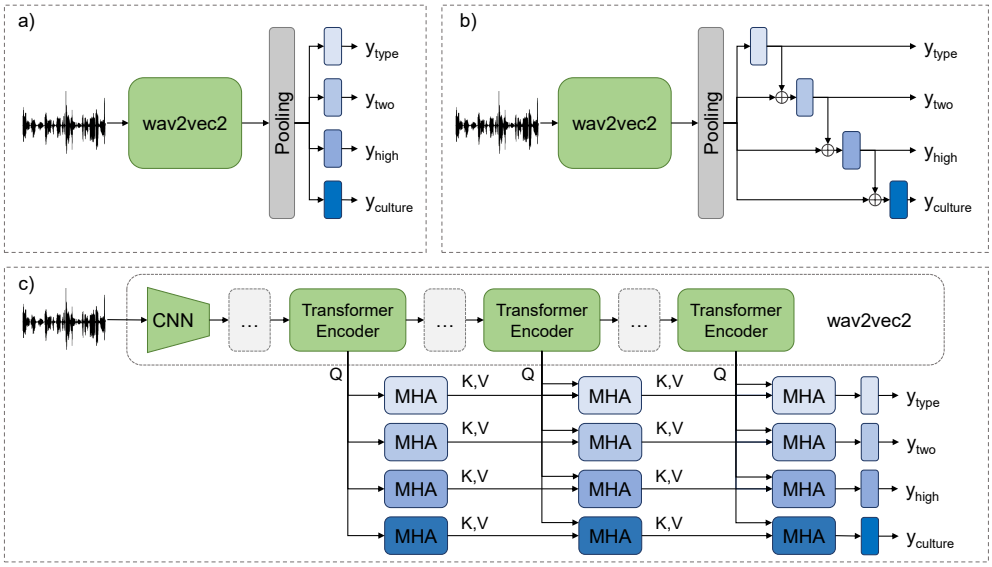



FIGURE 4.3: Multi-task models used for analysing vocal bursts: a) basic MTL model, b) classifier chain model, c) branching attention model.

On top of the WAV2VEC2 backbone, three model variants are constructed. They are illustrated in fig. 4.3 and described below.

#### Basic Multi-Task model

This basic model consists of four parallel fully connected networks which take the last transformer layer embedding as input and each predict one of the tasks. The prediction heads have identical size and are designed to have a low parameter count, consisting only of one hidden layer and one output layer.

#### Chain model

In this model, the basic architecture is extended by chaining prediction heads, as described in section 3.4.5. The WAV2VEC2 embeddings of the final transformer layer are used as input to the first task, and for each following task the embeddings are concatenated with the predictions of the previous tasks. In order to avoid confusing the model with inaccurate information during training, the ground truth labels instead of the predicted values are used as inputs to the network.

[6] https://huggingface.co/facebook/wav2vec2-base

The order of the tasks in the chain can be freely chosen. It is also possible to use a hybrid approach where some tasks are predicted in parallel, while others are chained. An initial assumption for the ordering is that easier tasks can be placed at the front, where they provide input to more complex tasks downstream. The type of a vocal burst may be easiest to recognise (A-VB-TYPE), followed by two-dimensional affective state (A-VB-TWO), then multiple emotional states (A-VB-HIGH), and finally cultural aspects of emotion (A-VB-CULTURE).

#### Branching Multi-Head Attention Model

In contrast to the previous architectures, this model uses multiple hidden states of WAV2VEC2 as task-specific inputs. The rationale behind this approach is that useful information may be contained at different depths for each task. Thus, inspired by multi-task attention network (Liu, Johns, and Davison, 2019), MHA blocks are used to combine features from selected layers of WAV2VEC2. The features serve as queries, while the output of each block forms the key-value pairs for the next one.

### 4.3.3 Training

The models are implemented in PyTorch and trained on Nvidia RTX3090 and A40 GPUs. Each training run is repeated multiple times ($N = 3$) with different seeds for the random weight initialisations, as these can have a significant impact on the outcome.

TABLE 4.19: Hyperparameter search space used in the vocal burst analysis experiments on the HUME-VB dataset. Various optimisation strategies including different loss weights are applied. For the task chaining architecture, different task orders (standard order or based on descending performance) and levels of chaining/parallelisation are used. In the branching attention architecture, both multi-head attention parameters and selections of hidden layers to branch out from are varied.

| **Hyperparameter** | **Value Range** |
|---|---|
| random seed | [1-3] |
| *Task heads* | |
| $d_{hidden}$ | [32, 64, 128, 256] |
| activation | [GELU, SELU] |
| *Optimisation* | |
| weight decay | [$10^{-4}$-$10^{-3}$] |
| $lr_{Transformer}$ | [$10^{-5}$-$10^{-4}$] |
| $lr_{downstream}$ | [$10^{-4}$-$10^{-3}$] |
| loss balancing | [mean, dwa, rruw, druw] |
| *Chain Models* | |
| task order | [(TYPE,LOW,HIGH,CULTURE), (LOW,HIGH,CULTURE,TYPE)] |
| internal task structure | [sequential, parallel] |
| internal task order | [default, performance desc.] |
| chain structure | [sequential, partly parallel] |
| *Branching Attention Models* | |
| $d_{embedding}$ | [32, 64, 128] |
| $n_{heads}$ | [4, 8] |
| branching layer depth | [1-12] |

For data augmentation, the equivalent of SpecAugment (D. S. Park et al., 2019) i. e., setting randomly selected slices of data to 0, is used on the CNN embeddings of wav2vec2 before they are passed to the transformer. This masking is applied to both the temporal and feature dimensions, and the default values of the Huggingface implementation (masking probability $p = 0.05$, mask length 10) are chosen.

All models are trained end-to-end for 30 epochs, with a learning rate scheduler set to halve the rate if the validation set performance fails to improve for 5 epochs. The choice of optimiser is AdamW, with weight decay set between $10^{-4}$ and $10^{-3}$. The initial learning rate is chosen in the interval $10^{-5}$–$10^{-4}$ for the pre-trained feature extraction network, and between $10^{-4}$–$10^{-3}$ for the randomly initialised downstream task prediction networks. The batch size is fixed to 8 based on GPU VRAM limitations. For the task prediction networks, hidden layer size is varied from 32 to 256, and GELU and SELU are chosen as activation functions. An overview on the hyperparameter search space is given in table 4.19.

### 4.3.4 Results

Here the results of the experiments on vocal burst data are presented. Since the test set is hidden, the focus is on the validation set.

TABLE 4.20: Validation set results on the four tasks for the basic MTL architecture with different transformer backbones: WAV2VEC2-BASE, WAV2VEC2-LARGE, WAV2VEC2-LARGE pruned and fine-tuned on MSP-Podcast, and HUBERT-BASE.

| **Backbone model** | **A-VB-TYPE** UAR | **A-VB-TWO** CCC | **A-VB-HIGH** CCC | **A-VB-CULTURE** CCC |
|---|---|---|---|---|
| wav2vec2-base | **.5547** | **0.7026** | **0.7271** | **0.6025** |
| wav2vec2-large | 0.5363 | 0.7018 | 0.7271 | 0.5969 |
| wav2vec2-large-ft | 0.5217 | 0.6898 | 0.7179 | 0.584 |
| hubert-base | 0.5492 | 0.6971 | 0.7231 | 0.5913 |

In order to examine whether WAV2VEC2-BASE is a suitable choice as feature extractor, experiments are conducted with the basic MTL architecture, using various transformer based models. These include WAV2VEC2-LARGE [7], WAV2VEC2-LARGE fine-tuned on MSP-Podcast and pruned to 12 layers [8] and HUBERT-BASE [9]. For each set of experiments, the transformer and the downstream prediction heads are trained end-to-end. Following hyperparameter optimisation and combining the predictions of $N = 3$ randomly initialised runs, the best results per type of model on each of the four tasks are selected. A comparison of the performances of the basic MTL architecture with different transformer backbones is shown in table 4.20. Given the similarity of the results, using the larger or fine-tuned models is not considered beneficial and WAV2VEC2-BASE is kept as the backbone for all experiments.

Next, for each downstream multi-task classifier/regressor architecture, the results for each task presented in detail.

[7] https://huggingface.co/facebook/wav2vec2-large
[8] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
[9] https://huggingface.co/facebook/hubert-base-ls960

### Basic Multi-Task Model

Results for the A-VB-TYPE task are shown in fig. 4.4. The best achieved values of recall per class are 0.8372 for Laugh, 0.4357 for Grunt, 0.4976 for Cry, 0.5131 for Pant, 0.7584 for Gasp, 0.3956 for Groan, 0.566 for Scream, and 0.5298 for Other, respectively.

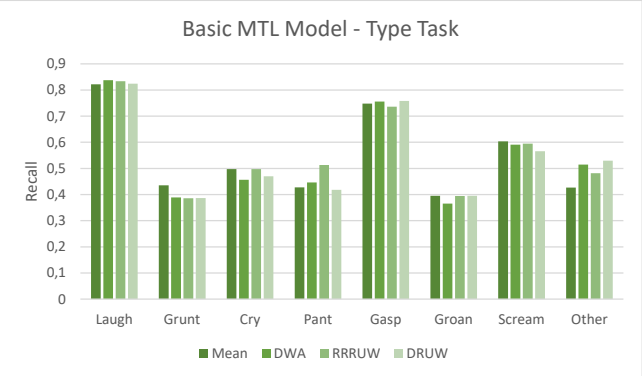
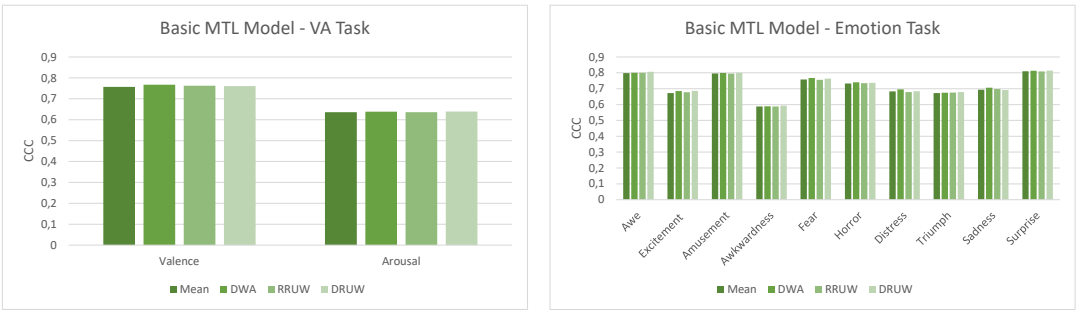



FIGURE 4.4: Validation set results of the best performing basic multi-task models on the A-VB-TYPE task. Shown are recall scores per class and loss balancing strategy.




(A) Two-dimensional affect

(B) 10 emotions

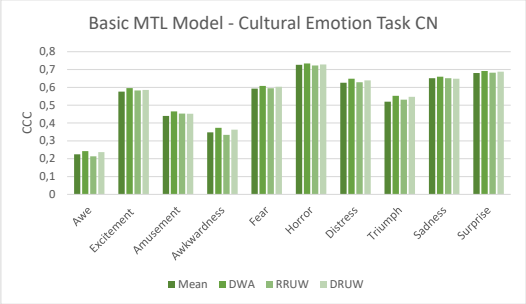
FIGURE 4.5: Validation set results of best performing base architecture models for A-VB-TWO and A-VB-HIGH tasks. Shown are CCC scores for: a) arousal-valence, b) 10 continuous-valued emotions.

In fig. 4.5a, the results of the A-VB-TWO task are shown, with maximum CCC values of 0.7677 and 0.6393 for valence and arousal, respectively.

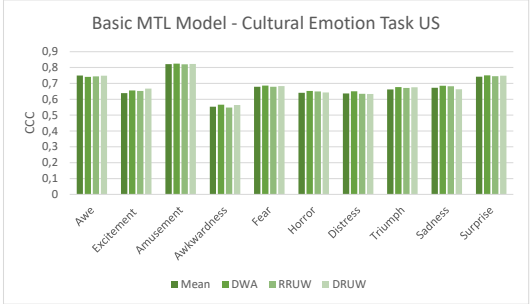
fig. 4.5b depicts the CCC values of the 10 annotated emotions in the A-VB-HIGH task. The best results are 0.8066 for Amusement, 0.6867 for Excitement, 0.8 for Amusement, 0.5935 for Awkwardness, 0.7675 for Fear, 0.7406 for Horror, 0.6837 for Distress, 0.6785 for Triumph, 0.7062 for Sadness, and 0.8131 for Surprise.

Finally, for the A-VB-CULTURE task, the results of the basic multi-task models are depicted in fig. 4.6. For each culture and annotated emotion, the best CCC value is also given in table 4.21.

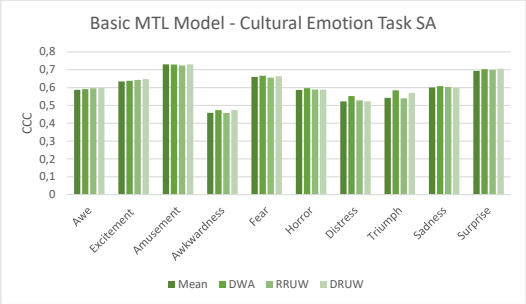
The top CCC score for Awe is 0.7495, for Excitement it is 0.6668. Amusement and Awkwardness achieve 0.8246 and 0.5654, respectively. Fear, Horror, and Distress are recognised at 0.6867, 0.7343 and 0.6508. Triumph, Sadness and Surprise are predicted with CCCs of 0.6767, 0.6855, and 0.7502. With the exception of Horror, all top scores come from the United States culture. The best average score across all cultures and emotions, achieved with DWA loss, is $CCC = 0.6025$.



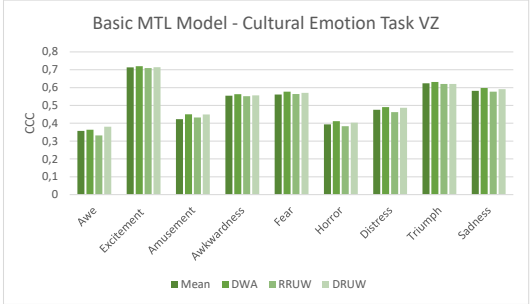

(A) China emotions




(B) United States emotions




(C) South Africa emotions




(D) Venezuela emotions

FIGURE 4.6: Validation set results of the best performing basic multi-task models on the A-VB-CULTURE task. Shown are CCC scores for the 10 annotated emotions per culture: a) Chinese, b) United States, c) South Africa, and d) Venezuela.

TABLE 4.21: Validation set results on the A-VB dataset per culture and annotated emotion in the A-VB-CULTURE task for the best performing basic multi-task models.

| **Culture** | **Awe** CCC | **Excitement** CCC | **Amusement** CCC | **Awkwardness** CCC | **Fear** CCC |
|---|---|---|---|---|---|
| China | .2428 | .5963 | .4649 | .3731 | .6078 |
| South Africa | .5987 | .6473 | .7304 | .4746 | .6664 |
| United States | **.7495** | **.6668** | **.8246** | **.5654** | **.6867** |
| Venezuela | .6897 | .3804 | .7201 | .4498 | .5627 |
| | **Horror** CCC | **Distress** CCC | **Triumph** CCC | **Sadness** CCC | **Surprise** CCC |
| China | **.7343** | .6481 | .5529 | .6602 | .6913 |
| South Africa | .5967 | .553 | .585 | .6081 | .7058 |
| United States | .6526 | **.6508** | **.6767** | **.6855** | **.7502** |
| Venezuela | .5774 | .4119 | .4911 | .6312 | .598 |

### Classifier chain models

In the classifier chain approach, there are many options for assembling the chain. Thus, various models with unique structures were evaluated. These included different orders of the tasks, as well as choosing which predictions to chain or run in parallel. Chains can be created at the top level of the four A-VB tasks, as seen in fig. 4.3, but also inside the task heads.

Regarding task-level chain order, preliminary experiments showed low performance in the A-VB-TYPE task if it was at the end of the chain, thus the task order *type* $\rightarrow$

TABLE 4.22: Configurations of classifier chain models used for multi-task learning on the A-VB dataset. Within a given task, the predictions can themselves be chained or predicted in parallel. Chains can be set in standard order provided by the annotation files, or ordered by descending validation set performance of the basic MTL model. The tasks are predicted in sequence, but the A-VB-CULTURE head may be set in parallel to the A-VB-HIGH head to reduce overall chain length (rightmost column).

| **Config** | **Task** | | | | | | **Fork Culture** |
|---|---|---|---|---|---|---|---|
| | A-VB-TWO | | A-VB-HIGH | | A-VB-CULTURE | | |
| | Chain | Order | Chain | Order | Chain | Order | |
| A | ✗ | - | ✗ | - | ✗ | - | ✗ |
| B | ✓ | perf | ✗ | - | ✗ | - | ✓ |
| C | ✓ | perf | ✗ | - | ✓ | perf | ✓ |
| D | ✓ | perf | ✓ | perf | ✗ | - | ✗ |
| E | ✓ | perf | ✓ | perf | ✓ | perf | ✓ |
| F | ✓ | - | ✓ | - | ✓ | - | ✓ |

*low* $\rightarrow$ *high* $\rightarrow$ *culture* was chosen. In order to shorten the length of the classifier chain, if the A-VB-CULTURE task was chained internally, it contained 4 parallel chains of length 10, one per culture. In addition, the A-VB-CULTURE task was placed in parallel to the A-VB-HIGH task in some configurations.

For ordering the predictions heads when chaining internally within a task, e. g., emotions for A-VB-HIGH, two options are used: standard order of annotations in the dataset, and descending order by performance, based on the results of the basic MTL model shown above. Chaining the A-VB-TYPE task internally showed worse results in initial experiments and was thus avoided.

In total, six different configurations were used for experimentation, labelled A–F. Configuration A is the least complex, with inter-task chaining and parallel prediction of the affect dimensions or emotions within the respective tasks. The others include variations of intra-task chaining and ordering as described above. Detailed configuration settings are listed in table 4.22.

Validation set results of the various classifier chain configurations, in terms of UAR for A-VB-TYPE and averaged CCC as well as averaged $\rho$ for the other tasks, are given in table 4.23. On A-VB-TYPE, the top result is $UAR = 0.5687$ (model D). For A-VB-TWO, the best CCC score is 0.7071 (model F). The best average score on the 10 emotions task A-VB-HIGH is $CCC = 0.7299$ (model B). Finally, the A-VB-CULTURE task achieves $CCC = 0.6072$ (model E). Regarding impact of loss balancing methods, the methods that take into account the loss development in preceding steps (DWA and DRUW) achieve the best results. They outperform RRUW, which in turn performs better than uniform weighing.

For each of the four tasks, validation set results of the top models presented above are now shown in more detail, beginning with A-VB-TYPE.

The results for the best performing classifier chain models on A-VB-TYPE are given in fig. 4.7. The maximum class recalls are 0.845 for Laugh, 0.4644 for Grunt, 0.5513 for Cry, 0.5083 for Pant, 0.7725 for Gasp, 0.4546 for Groan, 0.6409 for Scream, and 0.5779 for Other, respectively.

TABLE 4.23: Comparison of multiple classifier chain arrangements in terms of validation set performance (CCC and Pearson correlation coefficient) on the A-VB dataset.

| Config | A-VB-TYPE | A-VB-TWO | | A-VB-HIGH | | A-VB-CULTURE | |
|---|---|---|---|---|---|---|---|
| | UAR | CCC | $\rho$ | CCC | $\rho$ | CCC | $\rho$ |
| *Uniform Weighting* | | | | | | | |
| A | .5483 | .6948 | .6979 | .7103 | .718 | .5619 | .5844 |
| B | .553 | .6913 | .6974 | .7103 | .7181 | .5624 | .585 |
| C | .517 | .6912 | .6932 | .713 | .7227 | .5747 | .5922 |
| D | .5533 | .6982 | .6991 | .6047 | .6703 | .0781 | .1444 |
| E | .5534 | .6928 | .6952 | .592 | .671 | .5714 | .586 |
| F | .5301 | .6915 | .6933 | .713 | .722 | .5756 | .5926 |
| *Dynamic Weight Average* | | | | | | | |
| A | .5686 | .7068 | .7064 | .7276 | .7383 | .5922 | .6162 |
| B | .5612 | .7048 | .7054 | **.7299** | **.7387** | .5934 | .6157 |
| C | .5411 | .7037 | .7054 | .725 | .7377 | .6006 | .6174 |
| D | .562 | .7001 | .7046 | .6363 | .6943 | .1278 | .2147 |
| E | .557 | .7 | .7009 | .6454 | .7041 | .607 | .6177 |
| F | .5492 | .7063 | .7071 | .7241 | .737 | .6018 | .6161 |
| *Restrained Revised Uncertainty Weighting* | | | | | | | |
| A | .551 | .6981 | .6989 | .7168 | .7237 | .5713 | .5935 |
| B | .556 | .6968 | .6968 | .7122 | .7186 | .5637 | .5873 |
| C | .5309 | .6949 | .6961 | .7194 | .7271 | .5807 | .5988 |
| D | .5396 | .6875 | .691 | .5809 | .6717 | .0776 | .1433 |
| E | .5588 | .6963 | .698 | .5902 | .679 | .5819 | .5961 |
| F | .5376 | .6952 | .6976 | .7142 | .7233 | .5788 | .5962 |
| *Dynamic Restrained Uncertainty Weighting* | | | | | | | |
| A | .5603 | .7061 | **.7084** | .7291 | .7372 | .5391 | .6155 |
| B | .5649 | .7043 | .7075 | .7285 | .7365 | .5933 | .6174 |
| C | .5446 | .702 | .7028 | .7259 | .7367 | .6015 | .6155 |
| D | **.5687** | .6976 | .7025 | .6178 | .6963 | .1148 | .1936 |
| E | .5638 | .7019 | .7033 | .6468 | .7034 | **.6072** | **.6188** |
| F | .5542 | **.7071** | .7077 | .7262 | .7373 | .6066 | .6158 |

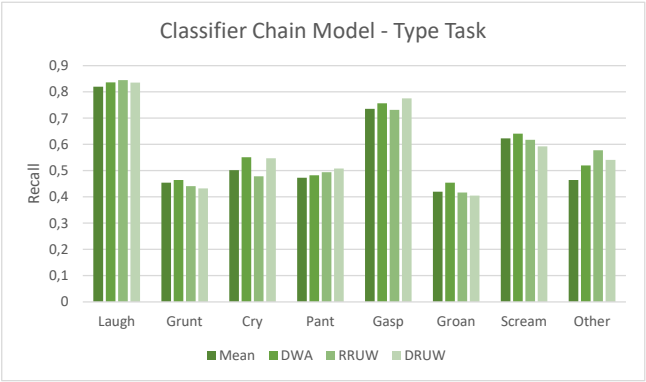
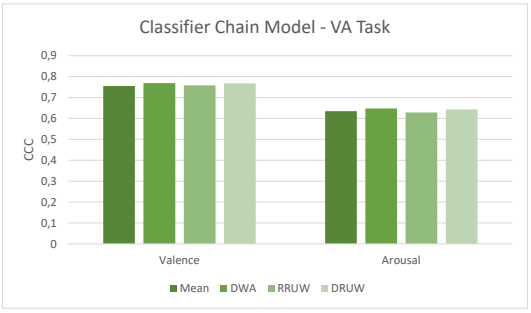
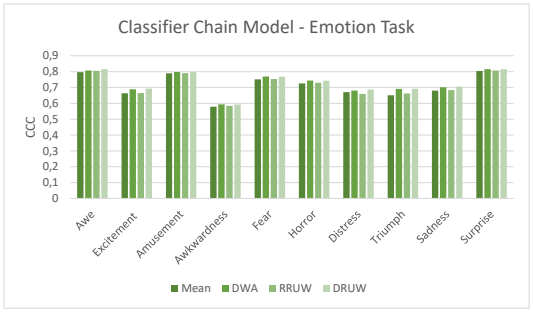



FIGURE 4.7: Validation set performance of the best performing classifier chain models on the A-VB-TYPE task. Shown are recall scores per class and loss balancing strategy.

Results for valence-arousal prediction and for the 10 annotated emotions on the validation set are given in fig. 4.8a and fig. 4.8b, respectively. For the A-VB-TWO task,



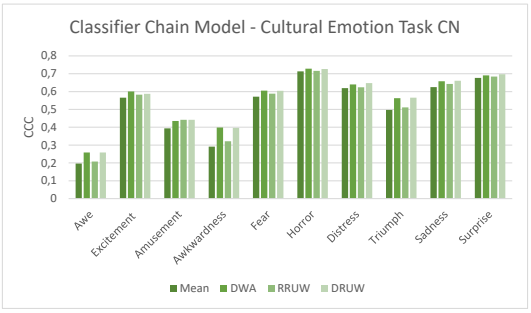

(A) Two-dimensional affect



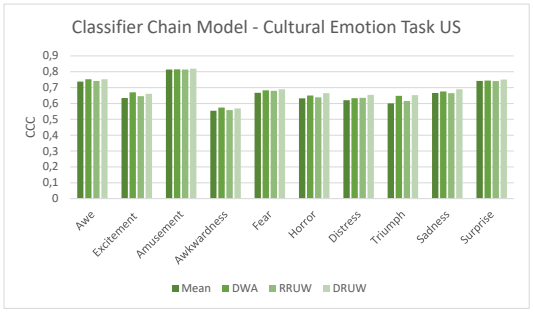

(B) 10 emotions

FIGURE 4.8: Validation set performance of the best classifier chain models on the A-VB-TWO and A-VB-HIGH tasks.

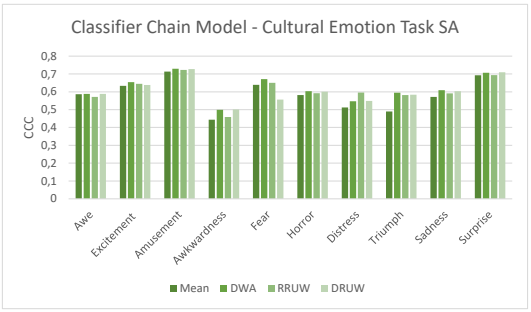
the best valence and arousal scores were 0.7686 and 0.6476 CCC, respectively. In the A-VB-HIGH task, the best achieved CCC values were 0.8157 for Awe, 0.6934 for Excitement, 0.7975 for Amusement, 0.5932 for Awkwardness, 0.7687 for Fear, 0.7432 for Horror, 0.6861 for Distress, 0.692 for Triumph, 0.7056 for Sadness, and 0.8161 for Surprise.



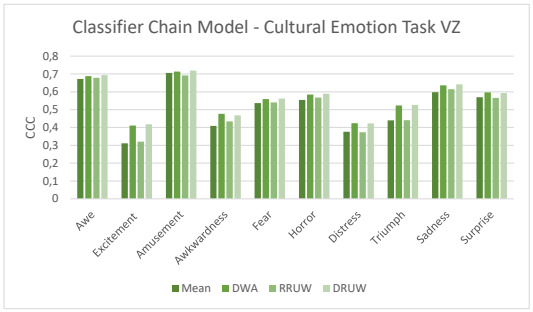

(A) China




(B) United States




(C) South Africa




(D) Venezuela

FIGURE 4.9: Validation set performance of the best classifier chain models on the A-VB-CULTURE task. Shown are CCC scores for the 10 annotated emotions per culture: a) Chinese, b) United States, c) South Africa, and d) Venezuela.

For each of the four cultures in the A-VB-CULTURE task, the validation set results are shown in fig. 4.9. In addition, the highest CCC scores per culture and emotion are summarised in table 4.24. The top score for Awe is $CCC = 0.7528$, for Excitement and Amusement it is 0.6701 and 0.819, respectively. Awkwardness achieved $CCC = 0.5742$. For Fear, Horror and Distress, the top results are $CCC = 0.6897$, $CCC = 0.7282$ and $CCC = 0.654$, respectively. Triumph, Sadness, and Surprise achieve CCC scores of 0.6528, 0.6897, and 0.7503, respectively. Just like the basic MTL model results in table 4.21, the top scores per emotion mostly originate from

TABLE 4.24: Comparison of the best performing classifier chain models on the A-VB-CULTURE task. Shown are CCC scores on the validation set for the best performing models per culture and annotated emotion.

| **Culture** | **Awe** | **Excitement** | **Amusement** | **Awkwardness** | **Fear** |
|---|---|---|---|---|---|
| | CCC | CCC | CCC | CCC | CCC |
| China | .2583 | .6003 | .4416 | .3987 | .605 |
| United States | **.7528** | **.6701** | **.819** | **.5742** | **.6897** |
| South Africa | .5887 | .6542 | .7293 | .5013 | .6711 |
| Venezuela | .6938 | .4184 | .7192 | .4769 | .5618 |
| **Culture** | **Horror** | **Distress** | **Triumph** | **Sadness** | **Surprise** |
| | CCC | CCC | CCC | CCC | CCC |
| China | **.7282** | .6473 | .5657 | .6604 | .697 |
| United States | .6654 | **.654** | **.6528** | **.6897** | **.7503** |
| South Africa | .603 | .596 | .595 | .6087 | .7093 |
| Venezuela | .5895 | .424 | .526 | .6416 | .5964 |

the United States culture, which considerably outperforms the others. The exception to this is again Horror, which is best recognised on Chinese data by a considerable margin (0.0618 over the next best result, 0.1377 over the lowest).

### Branching attention architecture

TABLE 4.25: Comparison of validation set performances (CCC and Pearson correlation coefficient) for branching models with varying selections of feature embeddings from the Transformer backbone: A (last 4 layers), B (even-numbered layers), C (first 4 layers)

| **Config** | **A-VB-TYPE** | **A-VB-TWO** | | **A-VB-HIGH** | | **A-VB-CULTURE** | |
|---|---|---|---|---|---|---|---|
| | UAR | CCC | $\rho$ | CCC | $\rho$ | CCC | $\rho$ |
| *Uniform Weighting* | | | | | | | |
| A | .5571 | .6934 | .6981 | .7114 | .7172 | .5791 | .5898 |
| B | **.5593** | .687 | .6925 | .7035 | .7055 | .5695 | .5796 |
| C | .4766 | .6609 | .6645 | .6764 | .6833 | .5468 | .5618 |
| *Dynamic Weight Average* | | | | | | | |
| A | .5479 | **.6966** | **.7008** | **.7214** | **.7272** | **.5931** | **.6043** |
| B | .5372 | .688 | .6914 | .7128 | .7186 | .582 | .5913 |
| C | .5126 | .6719 | .6732 | .6861 | .6909 | .5587 | .5704 |
| *Restrained Revised Uncertainty Weighting* | | | | | | | |
| A | .5476 | .6955 | .6997 | .7123 | .7203 | .5698 | .583 |
| B | .5583 | .6915 | .693 | .7046 | .7118 | .5678 | .5804 |
| C | .5017 | .665 | .6676 | .6829 | .6882 | .55536 | .5854 |
| *Dynamic Restrained Uncertainty Weighting* | | | | | | | |
| A | .5513 | .6951 | .6998 | .7204 | .727 | .5917 | .601 |
| B | .5437 | .6891 | .6933 | .7128 | .7194 | .581 | .5915 |
| C | .5163 | .6741 | .6756 | .6882 | .6939 | .5606 | .5728 |

For the branching architecture, the main design choice is selecting the branching points, i. e., the hidden layers in the Transformer backbone whose activations serve as inputs into the MHA branches. The base version of WAV2VEC2 has 12 Transformer

encoder layers. In preliminary experiments, various models using different combinations of layer depths are constructed. Three of them are presented here: Top-4 layers (A), highest evenly numbered layers (B) and bottom-4 layers (C). Validation set results for the best performing models organised by loss balancing method are given in table 4.25.

The top result for A-VB-TYPE is $UAR = 0.5583$ (model B). For A-VB-TWO, the top score is $CCC = 0.6966$, for A-VB-HIGH it is $CCC = 0.7214$, and for A-VB-CULTURE it is $CCC = 0.5931$. All top scores for the emotion tasks are achieved by model A, using DWA balancing. With the exception of A-VB-TYPE, training with dynamic task balancing again outperforms uniform loss weights. Comparing the three model configurations, model C performs consistently below the other two variants.

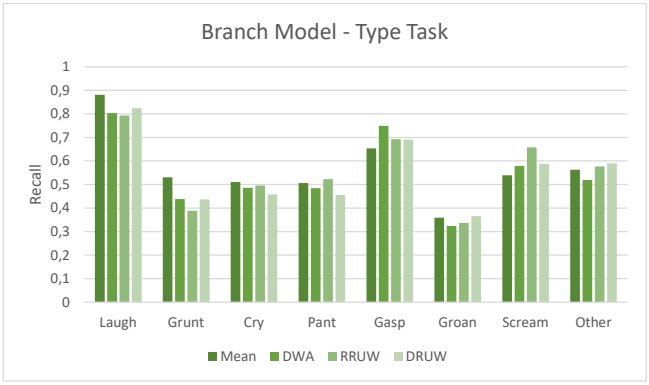
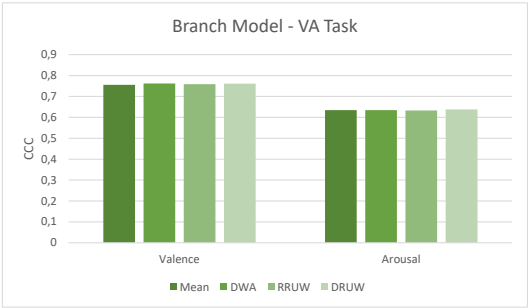
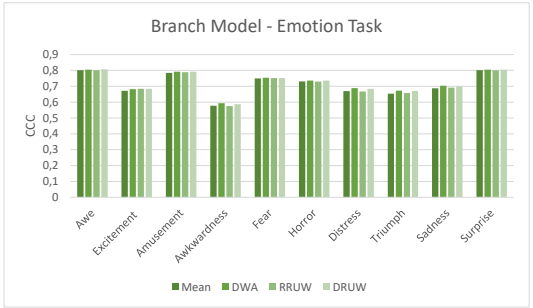



FIGURE 4.10: Validation set performance of the best performing branching attention models on the A-VB-TYPE task. Shown are recall scores per class and loss balancing strategy.

The validation set results for branching models on the A-VB-TYPE task are given in fig. 4.10. Best recalls per class are 0.8813 for Laugh, 0.5308 for Grunt, 0.5104 for Cry, 0.5226 for Pant, 0.7492 for Gasp, 0.3661 for Groan, 0.6579 for Scream, and 0.5901 for Other.




(A) Two-dimensional affect

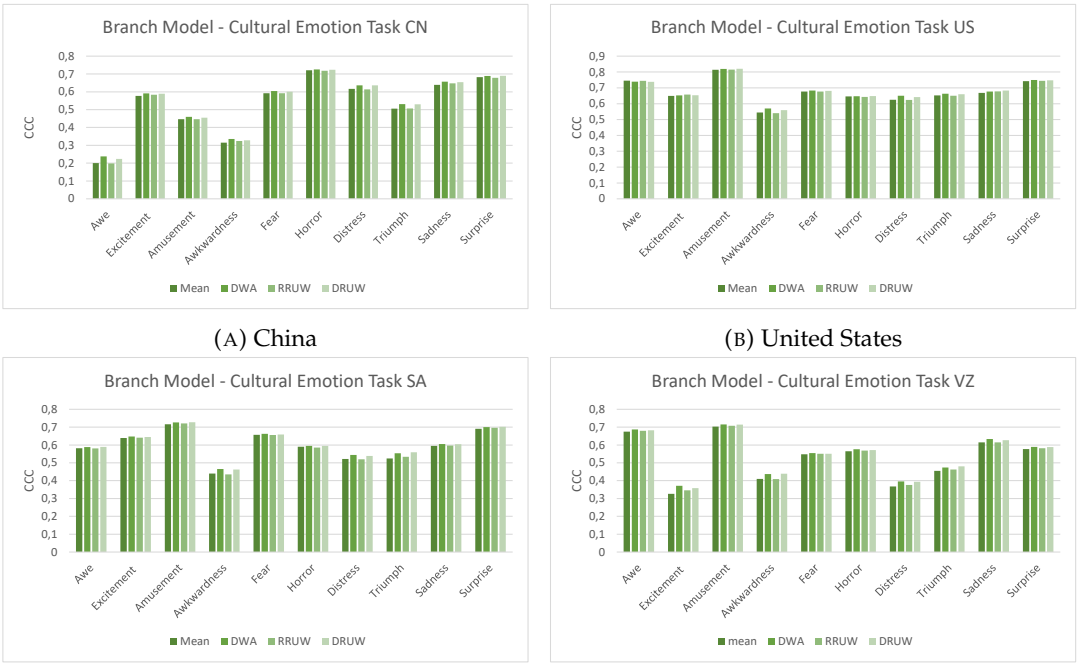



(B) 10 emotions

FIGURE 4.11: Validation set performance of the best branching models for A-VB-TWO and A-VB-HIGH tasks. Shown are CCC scores for: a) valence-arousal, b) 10 continuous emotions

For the A-VB-TWO task, the validation set results are shown in fig. 4.11a, with best CCC results of 0.7622 and 0.6378 for valence and arousal respectively. The A-VB-HIGH task results are shown in fig. 4.11b. There, the best CCC scores are 0.8072 for Awe, 0.6839 for Excitement, 0.7921 for Amusement, 0.5931 for Awkwardness,

0.7543 for Fear, 0.7355 for Horror, 0.6881 for Distress, 0.6725 for Triumph, 0.7029 for Sadness, 0.805 for Surprise.




(A) China (B) United States

(C) South Africa (D) Venezuela

FIGURE 4.12: Validation set results of the best performing branching models on the A-VB-CULTURE task. Shown are CCC scores for the 10 annotated emotions per culture: a) Chinese, b) United States, c) South Africa, and d) Venezuela.

TABLE 4.26: Comparison of the best performing branching models on the A-VB-CULTURE task. Shown are CCC scores on the validation set for the best performing models per culture and annotated emotion.

| **Culture** | **Awe** CCC | **Excitement** CCC | **Amusement** CCC | **Awkwardness** CCC | **Fear** CCC |
|---|---|---|---|---|---|
| China | .2379 | .5914 | .4597 | .3558 | .604 |
| United States | **.745** | .658 | **.8204** | **.5704** | **.6828** |
| South Africa | .5887 | .6471 | .7274 | .4665 | .6623 |
| Venezuela | .6869 | **.7153** | .4391 | .5547 | .5764 |
| **Culture** | **Horror** CCC | **Distress** CCC | **Triumph** CCC | **Sadness** CCC | **Surprise** CCC |
| China | **.7259** | .6367 | .5313 | .6573 | .6901 |
| United States | .6482 | **.6501** | **.6629** | **.683** | **.7495** |
| South Africa | .5949 | .5445 | .5588 | .6051 | .7021 |
| Venezuela | .5764 | .3954 | .4804 | .633 | .5892 |

In addition, the validation set results on the A-VB-CULTURE task are given in fig. 4.12. The best CCC scores per culture and annotated emotion are summarised in table 4.26.

The top score for Awe was $CCC = 0.745$, for Excitement and Amusement is was 0.7153 and 0.8204 respectively. Awkwardness scored $CCC = 0.5704$. Fear, Horror, and Distress yielded 0.6828, 0.7259 and 0.6501 respectively. Triumph, Sadness and

Surprise achieved $CCC = 0.6629$, $CCC = 0.683$ and $CCC = 0.7495$. The top result for Horror was again predicted on Chinese, while the top score for Excitement was found on Venezuelan ($+0.0777$ and $+0.0652$ over the next best results from the United States predictions, respectively). For the remaining 8 emotions, the best results were obtained on United States culture, just like for the other two architectures.

**Test set results**

After presenting detailed results for each of the three architectures, an overview of the performance on the validation set is now given. Results in terms of the evaluation metrics used in each of the four sub-challenges of the ACII'22 A-VB competition (see section 2.6) are shown in table 4.27. The numbers represent aggregated results of $N = 3$ runs, for the best models of each architecture and for each loss weighing method. Also shown for comparison are the baseline results achieved with end-to-end trained CNN-LSTM via END2YOU by the competition organisers (Baird, Tzirakis, Brooks, et al., 2022).

TABLE 4.27: Validation set results in terms of UAR for A-VB-TYPE, mean CCC, and mean $\rho$ for A-VB-TWO, A-VB-HIGH and A-VB-CULTURE, respectively. Shown are the best performing models for each task per architecture and loss weighing strategy, as well as the baseline score achieved by the organisers of the A-VB challenge with END2YOU.

| **Model** | **A-VB-TYPE** | **A-VB-TWO** | | **A-VB-HIGH** | | **A-VB-CULTURE** | |
|---|---|---|---|---|---|---|---|
| | UAR | CCC | $\rho$ | CCC | $\rho$ | CCC | $\rho$ |
| END2YOU | .4166 | .4988 | - | .5638 | - | .4401 | - |
| *Uniform Weighting* | | | | | | | |
| *BASIC MTL* | .5443 | .6964 | .6992 | .7205 | .7265 | .5892 | .5999 |
| *CHAIN* | .5534 | .6948 | .6979 | .7103 | .7180 | .5619 | .5844 |
| *BRANCH* | .5593 | .6934 | .6981 | .7114 | .7172 | .5791 | .5898 |
| *Dynamic Weight Average* | | | | | | | |
| *BASIC MTL* | .5446 | .7026 | .7034 | .7271 | .7347 | .6025 | .6128 |
| *CHAIN* | **.5686** | **.7068** | **.7074** | **.7276** | **.7383** | 6070 | .6177 |
| *BRANCH* | .5479 | .6966 | .7008 | .7214 | .7272 | .5931 | 6043 |
| *Restrained Revised Uncertainty Weighting* | | | | | | | |
| *BASIC MTL* | .5547 | .6992 | .7000 | .7213 | .7267 | .5892 | .5991 |
| *CHAIN* | .5588 | .6993 | .6989 | .7186 | .7237 | .5819 | .5961 |
| *BRANCH* | .5583 | .6955 | .6997 | .7123 | .7203 | .5850 | .5974 |
| *Dynamic Restrained Uncertainty Weighting* | | | | | | | |
| *BASIC MTL* | .5447 | .6950 | .7002 | .7243 | .7341 | .6006 | .6130 |
| *CHAIN* | .5638 | .7019 | .7033 | .7291 | .7372 | **.6072** | **.6188** |
| *BRANCH* | .5513 | .6951 | .6998 | .7204 | .7270 | .5917 | .6010 |

The validation set results are as follows: On the vocal burst classification task A-VB-TYPE the achievement was $UAR = 0.5686$. A-VB-TWO $CCC = 0.7068$, A-VB-HIGH $CCC = 0.7276$, A-VB-CULTURE $CCC = 0.6072$.

The test set results are shown in table 4.28. Results are based on test set predictions submitted to the organisers of the ACII'22 A-VB competition, and were first presented in a paper for the associated workshop (Karas, Triantafyllopoulos, et al., 2022). Since the number of submissions was limited, only the scores for the best models per architecture and task are listed. In addition, the predictions of those

models were combined into an ensemble, which was also submitted for evaluation. Again the END2YOU baseline is listed for comparison. The best results are $UAR = 0.5618$ for A-VB-TYPE, $CCC = 0.7066$ for A-VB-TWO, $CCC = 0.7363$ for A-VB-HIGH, and $CCC = 0.6195$ for A-VB-CULTURE, respectively.

TABLE 4.28: Test set results on the Hume-VB dataset of the best performing models per each architecture per task, as well as an ensemble created by combining predictions of those models by majority voting and averaging for classification and regression respectively. The END2YOU baseline results from the ACII'22 A-VB competition (Baird, Tzirakis, Brooks, et al., 2022) are shown for comparison.

| **Model** | **A-VB-TYPE** UAR | **A-VB-TWO** CCC | **A-VB-HIGH** CCC | **A-VB-CULTURE** CCC |
|---|---|---|---|---|
| END2YOU | .4172 | .5084 | .5686 | .4401 |
| *BASIC MTL* | .5377 | .6938 | .7209 | .6020 |
| *CHAIN* | **.5618** | .6942 | .7261 | .6002 |
| *BRANCH* | .5418 | .6888 | .7148 | .5945 |
| *ENSEMBLE* | .5560 | **.7066** | **.7363** | **.6195** |

# Chapter 5

# Discussion

In this chapter, the experiments of the previous chapter are discussed. The analysis is split into sections corresponding to the respective experimentation on multi-modal and cross-modal valence-arousal recognition on the Aff-Wild2 corpus, the cross-cultural valence-arousal prediction on SEWA and the multi-task learning on vocal burst audio data from the A-VB dataset.

## 5.1 Multi-Modal and Cross-Modal Emotion Recognition

This section interprets the experiments from section 4.1 on various sequence-to-sequence architecture processing CNN features for predicting affect on the videos of the Aff-Wild2 dataset in its ABAW'22 iteration.

### 5.1.1 Comparison between small and large feature extraction networks

Comparing the larger Inception-based network and the smaller MobileFaceNet directly showed the smaller CNN outperforming its larger counterpart. While the gain was small for valence, it was considerable for arousal (0.414 vs 0.203) when using the self-attention architecture.

This result is remarkable given that both networks are trained on facial recognition tasks and their layers were frozen in this experiment. The features from MobileFaceNet appear to transfer better to the emotion recognition task, especially for arousal. One might expect that the larger model would be able to generalise better. Instead, a CNN with less than 1M parameters was able to outperform a network 27 times its size. This result shows the optimisation potential of deep neural networks, and is encouraging for the development of computationally efficient affective solutions that can run on end user devices with limited resources. The ability to run locally is important for real-time applications, and may be mandatory due to the sensitive nature of the processed video data (see section 2.3 and section 6.1.1).

Owing to this strong performance, a MobileFaceNet fine-tuned on Aff-Wild2 was re-used in the experiments on SEWA described in section 4.2.

### 5.1.2 Comparison between frozen feature extraction networks and end-to-end learning

Comparing the results from table 4.2 and table 4.3 indicates that the end-to-end learning approach outperforms the frozen extractor approach on arousal prediction (top scores of 0.551 vs. 0.529 CCC), while the valence scores are similar and marginally better for the frozen network (0.393 vs. 0.388 CCC).

TABLE 5.1: Size of the best-performing models on the Aff-wild2 corpus validation set, adapted from Karas, Tellamekala, et al., 2022. Shown are the total number of parameters for the audiovisual models, grouped by seq-2-seq architecture (RNN, self-attention, cross-modal attention). For clarity, the number of parameters in the sequence models and the full number of parameters including CNNs are reported separately.

| **Method** | **Visual Encoder** | $\mathbf{P}_{sequence}$ | $\mathbf{P}_{total}$ |
|---|---|---|---|
| *Recurrent Models (RNNs)* | | | |
| AV-RNN | Inception | 109 K | 28.8 M |
| AV-RNN | MobileFaceNet | 4.4 M | 5.4 M |
| E2E-AV-RNN | MobileFaceNet | 76 K | 1.1 M |
| *Self-Attention (SA) Models* | | | |
| AV-SA | Inception | 765 K | 28.1 M |
| AV-SA | MobileFaceNet | 482 K | 1.51 M |
| E2E-AV-SA | MobileFaceNet | 193 K | 1.2 M |
| *Cross-Modal Attention (CMA) Models* | | | |
| AV-CMA | Inception | 134 K | 28.1 M |
| AV-CMA | MobileFaceNet | 2.1 M | 3.1 M |
| E2E-AV-CMA | MobileFaceNet | 2.4 M | 3.4 M |

The lack of improvement on overall valence prediction is due to the strong results of the CMA model with Inception-based features in the frozen experiments. In a direct comparison of the models using MobileFaceNet as the visual CNN however, a rise on valence prediction performance is apparent: 0.319 $\rightarrow$ 0.361 for RNN, 0.248 $\rightarrow$ 0.380 for self-attention, and 0.324 $\rightarrow$ 0.388 for cross-modal attention.

Notably, the RNN-based models benefited greatly from the end-to-end learning. When using frozen networks, the InceptionResNetv1 features outperformed the MobileFaceNet features (0.413 vs. 0.378), with a far smaller RNN model on top (109$K$ vs 4.4$M$), see table 5.1. This is interpreted as a difficulty to adequately model valence and arousal with the features provided by the smaller MobileFaceNet (Karas, Tellamekala, et al., 2022). However, once the CNN layers are unlocked, the performance is boosted, and the hyperparameter optimisation found an even more compact RNN model at only 76$k$ parameters, a reduction by 30.27% in the seq-2-seq model.

The self-attention models showed a similar trend when going from frozen CNN to end-to-end learning. The average CCC score increased from 0.378 to 0.450 (+0.072), while the parameter count in the transformer encoder stack decreased from 482$K$ to 193$K$. This result represents a considerable reduction of 59.95% in the seq-2-seq component.

The exception to end-to-end optimisation leading to smaller networks was CMA. The best E2E-CMA model gained 300$K$ parameters compared to its counterpart trained with frozen feature extractors, an increase by 14.7% in its cross-modal fusion module. Nevertheless, the end-to-end learning was effective in increasing validation performance, from 0.392 to 0.440 CCC.

From these results, the conclusion is drawn that end-to-end learning is an effective strategy to boost affect recognition on in-the-wild data, at the cost of additional computational resources for training the feature CNN parameters. In the experiments performed herein, using the compact MobileFaceNet and 1D audio CNN helped

reduce this cost. Additionally, training end-to-end led to the discovery of more compact seq-2-seq models, which saves computing power and memory in deployment.

### 5.1.3 Comparison between recurrence and attention-based sequence modelling

When comparing the different sequence modelling architectures in terms of recognition performance, the SA networks using early fusion followed by a transformer encoder and the CMA networks using cross-modal attention layers based on Tsai et al., 2019 delivered comparable results on the validation set, while on the test set, SA performed better (0.386 CCC vs 0.343 average CCC score). At the same time, as shown in table 5.1, the best SA model is much smaller than the best CMA model. The additional complexity introduced by the CMA architecture appears to not be beneficial in this case.

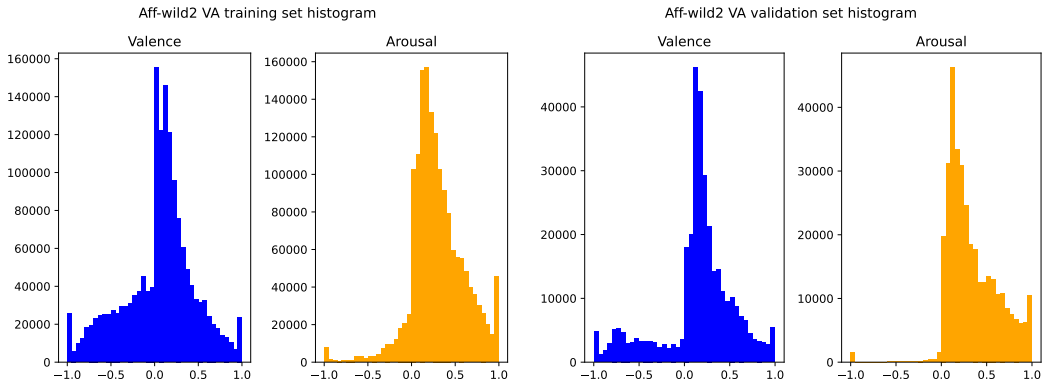
Another somewhat counter-intuitive result is the strong performance of the RNN models compared to the attention-based models. Given that Transformers have displaced recurrent models as state-of-the-art solution in many sequence modelling problems, most notably in NLP, one might also expect a clearly observable advantage on continuous emotion recognition. In this case, the experiments did not show the Transformer models consistently outperforming the RNNs.

An implementation error in the attention-based models is highly unlikely, since the code makes use of pre-defined PyTorch layers and extensive checks were performed. Another possibility is that the transformers have greater potential than the RNNs, but the training process happened to not discover those optimal configurations. This cannot be ruled out, despite the large number of automated trials sampled via Ray. For now, it can be concluded that with the approach used herein, RNNs show similar capability to model valence and arousal on the Aff-Wild2 data as attention models. This may change in future experiments when different feature sets or training strategies are used.

### 5.1.4 Comparison of uni-modal and multi-modal performance

Comparing the models using both audio and visual data with the ablation experiments that only have access to a single modality shows that the multi-modal models consistently outperform their uni-modal counterparts in terms of averaged CCC score. This matches the assumption that the two modalities contain complementary information which helps the model predict the affective state.

Notably, when comparing the results of audio and visual experiments with frozen CNN features in table 4.2, models using the MobileFaceNet features outperform the 1D Audio CNN on both valence and arousal prediction, by a wide margin. For instance, the visual self-attention model yields 0.529 CCC on arousal, while its counterpart only achieves 0.317. This result is counter-intuitive under the assumption that arousal is better predicted from audio, via the voice. However, an explanation is found in the characteristics of the Aff-Wild2 dataset. The data, being collected "in the wild" from diverse YouTube videos, is quite noisy. In the visual modality, while image quality varies greatly, the faces of the subjects are usually visible (some occlusions occur, e. g., from gestures). In the audio modality, there may be disturbances from other sound sources, as well as prolonged periods of silence that provide no information to the model. In particular, a challenging case for purely audio-based models encountered in the dataset is reaction-style content, in which




(A) Training Set (B) Validation set

FIGURE 5.1: Histograms of the valence and arousal annotations on the training and validation sets of the Aff-wild2 dataset.

the subject watches a different video while being silent most of the time. Possible steps to address these issues are extensive preprocessing to separate the subjects' voices from background noise, and removing extended silent sections, training only with data that contains useful information. Another promising strategy is the use of a more complex feature extractor, such as an audio Transformer. As demonstrated by the results in section 4.3, models like WAV2VEC2 are well suited for handling noisy audio data.

### 5.1.5 Performance discrepancy between validation and test sets

Another remarkable result is the change in model performance between validation and test sets. While all architectures achieved averaged CCC scores above 0.4 on validation, none did so on test (with the exception of the final ensemble). Comparison of table 4.3 and table 4.5 shows that the drop in overall score is due to arousal decreasing considerably, while valence increased for RNN and SA models. The exception is the CMA model, which also suffered a loss in valence prediction performance (0.327 vs. 0.388 CCC).

With the test labels hidden, the exact cause of this effect cannot be determined. Possibly, the statistical properties of the dataset lead to arousal being easier to predict on the validation set than on the test set. Histograms of the training and validation labels are shown in fig. 5.1. During training, the hyperparameter optimisation process will prefer configurations well suited to predicting the distribution of the validation set. The validation set contains mostly samples with positive arousal, peaking around 0.1. The test data may show a wider distribution with more instances having high or low arousal scores.

### 5.1.6 Comparison with the field and limitations of the approach

In this final subsection, the results obtained on Aff-Wild2 are compared to other works in the literature addressing valence-arousal estimation on that dataset.

As stated in section 4.1, once the best-performing models were identified, test set predictions of those models and and of an ensemble combining them were entered into the third ABAW competition. The submission code can be found on Github [1],

[1] https://github.com/VincentKaras/abaw3_rnn_attn

and the corresponding paper is Karas, Tellamekala, et al., 2022.

A total of 33 teams participated in that challenge, and 16 submitted test set predictions. The results presented herein achieved fifth place.

The winner of the ABAW'22 challenge was Meng et al., 2022, who achieved valence CCC of 0.606 and arousal CCC of 0.596 on the test set. Their approach was similar to the one used herein, using early fusion followed by either LSTM or transformer encoder, as well as strided sampling of the dataset.

There is a considerable performance gap between the results of the two papers, despite the fundamental similarity of the models. A possible reason for this are the feature choices. Meng et al., 2022 used the activations of a DenseNet and an InceptionResNet100, each pre-trained on multiple facial emotion datasets including FER+ and AffectNet (Mollahosseini, Hasani, and Mahoor, 2019) as visual features. FAUs were also extracted. In the audio modality, eGeMaps, ComParE, VGGish (Hershey et al., 2017) and wav2vec2 were used. It seems that the combination of these handcrafted and deep features allowed the model to discover more useful information.

The other likely reason for the performance gap lies in Meng et al., 2022 choosing a larger sequence length of 100, giving their models more temporal context. In hindsight, the sequence length was set too low for the experiments described herein. Originally this choice was made to conserve memory in the parallelised search trials.

In the fifth ABAW challenge (Kollias, Tzirakis, Baird, et al., 2023), the same team won again, this time achieving CCC scores of 0.619 on valence and 0.6634 on arousal. Once again, a similar feature set was combined via early fusion, and processed with Transformers or RNNs, as well as a combination thereof. An ensemble strategy was also employed to further boost performance.

The conclusion from this is that the approach used in this thesis for predicting continuous emotions is sound in principle, but further optimisation is needed. Once this is performed, competitive results on Aff-Wild2 are expected to be reached. Another clear advantage is that the approach is easily transferable to other datasets.

Nevertheless, the findings of the analysis still stand: Smaller feature extractors can be competitive with larger models, end-to-end learning is beneficial as is forming model ensembles, and recurrent encoders are still competitive with attention-based ones. These conclusions were used in the experiments for contributions C–2 and C–3, see section 4.3 and section 4.2 respectively.

## 5.2 Vocal Burst Emotion Recognition

In this section, the results of the vocal burst type classification, two-dimensional affect, emotions and culture specific emotions prediction experiments performed on the A-VB dataset in section 4.3 are discussed. Together they form contribution C–2 and are primarily used for answering research question RQ–4, and to a lesser extent RQ1–3.

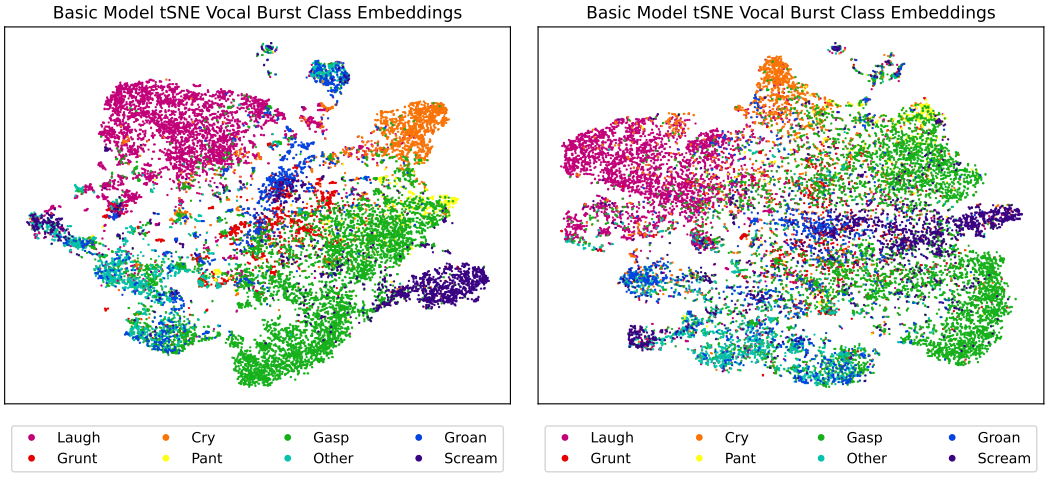
### 5.2.1 Feature embedding analysis

In order to better comprehend how the Transformer-based models solve the four tasks, features extracted by the WAV2VEC2 backbone are visualised using *t-distributed stochastic neighbour embedding (t-SNE)*.

For the basic MTL and classifier chain architectures, the activations of the last layer are extracted. For the branching model architecture, the same is done to each layer that serves as a branching point and input to the multi-head attention stack. The derived feature representations are then averaged over time, so that each sample is represented by a single vector of size 768.

The scikit-learn toolkit [2] is used to compute the t-SNE embeddings from the features. Following the recommendations of the algorithm's authors, the high dimensional features are first reduced to $N = 50$ by *principal component analysis (PCA)*.

Due to the test set labels of Hume-VB being hidden, this analysis is restricted to the training and validation partitions.



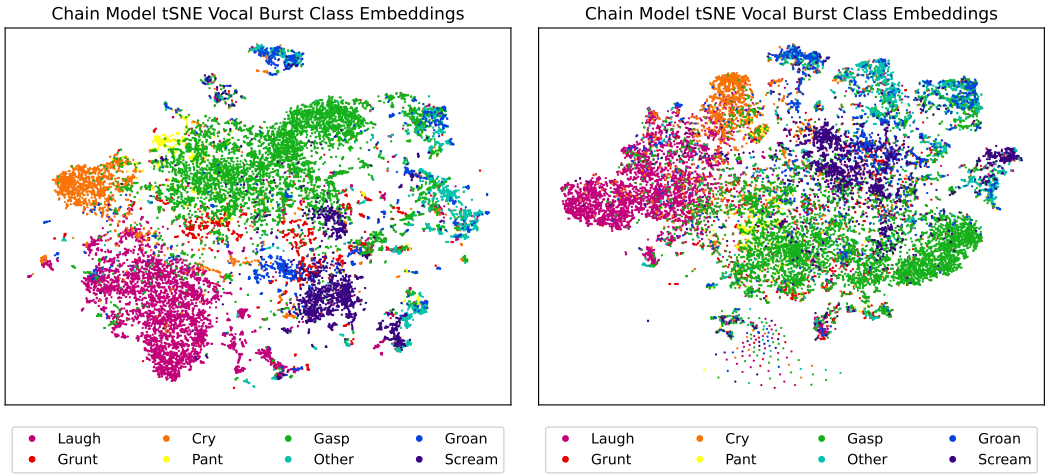

(A) Training Set (B) Validation set

FIGURE 5.2: t-SNE visualisation of the features extracted by the best performing basic MTL model. Data points are coloured for each of the 8 annotated classes of the A-VB-TYPE task.

For the visualisation of the A-VB TYPE task, each type of vocal burst is represented as a distinct colour, leading to a scatter plot with 8 classes. The t-SNE embeddings of the basic MTL model are depicted in fig. 5.2. For comparison, t-SNE plots of the best performing classifier chain and branch architectures are shown in fig. 5.3 and fig. 5.4, respectively.

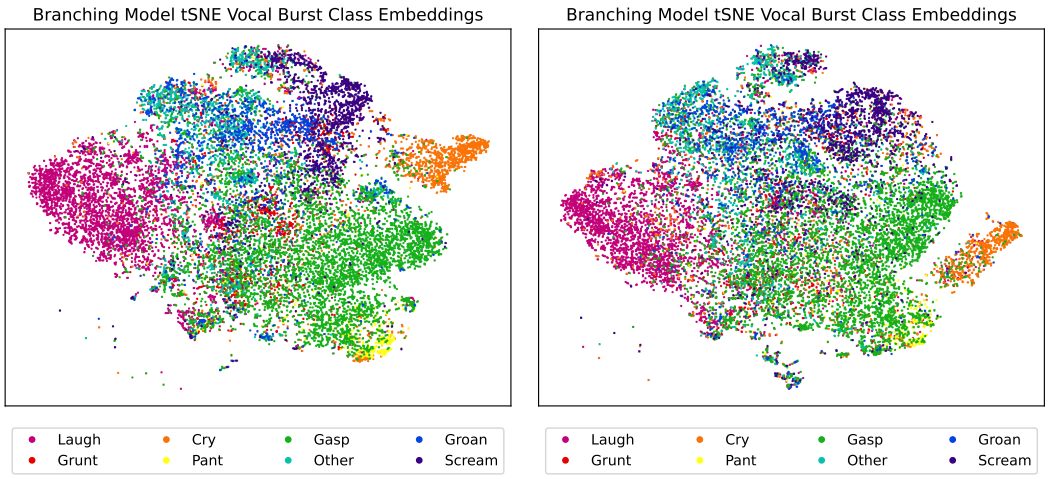
The plots show distinctive regions of uniform colour, indicate that the fine-tuned WAV2VEC2 model has learned to cluster at least some of the classes. In particular, Laugh (magenta), Cry (orange), Gasp (green), and Scream (indigo) are grouped in fairly well separated regions. Notably, this group consists of the largest classes in both the training and validation sets. By contrast, the data points for Grunt (red), Pant (yellow), Other (cyan), and Groan (blue) appear more dispersed. It is also worth mentioning that the separations between the various regions appear more distinct

[2] https://scikit-learn.org/stable/index.html




(A) Training Set (B) Validation set

FIGURE 5.3: t-SNE visualisation of the features extracted by the best performing classifier chain model. Data points are coloured for each of the 8 annotated classes of the A-VB-TYPE task.




(A) Training Set (B) Validation set

FIGURE 5.4: t-SNE visualisation of the features extracted by the best performing branching attention model. Data points are coloured for each of the 8 annotated classes of the A-VB-TYPE task.

for the basic MTL and classifier chain models than they do for the branching model, with the exception of the Cry cluster.

In the cases of the A-VB LOW, A-VB HIGH and A-VB CULTURE tasks, since the annotations are continuous, an additional sorting step is needed to assign distinctive colours to the samples.

For A-VB LOW, based on the quadrants of the arousal-valence circumplex (see section 2.1), the two affect dimensions are each discretised into two value ranges. For valence, these are $[0, 0.5[$ and $[0.5, 1.0]$ , for arousal they are $[0, 0.75[$ and $[0.75, 1.0]$ respectively. The ranges for arousal are chosen asymmetrically since annotations below 0.5 practically do not occur, see fig. 5.11. This results in 4 categories. The results for the basic MTL model are depicted in fig. 5.5.

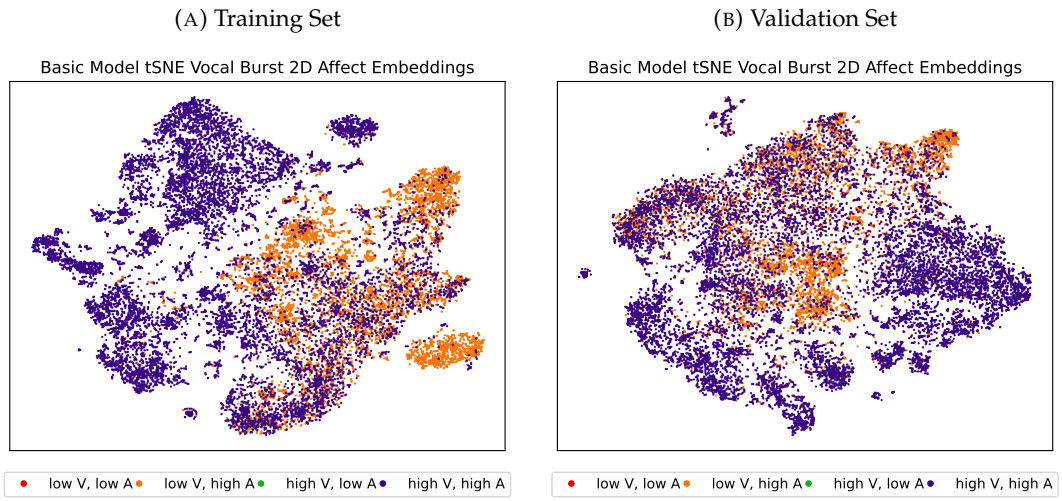



FIGURE 5.5: t-SNE visualisation of of the features extracted by the best performing basic MTL model on the A-VB-TWO task. Samples are coloured based on 4 clusters (high valence – high arousal, high valence – low arousal, low valence – high arousal and low valence – low arousal.

In this case, some similarities with the corresponding plots of the A-VB-TYPE task are visible. For instance, the cluster for laughter is assigned high valence–high arousal, and the cluster for screaming is assigned low valence–high arousal. Besides this, there are large regions where the classes intermingle. Changing the discretisation to 9 clusters increased the effect, making the plots look quite noisy. The conclusion from this is that while some general trends are visible, valence and arousal vary considerably locally. The plots for the other architectures are omitted here, as they provide no additional information, and the focus is instead placed on the two remaining tasks.

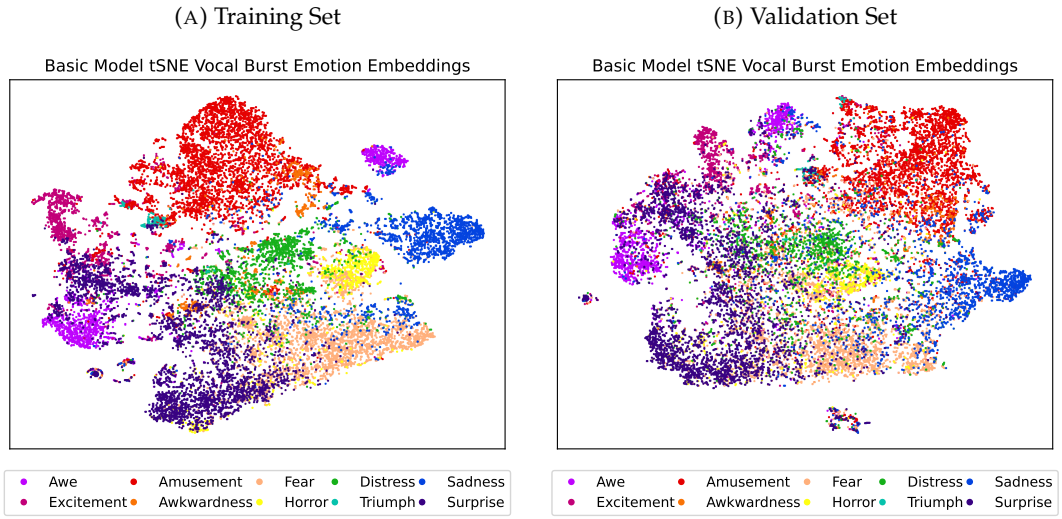
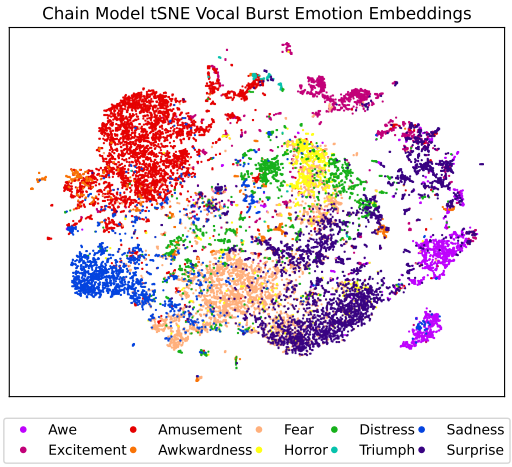



FIGURE 5.6: t-SNE visualisation of features extracted by the best performing basic MTL model for the 10 annotated emotions in the A-VB-HIGH task. Samples are assigned to a category by their dominant (highest-rated) emotion.

Following Baird, Tzirakis, Brooks, et al., 2022, for A-VB HIGH and A-VB CULTURE, the dominant i. e., highest rated emotion is chosen to assign the samples to a colour,

(A) Training Set




(B) Validation Set

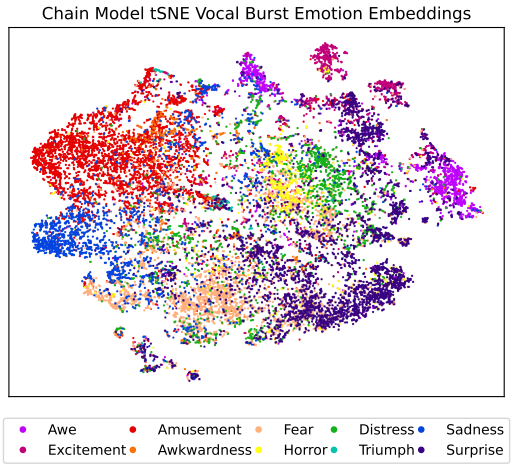
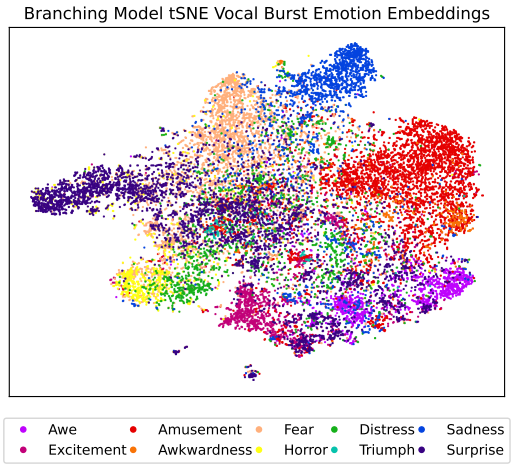



FIGURE 5.7: t-SNE visualisation of features extracted by the best performing classifier chain model for the 10 annotated emotions in the A-VB-HIGH task. Samples are assigned to a category by their dominant (highest-rated) emotion.

(A) Training Set




(B) Validation Set

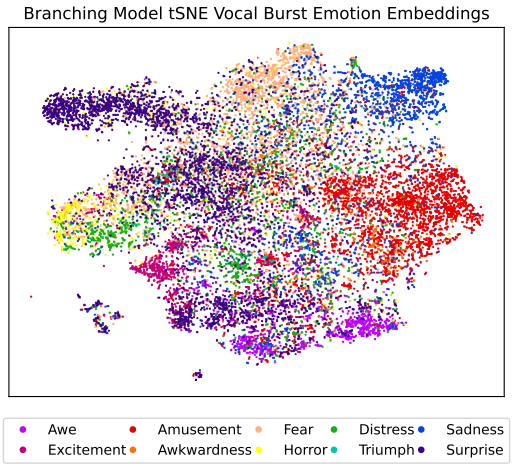



FIGURE 5.8: t-SNE visualisation of features extracted by the best performing branching attention model for the 10 annotated emotions in the A-VB-HIGH task. Samples are assigned to a category by their dominant (highest-rated) emotion.

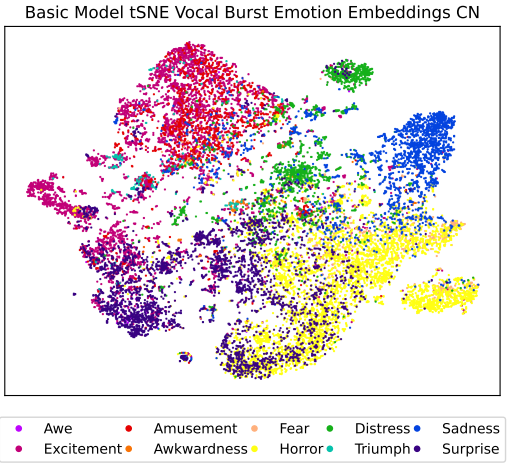
which results in scatter plots with 10 classes (in the A-VB-CULTURE task, different cultures are visualised in separate sub-plots).

The results for the A-VB-HIGH task are depicted in fig. 5.6 for the basic MTL model, fig. 5.7 for the classifier chain model, and fig. 5.8 for the branching attention model, respectively. It can be observed that the model clusters the samples into regions corresponding to specific dominant emotions, e. g., the purple, red and blue clusters in fig. 5.6 corresponding to awe, amusement, and sadness, respectively. Also, it is visible that regions of emotions blend into each other, see for instance the locations of fear (peach), horror (yellow) and distress (green) in fig. 5.6. This smooth variation can be explained from the gradual transition of one dominant emotion to the next. It matches the findings of A. S. Cowen and Keltner, 2017, see section 2.1.

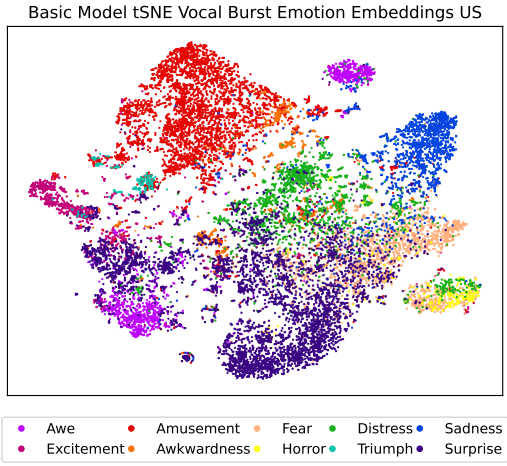
Comparing the plots by architecture, clusters derived by the basic and chain models

appear tighter and better separated, while those for the branching model are more dispersed. Awkwardness (orange) and triumph (cyan) are hardly observable in any of the plots. An explanation for this is that samples where these subtle emotions dominate are underrepresented in the data (they account for 3.782% and 0.515% of the training set, respectively).

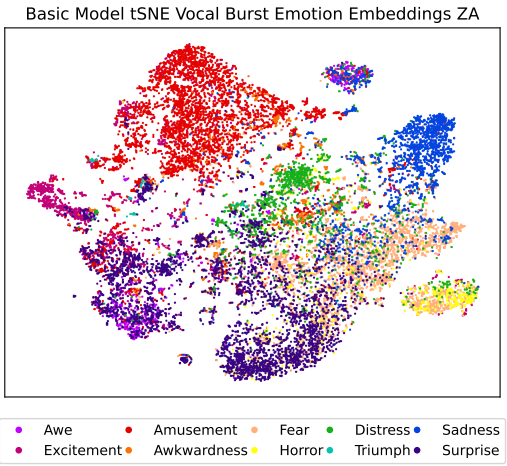
(A) China




(B) United States




(C) South Africa




(D) Venezuela

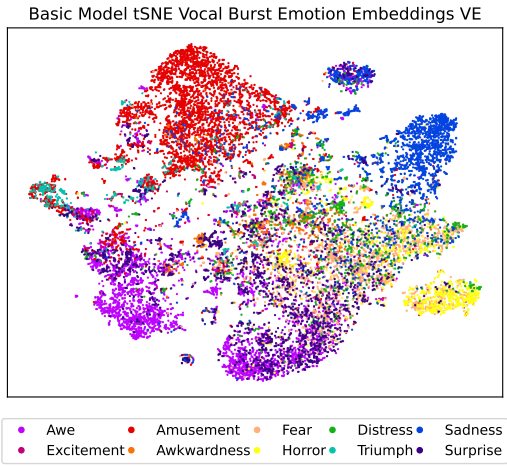



FIGURE 5.9: t-SNE visualisation of the training data set for the best performing basic MTL model on the A-VB-CULTURE task. Samples are coloured by their dominant emotion in the annotations of the respective culture.

Visualisations for the emotions of the four cultures in the A-VB-CULTURE task are given in fig. 5.9 and fig. 5.10, for the training and validation sets respectively. Depicted are the features learned by the basic MTL model. Notably, there are some clearly visible similarities and differences between the cultures.

In terms of similarities, samples that are dominantly sadness (blue) appear in a distinct region on the top right side across all cultures. The plots for the United States and South Africa appear quite similar in their clustering of the emotions, see the regions for amusement (red), surprise (indigo), and the grouping of negative emotions from distress (green) to fear (peach) and horror (yellow).

By comparison, the embeddings for Venezuela appear less uniformly clustered, and

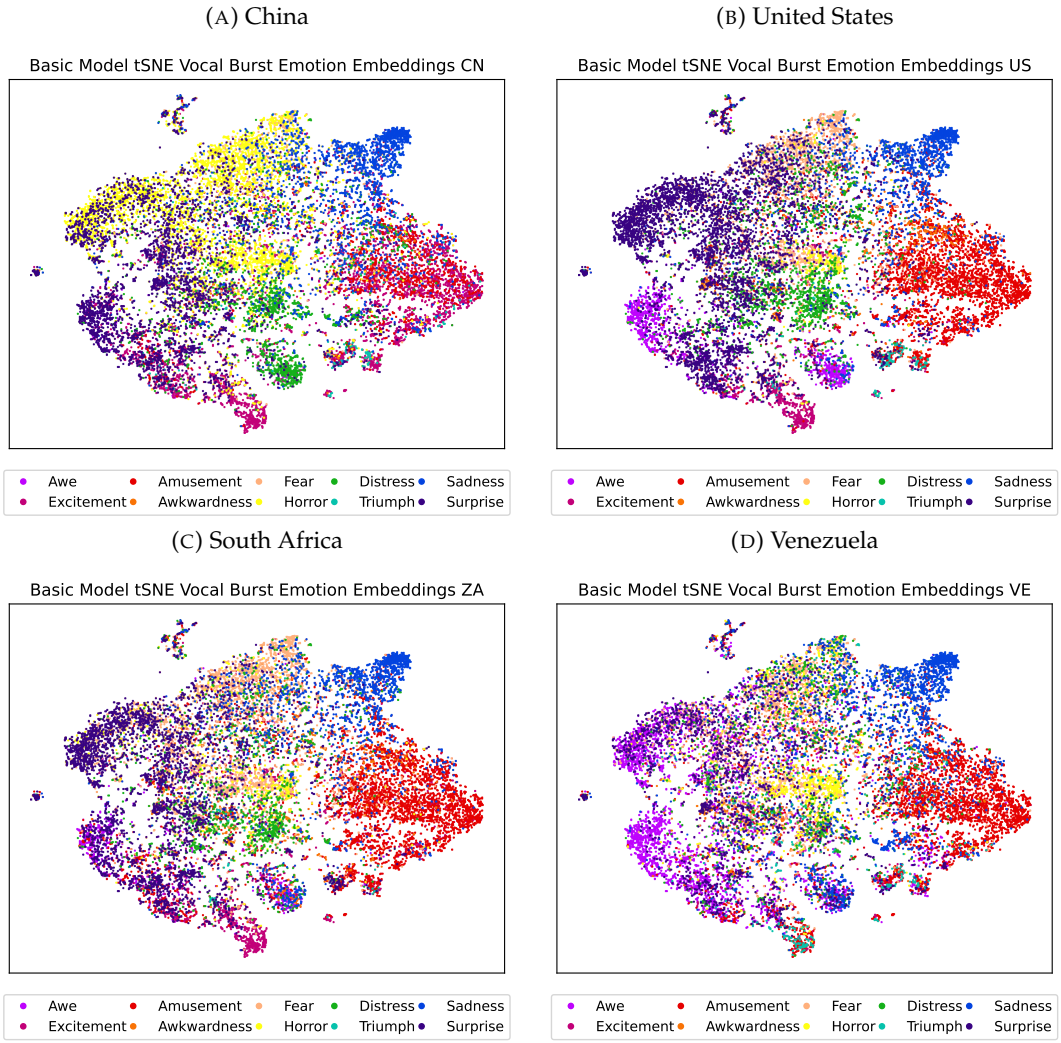



FIGURE 5.10: t-SNE visualisation of the validation data set for the best performing basic MTL model on the A-VB-CULTURE task. Samples are coloured by their dominant emotion in the annotations of the respective culture.

more samples are assigned to awe (purple) at the expense of surprise. For the Chinese culture, many samples that are amusement dominant in other cultures are assigned to excitement (magenta), and the two clusters intermingle. In addition, for many samples that are rated dominantly as fear by other cultures, horror is rated as the strongest emotional attribute in Chinese.

These results indicate that while there are general similarities in how members of different cultures perceive affect in non-verbal vocalisations, nuanced differences exist regarding the emotional concepts. This reinforces the need to study emotion recognition in cross-cultural settings, see RQ–2, and also validates the approach of modelling emotions as continuous quantities that vary smoothly.

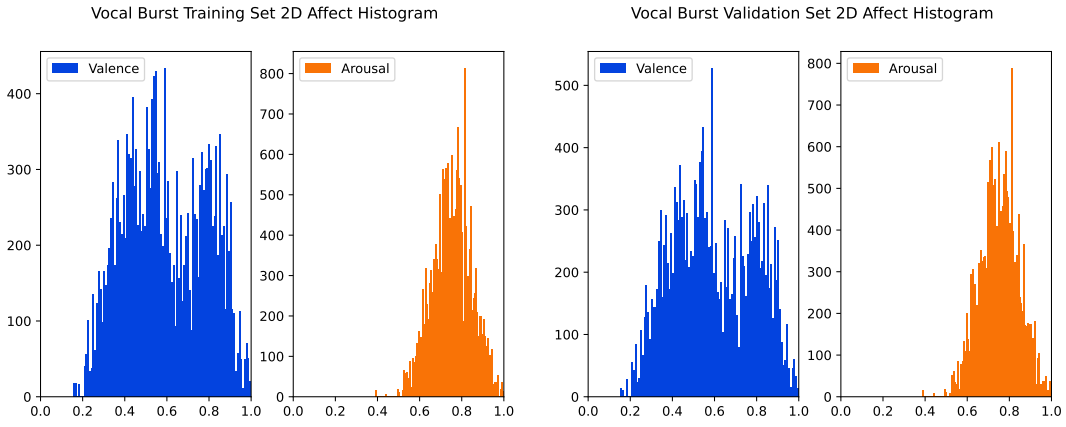
### 5.2.2 Recognising affect in vocal bursts

In this subsection, the performances of the models on each of the four tasks as discussed, along with observed general trends and limitations of the chosen transformer-based approach.

### Vocal Burst Classification

For the A-VB-TYPE vocal burst classification task, when the 8 individual class recalls are examined, the models show that some classes are recognised considerably better than others, see fig. 4.4, fig. 4.7, fig. 4.10 for the basic, chain, and branching models respectively. Laughter achieves the best results with recall above 0.8, with gasping being the runner-up at around 0.7, depending on architecture and loss balancing. Screaming and the "Other" categories achieve a recall of approximately 0.6, while the remaining classes perform considerably worse, rarely exceeding a score of 0.5. The Groan class shows the worst performance, with recalls around 0.4.

This behaviour remains fairly consistent across model runs, as well as different classifier architectures and loss balancing strategies. These results indicate that the model performance is impacted by the imbalance in the dataset, as the two best-performing classes account for more than 50% of the training data. This is supported by the t-SNE analysis, which demonstrated that the model has focused on clustering the data points of the majority classes.




(A) Training Set (B) Validation set

FIGURE 5.11: Histograms of the valence and arousal annotations of the training and validation sets for the A-VB-TWO task.

### Valence-Arousal Estimation

In the A-VB-TWO dimensional affect recognition task, the models show a considerable difference between valence and arousal prediction, with the validation CCC for valence usually exceeding 0.75 while arousal remains below 0.65. This trend is observed in all model architectures and loss weighing methods used for the experiments (fig. 4.5a, fig. 4.11a, fig. 4.8a), and remains consistent across multiple runs with different random initialisations of the downstream model layers.

Following the reasoning above for the A-VB-TYPE task, the discrepancy could be due to dataset properties. Given that A-VB-TWO is a regression problem, histograms for the valence and arousal annotations are computed. They are shown in fig. 5.11 for training and validation sets respectively. The distribution for valence is clearly wider than that for arousal ($\sigma = 0.193$ vs $\sigma = 0.093$ on the training set). For both affect dimensions, the distributions closely resemble each other for training and validation sets. In case of arousal, the mean and standard deviations are $\mu = 0.755$, $\sigma = 0.093$ and $\mu = 0.752$, $\sigma = 0.092$ on training and validation data respectively.

Thus the lower performance on arousal cannot be explained as a consequence of the dataset structure.

Given that audio-based models have traditionally performed better on arousal prediction, the results are somewhat unexpected. However, it has already been demonstrated that transformers, including WAV2VEC2, can achieve strong performance on valence in SER (Wagner et al., 2023). While the model demonstrates its capability of predicting valence, the vocal burst data is considerably different from the speech corpora the transformer backbone was trained on, which may negatively impact arousal.

The suspected root cause of the issue is that arousal is simply harder to predict from vocal bursts. Baird, Tzirakis, Brooks, et al., 2022 arrive at this conclusion based on the short duration of the samples compared to the longer utterances commonly used in SER, and on the diversity of the recording environments. Despite the relative loss of performance on arousal, the chosen approach is still clearly effective.

#### Emotion Estimation

For the A-VB-HIGH task of predicting the intensities of 10 emotions, differences between the classes are apparent, see fig. 4.5b,fig. 4.8b, fig. 4.11b for the three model architectures. The best results are achieved for awe, amusement and surprise, around $CCC = 0.8$. Fear and horror also perform relatively well, at approximately 0.76 and 0.74 respectively for the basic MTL (vanilla) model architecture. They are followed by excitement, distress, triumph and sadness, approaching but not exceeding $CCC = 0.7$. Finally, the worst performance is awkwardness, at slightly lower than $CCC = 0.6$. This pattern is also seen with the other model architectures, and reproduces across runs.

A possible explanation for this behaviour is that some of the more subtle emotions, like awkwardness and triumph, are simply harder to detect from vocal bursts compared to e. g., horror and surprise. This fits with the models being more capable of predicting valence than arousal, as evidenced by the results on the A-VB-TWO task. However, it does not explain why awe and amusement outperform excitement by around 0.1 CCC. Another, explanation is found in predominantly awkward or triumphant samples being underrepresented in the dataset, as seen in the t-SNE analysis above. This will negatively impact the model's success at predicting those emotions.

#### Cultural Emotion Estimation

For the 4-country A-VB-CULTURE task, the results in terms of averaged CCC fall considerably below those of the A-VB-HIGH task on both the validation and test partitions, cf. table 4.27, table 4.28.

Given that the labels of this task represent the culture-specific gold standards of the respective cultures that contributed to the dataset (Baird, Tzirakis, Brooks, et al., 2022), one might assume that the models have difficulty predicting the three out-of-culture annotations on the samples. However, examining the CCC scores of on a per-culture basis shows that the performance is not uniformly worse compared to the A-VB-HIGH task.

Instead, comparison of CCC scores achieved by the basic MTL models in fig. 4.5b and fig. 4.6 shows that for the United States, results are quite similar to A-VB-HIGH .

On the contrary, South African, Venezuelan, and Chinese show worse performances, leading to the overall lower average scores. The same pattern can be seen in the results of the classifier chains and branching models, cf. fig. 4.8b, fig. 4.9, and fig. 4.11b, fig. 4.12, respectively.

TABLE 5.2: Pearson correlation coefficients between the culture specific emotion scores from A-VB-CULTURE and the emotion scores from A-VB-HIGH , calculated for the best performing models of each of the three architectures (basic MTL, classifier chain, branching attention).

| **Model** | **China** | **United States** | **South Africa** | **Venezuela** |
|---|---|---|---|---|
| | $\rho$ | $\rho$ | $\rho$ | $\rho$ |
| *VANILLA* | .027 | .8719 | .8013 | .7461 |
| *CHAIN* | -.0422 | .8932 | .7381 | .7147 |
| *BRANCH* | .0495 | .8789 | .7810 | .7521 |

Pearson correlation scores between the 10 emotions' results on the A-VB-HIGH task and the 4 groups of scores on the A-VB-CULTURE task are given in table 5.2. These are calculated on the top validation set results for each of the three architectures. It can be seen that the United States scores show the highest per-model correlation with the 10 emotions task. For South Africa and Venezuela correlations are lower, and China does not correlate at all.

A possible explanation for these inter-cultural performance discrepancies, with the US performing best, could be that the transformer model was pre-trained on an English-language dataset (LibriSpeech). However, the same deficit in prediction performance is also reported in Baird, Tzirakis, Brooks, et al., 2022, who used EGEMAPS, COMPARE, and randomly initialised CNN-LSTMs from the END2YOU toolkit i. e., models which had no language-specific pre-training.

Given that the number of speakers in the training and validation sets from the USA and South Africa (206 and 244), both of which have English as an official language, is much greater than those of China and Venezuela (79 and 42, 76 and 42 on training and validation sets respectively), it is more likely that the main cause is the imbalance of the dataset, see table 3.3.

Another hypothesis to explain these results can be based on cultural distinctions of how emotions themselves are perceived and expressed. It is worth noting that the 10 labelled emotions are all taken from the English language. Thus the annotations from native English speakers may be easier to model, whereas cultural differences make Venezuelan and Chinese emotions harder. This point has been argued in the A-VB challenge baseline paper by Baird, Tzirakis, Brooks, et al., 2022. It may also explain why some individual emotions are still recognised well in the worse-performing cultures, e. g., CCC for Horror in Chinese outperforming the results for the United States. Due to culture-specific differences in how these emotions are experienced, the annotations are impacted, and the model may have an easier or harder time to learn them, see the t-SNE analysis for A-VB-CULTURE above.

**Comparison of configurations for the branching and classifier chain architectures**

Regarding the branching model architecture accessing multiple layers' embeddings of the transformer backbone, the experiments summarised in table 4.25 showed no statistically significant difference between model runs A (last 4 layers) and B (even

numbered layers). One sided t-tests with $\alpha = 0.05$ result in p-values of 0.129, 0.178, 0.26, and 0.148, for the average CCC scores on the four tasks, respectively. However, model C (first 4 layers), showed significantly worse performance than model A, with one-sided p-values of $6.4 * 10^{-8}$, $1.7 * 10^{-4}$, $1.6 * 10^{-6}$ and $5.3 * 10^{-6}$ respectively. From this, it can be concluded that the higher layer embeddings of the transformer model are better suited for predicting the vocal burst types and emotions. In addition, it provides an explanation for why the basic MTL architecture, which uses only the final layer's hidden states, performs comparable to or above the more complex branching architecture.

For the various classifier chain models, in order to compare their performances on the validation sets (cf. table 4.23), the simplest variant A (chaining the task-specific heads, no internal chaining) is measured against the variants B-F.

Model B, which also does not use task-internal chaining and predicts A-VB-HIGH and A-VB-CULTURE in parallel instead of sequentially, did not show statistically significant differences at $\alpha = 0.05$.

Model C, which uses internal chaining in the A-VB-CULTURE task and predicts that task in parallel to A-VB-CULTURE showed no statistically significant differences in the affect tasks, however, the A-VB-TYPE task was significantly worse ($p = 0.002$ at $\alpha = 0.05$).

Model D, which chains the A-VB-HIGH task internally, showed statistically significant ($\alpha = 0.05$) worse performance compared to model A on both that task and A-VB-CULTURE ($p = 5.6 * 10^{-}13$ and $p = 1.7 * 10^{-}34$, respectively). Notably, performance on the latter task did not exceed CCC scores of 0.12 on the validation set, compared to the approximately 0.6 normally achieved.

Model E, which predicted the A-VB-HIGH and A-VB-CULTURE tasks in parallel but internally chained both of them, performed significantly worse on A-VB-HIGH at $p = 4.5 * 10^{-}13$ and $\alpha = 0.05$.

Finally, model F, which was structured like model E but did not order the emotions in the chains by descending order of performance on the basic MTL model, showed no statistically significant difference to model A ($\alpha = 0.05$).

From these results, it can be concluded that more elaborate attempts at chaining the predictions were not beneficial or even harmful to model performance. Chains that are too long apparently lead to degradation on the following tasks, as evidenced by the failure of model D to predict A-VB-CULTURE Ḣowever, it is noteworthy that a chain of no less than 40 emotions produced no worse results on A-VB-CULTURE for model F than the prediction in a single layer with 40 outputs did for model A.

**General performance and limitations of the approach**

The analysis presented above was focused on the validation set, since the test labels of Hume-VB are hidden. In order to better gauge the effectiveness of the proposed approach, and to compare with the rest of the field, the best-performing models and an ensemble thereof were submitted to the organisers of the ACII'22 A-VB competition and workshop, see table 4.28.

Methods based on deep audio feature extraction, including fine-tuning WAV2VEC2 or related models, featured prominently in the workshop proceedings[3].

[3]https://arxiv.org/html/2210.15754/

TABLE 5.3: Results of the ACII'22 A-VB Competition. Shown for each of the four tasks are the results of two baselines (using handcrafted COMPARE features and CNN-RNN trained end-to-end with END2YOU ) provided by the organisers (Baird, Tzirakis, Brooks, et al., 2022), the scores of the winners and runner-ups, and the results of the approach presented in this thesis, first published in Karas, Triantafyllopoulos, et al., 2022. The latter were outside the official rankings due to affiliation with the organisers.

| **Team** | **A-VB-TYPE** UAR | **A-VB-TWO** CCC | **A-VB-HIGH** CCC | **A-VB-CULTURE** CCC |
|---|---|---|---|---|
| COMPARE | .3839 | .5214 | .4986 | .3887 |
| END2YOU | .4172 | .5686 | .5084 | .4401 |
| Winner | **.5856** | .7295 | .6854 | .6017 |
| Runner-Up | .519 | .7237 | .629 | .5495 |
| Ours | .5618 | **.7363** | **.7066** | **.6195** |

The leaderboards of the best-performing submissions on each of the four tasks can be found on the competition's website (Hume AI, 2022). The results are summarised in table 5.3. They show that the baseline was surpassed by a wide margin. The submission based on the approach presented here was excluded from the official rankings due to affiliation with the competition organisers. It performed very well, beating the winning teams on the A-VB-TWO , A-VB-HIGH and A-VB-CULTURE , and only falling below on the A-VB-TYPE task (0.5618 vs 0.5856).

These results show the method for vocal burst analysis used herein to be highly effective. A fine-tuned WAV2VEC2 provides powerful features for distinguishing both vocal burst types and emotions, while being originally trained on substantially different dataset containing English speech. Using different downstream architectures like chained prediction heads, and combining them into an ensemble strategy, further benefits the overall emotion recognition capabilities of the models. Furthermore, using advanced loss balancing strategies, including uncertainty measures and dynamic weighting during training, was helpful in obtaining additional performance.

The limitations of the approach used for C–2 include its ability to predict underrepresented types of vocal bursts. As noted above, A-VB-TYPE was the only task where it failed to outperform the A-VB competition winners (although it still beat the runner-up). It was also here that an ensemble failed to yield any benefits, see table 4.28. If additional steps like oversampling the dataset or deliberate optimisation towards minority classes were taken, performance could likely be improved.

Another limitation is the performance on culture-specific emotions. As noted above, Chinese emotions and Venezuelan emotions are challenging to predict, in addition to the cultures being underrepresented in the dataset. While the overall results on A-VB-CULTURE are still strong, additional measures could be taken to boost culture-specific performances. Domain adaptation for cross-cultural emotion recognition will be discussed in the following section.

## 5.3 Cross-Cultural Emotion Recognition

In this section, the set of experiments on the SEWA dataset from section 4.2 will be discussed. First, the supervised baseline results are analysed. Next, the semi-supervised DANN results are interpreted.

### 5.3.1 Supervised baseline

The multi-modal baselines trained on the complete source culture data are first discussed, followed by the ablation to uni-modal computation, and the effects of restricting the multi-modal models to training on fractions of the source data.

#### Audiovisual CNN baseline models

The experiments using CNN feature extractors and GRU or self attention with GRU as seq-2-seq encoder show several important results. A shown in table 4.7, models trained on German data tend to achieve higher validation set scores on both German and Hungarian than their counterparts trained on Hungarian videos. This holds for both arousal and valence, with the exception of the German-trained CNN-GRU model, which achieved only a top valence score of 0.376 on Hungarian, while the Hungarian models obtained a top score of 0.488.

On the test set, German trained models achieve the highest arousal (0.641 ) and valence (0.642) CCC scores on German data, as expected. Performance drops when predicting arousal on Hungarian data (0.516), and even more considerably on Chinese (0.391). On valence there is also a drop, to 0.443 CCC on Hungarian and 0.46 CCC on Chinese respectively. This demonstrates the effects of domain shift, the model has issues to recognise affective displays from cultures other than the one it was trained on (the source domain).

The Hungarian-trained models show lower performance on their own culture for the validation set compared to German, which is a counter-intuitive result. On the test set however, the models are strongest on their own culture, with 0.521 on arousal and 0.541 on valence respectively. Performances outside the source domain are strong for arousal (0.493 for German and 0.477 for Chinese). However, on valence, the Chinese predictions are much worse, obtaining only 0.34 CCC.

Based on these findings, the interpretation is that the German data allows the models to learn more effective representations, generalising better to unseen videos than models trained with Hungarian. Another possible cause is found in table 3.2. While German and Hungarian contain identical numbers of subjects in the training and validation partitions, German contains more video footage on both, approximately 35% and 30% more respectively.

#### Ablation study: uni-modal training

Performing an ablation from multi-modal to uni-modal models shows several expected and some unexpected effects. First, it becomes clear that the MobileFaceNet features used in the visual modality are quite powerful, as the visual-only model still shows strong performance, see table 4.8. This matches the results in section 4.1. On the validation set, the German-trained model drops to 0.684 and 0.54 CCC on arousal for German and Hungarian respectively. A drop in arousal is expected, as the audio modality is assumed to provide helpful information here. On the test sets,

for arousal the performance on German and Hungarian also drops, to 0.584 and 0.502 respectively. Interestingly, the arousal score for Chinese shows an increase for the models using self-attention with GRU, while dropping for the pure GRU model. For valence prediction, the German-trained models also exhibit a reduction in performance, to 0.699 and 0.441 respectively. On the test set, there is a slight increase for German (0.654), no change for Hungarian (0.443) and a drop for Chinese (0.429) CCC, respectively. The German-trained models behave mostly as expected under ablation, however the Hungarian models present a different picture. On arousal, the top validation scores actually increase, to 0.583 and 0.5 for German and Hungarian respectively. For valence, the German score increases to 0.639 and the Hungarian score drops to 0.442. On the test set, German performance increases on arousal (to 0.521) and even more strongly on valence (0.619). On the Hungarian test data, arousal performance drops, while valence improves, though no by a wide margin. Finally, Chinese improves on arousal prediction to 0.517 CCC and remains at similar performance on valence (0.334 CCC).

One might expect that the removal of information provided by the audio modality will have a harmful impact, in particular on arousal prediction across the test cultures. Performance increasing on the Chinese videos when audio is removed is likely due to a statistical effect, and further optimisation searches are expected to yield a multi-modal model that is strictly better than its visual-only counterpart in all respects.

Going the opposite route and ablating the visual features also shows that the current approach performs less well when restricted to the audio modality. Training on audio only with the features extracted by the light-weight 1D-CNN shows a considerable drop in performance. As evidenced in table 4.9, for models trained on German data the validation scores of arousal and valence on the same culture reach $CCC = 0.441$ and $CCC = 0.383$, respectively. On Hungarian validation data, a top arousal score of 0.27 CCC is reached, while on valence, the model fails to predict entirely ($CCC = 0.0$). Evaluating on the test set, the German culture yields arousal and valence scores of 0.359 and 0.332 CCC, where the multi-modal model achieved over 0.6. As expected, the performance outside the source domain decreases further, with Hungarian not achieving a score above 0.25 and Chinese not reaching $CCC = 0.2$ on arousal, and only marginally surpassing that result for valence.

In order to make a fair comparison between audio and visual modalities, the nature of the dataset needs to be taken into account. As SEWA consists of dyadic conversations, but the labels refer only to the speaker, the visual model will have speaker information always available via the face (some occasional detection failures in the data notwithstanding), while the audio model will only have the interlocutor voice over extended periods of time, while the speaker is being silent.

Nevertheless, it becomes clear that the 1D-CNN audio features are not the optimal choice, at least not in a uni-modal setting, when comparing models trained on them to the results obtained with the fine-tuned WAV2VEC2 audio transformer. As shown in table 4.9, models trained with these features surpass their 1D-CNN based counterparts by a wide margin. The German-trained models express strong performance in-domain (validation scores of 0.61 and 0.617 as well as test scores of 0.527 and 0.581 CCC on arousal and valence respectively). While the performance drops on Hungarian and Chinese test sets, the models still show some ability to generalise out of domain. Meanwhile, the Hungarian-trained models generalise reasonably well

on German and Hungarian test sets. Performance on Chinese is reduced for both arousal and valence.

In order to maintain fairness in the following discussion, only models that make use of the same feature extractors will be compared.

**Effect of limited source label availability on the baseline**

When the amount of labelled data available for training decreases, the ability of the models to generalise is expected to be negatively impacted. This degradation of performance can be seen for the baseline models in table 4.10, where the validation set CCC scores tend to decrease with the amount of available data, irrespective of the culture the model is trained on. For instance, when examining the German validation CCC of arousal, the best models trained with German as source culture achieve 0.767 using the complete data, 0.729 at 75%, 0.719 at 50% and 0.717 at 25%. However, it is worth mentioning a peculiarity in the results, regarding the extent of the performance drop across cultures and partitions. Going back to the above example, when looking at the German test data, the results do not seem to change: The top CCC scores are 0.621 at 100%, 0.615 at 75%, 0.621 at 50% and 0.619 at 25%. While the validation performance suffered, the ability of the models to generalise appears to not be impacted, at least not in-domain. For the other test set cultures, a performance drop is visible.

In some cases, the top scores actually increase when the amount of labelled data decreases. As the increases are mostly minor, this is interpreted as a statistical effect of the optimisation process. If even more training runs were performed, the models trained on more data are expected to consistently outperform those trained on smaller subsets.

### 5.3.2 Domain Adversarial Neural Networks

Next, the results of the DANN experiments are interpreted and compared to those of the baseline, beginning with models trained with CNN feature extractors and full sets of training labels.

TABLE 5.4: Pearson correlation coefficients $\rho$ across validation and test set results of DANNs trained with various combinations of cultures. Correlations are high, indicating models learn similar patterns irrespective of target culture.

| Source Culture | Target Culture | Reference Culture | $\rho$ |
|---|---|---|---|
| DE | EN | HU | .991 |
| DE | SR | HU | .994 |
| DE | GR | HU | .995 |
| HU | EN | DE | .889 |
| HU | SR | DE | .939 |
| HU | GR | DE | .979 |

In this analysis, the results of various DANNs trained with different target cultures e. g., English and Serbian but identical source cultures are jointly compared with the baseline. This is justified by the fact that the training yielded highly similar results in terms of variation across the validation and test set cultures, even when values in

absolute terms differed. Pearson correlation scores illustrating this are depicted in table 5.4.

**CNN-GRU based DANNs**

Looking at the CRNN-based DANN results from table 4.11 shows some interesting effects. When using German as the source culture, the top validation performance in-domain remains comparable or slightly below baseline for arousal, while increasing considerably for valence ($0.714 \rightarrow 0.769$. Out-of domain (Hungarian), the validation performance drops, considerably so for valence. One might expect the opposite effect, but this may be explained with the observation that achieving good results on the Hungarian data is difficult especially on valence, at least using the approach chosen herein.

The test set, however, presents a different picture. Here, gains in performance are visible, both for German $0.627 \rightarrow 0.66$ and $0.642 \rightarrow 0.688$ on arousal and valence respectively, and Hungarian $0.516 \rightarrow 0.546$, $0.443 \rightarrow 0.454$. On Chinese, the arousal performance drops ($0.471 \rightarrow 0.422$), while valence improves ($0.46 \rightarrow 0.5$).

Looking at the maximum scores from the tables may not show the complete picture, as the results on Chinese in particular have shown to be prone to strong swings in the experiments. Thus, additionally t-tests for statistical significance on the test sets are performed. All tests use $\alpha = 0.05$. The maximum results will still be referenced as indicators.

For the German-trained DANNs, the improvements on German arousal and valence are significant with $p < 0.05$ ($p = 0.026$ and $p = 0.003$ respectively). The same holds for Hungarian ($p = 0.04$ and $0.037$). On Chinese, there is no significant difference for arousal. The top baseline score of $0.471$ turned out to be unusual, with the mean result being $0.394$. For valence, the improvement is statistically significant with $p < 0.05$ ($p = 0.018$).

Analysing the DANNs trained with Hungarian as the source culture, there are again validation set improvements for German arousal ($0.556 \rightarrow 0.654$) and valence ($0.531 \rightarrow 0.636$), Hungarian arousal improves to $0.512$ from $0.437$, Hungarian valence does not. On the test set, beginning again with German, the top scores show considerable improvement for arousal ($0.493$ to $0.595$), which is also significant, $p = 0.01$. For valence, the increase is from $0.48$ to $0.638$ is significant, with $p = 0.002$. Continuing with the performance on the Hungarian source culture itself, the arousal increases from $0.521$ to $0.563$, a significant change with $p = 0.016$. With Hungarian valence, there is an increase from $0.518$ to $0.542$, which is significant, $p = 0.011$. Finally, on Chinese, the top arousal score jumps from $0.512$ to $0.561$, but this difference is not significant ($p = 0.16$). On valence, the increase from $0.34$ to $0.358$ is significant.

These results indicate that the DANN training is effective at learning representations of emotional content that are domain-invariant, the domains being the cultural backgrounds of the subjects. These representations allow for improved generalisation when the model is confronted with test data from a culture that was not present in the labelled training set, boosting affect recognition results on those cultures. Furthermore, the approach is shown to also improve generalisation within the source domain itself, as both German- and Hungarian-trained DANNs achieved better results in terms of CCC scores on German and Hungarian test data, respectively. A

possible explanation for this is that by removing domain-specific information from the representations, the model is able to focus more on the emotion task itself.

The exception here is the performance on the Chinese data. There was no improvement for arousal, and while valence improved, the results were still below those of the other cultures. This indicates that Chinese emotional displays are challenging for the model, and that further optimisation of the approach is required. A possible explanation for the lower results on the Chinese test set is that is is quite dissimilar from all other data in SEWA, being the only non-European culture, thus the domain shift that needs to be overcome is greater.

**Training DANNs with restricted label data**

Now, the experiments on training DANNs while limiting the emotion labels are analysed, in descending order of available label share. First, the DANNs trained on German are examined.

When the amount of labels is reduced to 75%, the validation set results behave similarly to the DANN trained on the full set, in that German increases (0.729 to 0.754 for arousal, 0.735 to 0.76 for valence, while Hungarian stays the same for arousal ($CCC = 0.53$) and drops on valence (to .396). The test set shows an improvement on German for arousal (0.615 to 0.621 and valence (0.662 to 0.669, both significant at $p < 0.05$. For Hungarian, the improvements on arousal (0.473 to 0.53) and valence (0.401 to 0.43) are both significant, $p = 0.000$ and $p = 0.001$, respectively. For Chinese, there is an increase in arousal score (0.501 to 0.513), and a drop for valence (0.497 to 0.485), but neither of those are significant at $p < 0.05$.

Once the labels are further reduced to 50%, the DANNs exhibit identical behaviour in terms of trends on validation set performance as above. For brevity, the details of the changes are omitted, referring back to the experiments chapter and table 4.15. Instead, the effects on the test results and their significance will be examined here. For German test data, neither arousal or valence results are significantly different from the baseline, $p = 0.13$ and $p = 0.457$, respectively. In case of the Hungarian culture, the improvement on arousal ($CCC = 0.545$) is significant, $p = 0.006$, while there is no such change on valence, $p = 0.10$. For Chinese, neither arousal nor valence show significant improvement ($p = 0.331$, $p = 0.485$).

When the labels were reduced even further, to 25% leading to the results in table 4.17, the test scores on German ($CCC = 0.648$ for arousal, $CCC = 0.667$ for valence) are again both not statistically better than those of the baseline at $p < 0.05$. On the Hungarian dataset, however, the performance increases, to $CCC = 0.505$ and $CCC = 0.499$ respectively, are both significant ($p = 0.001$ for arousal, $p = 0.026$ for valence). Finally, for Chinese, while the performance increases, to 0.471 on arousal and 0.442 on valence, there are again no significant differences at$p < 0.05$.

It is somewhat unexpected that the DANNs trained with full and 75% of the German labels significantly outperform the baseline both German and Hungarian test data, but their counterparts using 50% and 25% of labels do not. A likely explanation is that the supervised baseline training finds powerful models more easily, while the adversarial training process of the DANNs requires further experimentation to achieve optimal results. It is worth highlighting that the DANNs trained on 25% of German labels managed to outperform their corresponding baseline out-of-domain

on Hungarian, so the approach is shown to still be effective even with very small amounts of labelled data.

In order to examine the impact of the source culture choice, the DANNs trained on Hungarian source data with subsets of the labels are discussed. Again, in the interest of brevity and readability, not every effect on the results is repeated from the experiments chapter, and the reader is referred back to the respective tables. Instead, the focus is on the statistical significance of the changes per culture and their interpretation.

Using 75% of the Hungarian labels, see table 4.14, the performance on the test set for German arousal increases significantly with $p < 0.05$ (mean CCCs 0.523 and 0.554 for baseline and DANN respectively). On valence there is no significant difference (average CCCs 0.601 and 0.6.) In-domain, i. e., on the Hungarian test set, the improvements are significant at $p < 0.05$ for both arousal (average $CCC = 0.525$ to $CCC = 0.558$) and valence (average CCC 0.488 to 0.533). The Chinese test set shows no improvement at $p < 0.05$ on arousal (average performance is 0.489 for baseline, 0.489 for DANN), but there is a significant gain for valence (CCC 0.286 to 0.334, $p = 0.001$).

By reducing the labels to 50%, leading to the results in table 4.16, again a significant ($p = 0.001$) gain on German test set is achieved against the baseline ($CCC = 0.518$ to $CCC = 0.58$). For valence, the improvement from 0.583 to 0.597 is not significant at $p < 0.05$. Once more, the DANN outperforms the baseline in-domain, and the improvements are both significant at $p < 0.05$ for arousal (0.512 to 0.547) and valence (0.495 to 0.538). The trend from the models trained on more data also repeats on the Chinese culture, with the improvement on arousal from 0.448 to 0.465 not being significant at $p < 0.05$, while the gain on valence CCC is, from 0.291 to 0.34.

In the final reduction step to 25% of the labels, the DANNs presented in table 4.18 achieve statistically significant improvements at $p < 0.05$ for both German arousal ($\mu = 0.46$ to $\mu = 0.527$) and valence ($\mu = 0.512$ to 0.547). Performance on Hungarian also improves significantly at $p < 0.05$ for both arousal (0.512 to 0.547) and valence (0.48 to 0.499). On Chinese test data, the change in arousal from 0.463 to 0.507 is again not significant at $p < 0.05$, but the improvement on valence is (0.307 to 0.362)

The results of the Hungarian-trained DANNs with reduced label shares are consistent with the previous analysis on the full in that the source domain benefits for both arousal and valence prediction, and that German arousal and Chinese valence both see improvements.

Based on these experiments, the conclusion is that the DANN approach remains effective even as the relation of labelled samples to unlabelled samples in the source culture is greatly decreased. In some cases the DANNs even surpass baselines trained with more labelled samples. These results are promising for the development of affective computing solutions on larger datasets, where a complete annotation would not be feasible, see section 2.5.2.

#### Performance comparison and limitations of the approach

There are several limitations in the approach presented here, which offer room for improvement. First, while the baselines with WAV2VEC2 audio features clearly outperformed the ones using 1D-CNN, the advantage is less prominent with DANNs, suggesting that further hyper-parameter searches will lead to even better models.

Second, the DANNs are not easy to train, requiring specific adjustments of the losses and train loop to converge, see section 3.4.4.

The analysis presented here focused on recurrent encoders, and could be extended to attention-based models, including hybrid cross-modal fusion, as in section 4.1. Furthermore, the approach was tested only on SEWA, as it was the most suitable dataset available. It could be applied to other, larger corpora, or be extended to a combination of datasets, which would introduce additional domain shift.

From the results on the Chinese culture, it is expected that more optimisation will lead to even better performance. Comparing with related work, the models obtained here outperform the Transformer-based TEMMA (H. Chen, Jiang, and Sahli, 2020) on the Chinese culture for both arousal, with 0.561 vs 0.470, and valence (0.5 vs 0.459), respectively. TAADA (H. Chen, Y. Deng, and Jiang, 2021), which is also DANN-based, outperformed the models of this thesis on arousal (0.576), while falling below on valence (0.472). Notably, those other models were trained on the full set of labels from both German and Hungarian cultures simultaneously, while the DANNs in this thesis were trained on either German or Hungarian. Based on these results, it is concluded that the method used in this thesis is competitive and effective with limited access to labels.

## 5.4 Final considerations

Following the discussions of the individual contributions above, this section concludes the chapter by summarising lessons learned and relating the work done in this thesis back to the research questions defined in section 1.2.

Several general observations were made when training models for this thesis: It was found that individual model runs are highly dependent on the initialisations of the weights. Re-running multiple times with different random seeds helped mitigate this. Furthermore, combining the best-performing models into ensembles boosted performance, and is therefore highly recommended. Training models with multi-task learning and applying dynamic balancing between the task losses is also beneficial. Finally, there is considerable potential for hyper-parameter optimisation. Given the large number of possible parameter combinations, running automated searches with early stopping criteria for underperforming trials is recommended e. g., with toolkits like Ray.

Regarding research question RQ–1 i. e., methods for continuous emotion recognition in the wild, it has been established that deep feature extractors learning from raw data are effective. In particular, the light-weight CNN MobileFaceNet performed very well for detecting both arousal and valence. For fusion strategies, both early and hybrid fusion were successfully applied. Both recurrent and Transformer-based sequence models were used, and it was shown that the former are still competitive compared to their more recent attention counterparts. Finally, end-to-end learning was shown to be effective for emotion recognition both to boost performance, and to develop more light-weight models, which is relevant for applications with constrained resources.

On research question RQ–2 i. e., cross-cultural emotion recognition, it was shown that adversarial domain adaptation is an effective tool to help the models learn representations which generalise well on unseen cultures. The method was also found

to boost emotion recognition performance on the source domain. Various combinations of five source and target cultures were experimented with, and the approach was effective on all of them. Thus, adding self-supervised domain adaptation can be considered a promising strategy for cross-cultural affect recognition on diverse in-the-wild data.

Research question RQ–3 i. e., using unlabelled data for emotion recognition was addressed alongside the domain adaptation discussed above. It was found that data containing emotional displays, even when it had no annotations, could still be leveraged to improve recognition performance. By adding a self-supervised culture task, the models were able to learn implicit information that helped them generalise. This approach worked even when the unlabelled samples greatly outnumbered the labelled ones. Furthermore, it was also demonstrated that audio Transformers pre-trained without any emotion labels are effective at predicting affect.

On research question RQ–4 i. e., recognising affect from non-verbal vocalisations, it was found that multi-task learning with models based on fine-tuned audio Transformers is highly effective. The approach works even though the pre-training of the backbone model happened on a quite dissimilar dataset of English speech. It is also robust in challenging, diverse audio conditions. These findings are relevant for developing models that can recognise affect in the absence of speech, which can often be missing in real-life situations.

# Chapter 6

# Outlook and Future Work

In this chapter, an outlook on possible future research and applications is given, based on the findings of the previous chapters and ongoing trends in the field of affective computing and the industry.

For industrial applications the focus is placed on the automotive industry, due to the research for this thesis having been conducted in cooperation with the BMW Group. An outlook towards future generations of vehicles that integrate features for estimating and regulating affect is given in section 6.2.

## 6.1 Challenges and research opportunities

Although affective computing has matured considerably, see e. g., the progress in SER (B. W. Schuller, 2018), major challenges remain in order for the technology to become widely deployed and accepted in real life conditions. In this thesis, aspects of recognising emotions in the wild were investigated, focusing on sequence-based modelling of continuous affective states from audiovisual data. Beyond this many opportunities for further work exist, both on the technical level of data acquisition and emotion recognition, as well as in terms of applications and the ethical implications of wide-scale deployment of such systems. Several of those open questions are listed here, and research paths for addressing them are suggested.

### 6.1.1 Privacy

Privacy is a major concern, given that affective computing is becoming increasingly pervasive, and that it can be used to analyse the mental and physiological states of individuals. This may have far-reaching consequences if the technology is mishandled and data or the inferences derived from it leak.

The data, whether it is generated by cameras, microphones, or any number of wearable sensors, is inherently tied to a person. It cannot be easily anonymised, as e. g., blocking out the face region in videos would also remove salient information that the affective computing solution relies upon, making it, if not inoperable, far less effective.

One way to address privacy concerns is to run all computation locally on the user's hardware, and not to employ persistent storage. For instance, an mood estimator employed in a vehicle assistant could be a compact network running on an ECU, working on video frames that only exist transiently in a buffer. Implementing this kind of system, however, requires extensive optimisation to fulfil computational

resource constraints. Compressing deep neural networks down for edge computing, while maintaining their performance, is a highly topical research field. Popular methods include pruning parameters of the network, as well as quantisation of the remaining weights. Another very interesting strategy are knowledge distillation methods, e. g., learning a smaller network from a larger one in a teacher-student approach (Gou et al., 2021).

In case a transfer of data to a backend server cannot be avoided, one approach to preserve privacy while maintaining the salience of the data is to obscure the identity of persons by replacing them with generated content. For instance, *DeepPrivacy* (Hukkelås, Mester, and Lindseth, 2019) is a GAN-based approach that can anonymise faces. For emotion estimation, the user's face or body could be changed to a different appearance while maintaining the facial expression or pose. The new virtual identity may also be used in privacy-preserving in-cabin monitoring for autonomous vehicles (Gomez-Donoso et al., 2022). Conversely, in situations where a person knows or suspects their affective or cognitive state is being analysed, and that a negative assessment may have undesirable consequences, they may wish to deploy software that can obscure e. g., signs of stress or anxiety. This, however, is only viable in remote interactions where the person has control over the device that records and transmits their data. On the other hand, in online conversations, the other party may be interested in knowing whether the interlocutor is genuine, and deploy software trained to detect such "spoofing".

### 6.1.2 Distributed Learning

There are many reasons to keep data and computation on the end user devices, including latency, security, and privacy, see section 6.1.1. At the same time, it is desirable to have those devices share knowledge, so that they can learn from the data their peers have collected and processed.

*Federated Learning* is a distributed machine learning paradigm that aims to address this challenge. Its basic premise is to have multiple clients perform training locally and communicate with a server, which aggregates their information to learn a joint model with improved performance, and pushes weight updates back to the clients.

While this approach avoids clients having to share their data over the network, adversaries may still be able to exploit the transmitted information, e. g., by model weights leaking aspects of the underlying data. Privacy-preserving federated learning is an active area of research, with a trade-off existing between protection level and convergence performance that needs to be considered in system design (K. Wei et al., 2020). Federated learning has been shown to be effective in SER (Latif et al., 2020), and has great potential for improving the user experience in future connected vehicles, where manufacturers could push out OTA updates based on customer interactions learned from fleets of cars.

### 6.1.3 Beyond Supervised Learning

An ongoing challenge in affective computing and deep learning in general, which has also inspired RQ-2 of this thesis, is overcoming the bottleneck of labels and train networks on large quantities of unlabelled data. Self-supervised training on pretext tasks which let the model implicitly learn about the data has driven the major advancements in NLP, Computer Vision and audio analysis via Transformer-style

architectures in recent years. Commonly these large models are then fine-tuned to a downstream task (Macary et al., 2021), an approach that was also used in this thesis.

Alternatively, self-supervised training of novel multi-modal, cross-modal or cross-corpus models can lead to powerful features. An example is Shukla, Petridis, and Pantic, 2021, who used reconstruction of face based on speech as a cross modal pre-text task. In addition, swapping part of the audio was used as uni-modal pretext task. The combination of both approaches in a multi-task setting yielded improved performance. The approach was tested on several datasets including RECOLA, SEWA and IEMOCAP and shown to outperform existing self-supervised methods as well as fully supervised training. Choosing suitable pre-text tasks for affect remains an open research question.

### 6.1.4 Trust, Fairness and Explainability

One major issue for the widespread adoption of highly automated systems, e. g., autonomous vehicles, is a lack of trust. This may lead to potential users being insecure or apprehensive around the technology, and deciding not to engage with it. If there is sufficient concern, it may cause regulators to restrict or outright ban those systems until their trustworthiness has been sufficiently established.

Anthropomorphising technology is beneficial for user trust, e. g., when an AV possesses features like a name, gender and voice (Waytz, Heafner, and Epley, 2014). However, this approach should try to avoid the "uncanny valley" effect, where an artificial agent that appears almost but not quite human is perceived as unsettling.

Affective computing can help promote trust in automated systems by making their HMI more natural. Being responsive to the users' mood and in turn displaying emotional behaviour can encourage engagement. At the same time, the degree of responsiveness and the way emotions are integrated in the UI needs to be carefully designed and calibrated to the users' preferences to avoid upsetting them, e. g., a speech assistant exacerbating negative feelings through inappropriate comments. Beyond correctly estimating users' emotions, studying how to apply that information in an engaging and context-sensitive manner is a promising research path.

Another important factor that impacts both safety and user satisfaction is fairness of the machine learning models that enable the application. For instance, an emotional or cognitive state estimation component that is deployed globally should give reliable performance results across a wide range of people, not just a specific demographic. While collecting extensive and diverse training data can help mitigate this issue, biases may still creep into the model, e. g., via unaccounted skews in the data distribution or via properties of the algorithm itself. The results can range from annoying, e. g., for entertainment apps not working as expected, to dangerous, e. g., a medical system for mental health analysis misdiagnosing a patient. Therefore, it is important to consider potential harms of biases an automated system and develop methods to analyse and mitigate them. Fairness in machine learning is an active field of research. A recent survey with a taxonomy of approaches and open research directions can be found in Mehrabi et al., 2021.

Furthermore, adoption of systems based on machine learning may be hindered by a lack of interpretability. While deep learning has lead to considerable performance

improvements across many domains, state of the art models are so large that the relation between their inputs and outputs is no longer interpretable by humans (consider e. g., GPT-3, a LLM with 175 billion parameters), effectively turning them into black boxes. There is an inherent trade-off between model transparency and performance, and users may tolerate some opacity if the application works well or the stakes of its decisions are low (Adadi and Berrada, 2018). Nevertheless, interpretability is key to building trust with a technology and managing it effectively. It can be useful when decisions of the model need to be justified, e. g., an autonomous vehicle explaining its actions to avoid passengers experiencing frustration or fear from an unexpected manoeuvre (Wiegand et al., 2020). Interpretability can also contribute to improving the model by uncovering biases or performance issues. Furthermore, besides model errors, vulnerabilities against deliberate attacks can also be studied. The need to develop powerful models while still maintaining the ability to understand their results has given rise to the field of XAI (Barredo Arrieta et al., 2020). XAI encompasses a wide number of approaches, which can be model-specific or model-agnostic. Explanations can be obtained by choosing a model that is intrinsically interpretable by being small or of low complexity, or by adding a post-hoc analysis to a large and complex black box model. For instance, the model's internal representations can be visualised, consider e. g., the attention maps of Transformers. Taxonomies of XAI, along with open challenges in the field, can be found in Adadi and Berrada, 2018; Barredo Arrieta et al., 2020. For affective computing and medical AI, XAI is a highly relevant and topical research direction.

## 6.2 Outlook: The emotionally intelligent vehicle

As described in chapter 2, there is a trend in the automotive industry towards integrating sensors into the vehicle cabin for monitoring the occupants. These sensors enable two types of functions: Safety functions and comfort functions. The former are primarily driver-focused, due to the driver's responsibility for operating the vehicle, and include camera-based driver monitoring for attentiveness and fatigue (J. Wang, Chai, et al., 2022). Safety functions that extend to the passenger seats include e. g., seat belt reminders, intelligent (de)activation of airbag systems, and child or pet presence detection, based on pressure sensors, cameras, or radars. Comfort functions are more geared towards the passengers, allowing them to relax or distract themselves with other activities during the drive. They are controlled mainly by touch-sensitive screens replacing physical buttons (Breitschaft, Pastukhov, and Carbon, 2021), as well as by gesture and speech commands (Murali, Kaboli, and Dahiya, 2022).

In this section, an outlook is given for a future generation of vehicles, in which affective computing based on an in-cabin sensing system enables various use cases.

### 6.2.1 Driving Experience

The vehicle of the future will be capable of semi-autonomous, and later fully autonomous, driving. Affect recognition has important applications here for governing the driving behaviour in autonomous mode.

A car may adjust its driving style, to regulate or promote certain emotions in its passengers. For instance, a sporty and dynamic style may appeal to passengers who are energetic and excited for the drive, while a smooth, cruising style can help

foster relaxation and relieve anxiety (Ling et al., 2021). Personalised and mood-based driving style adaptation is a promising tool for building trust into autonomous vehicles (Sini et al., 2021; Alsaid et al., 2023).

Similarly, an autonomous car might also attempt to gauge the affect of other traffic participants, particularly pedestrians, and interact with them in a reassuring way. Some current-generation vehicles already include signalling capabilities beyond standard turn indicators and horns, in the form of matrix LED displays. A future vehicle could recognise the emotions of pedestrians, infer their expectations based on the context e. g., doubts whether the vehicle will yield, and signal its intent. A survey with recommendations for such interactions can be found in Y. Wang, Hespanhol, and Tomitsch, 2021.

Another way in which affective computing may be used to shape the driving experience is by customising the route based on the emotions of the passengers. The vehicle may suggest a scenic route, and then observe during the drive whether the user enjoys it. This implicit feedback may be used for future recommendations, including for new routes experienced positively by other people (H. Huang et al., 2014).

In order to improve acceptance, the affect-based recommendation could be combined with contextual knowledge, such as whether the user is travelling under a time constraint e. g., going to work, or in a more leisurely manner e. g., on a family vacation. In the former case, a fast and stress-minimising travel route is likely preferred, while for the latter, more emphasis can be placed on aesthetic sights and longer drives. Emotion-based route adjustment is an interesting topic with potential for greatly enhancing user satisfaction, but remains relatively understudied (Karas, D. M. Schuller, and B. W. Schuller, 2024).

### 6.2.2 Infotainment

As driving becomes more autonomous and the interior more connected, new forms of information and entertainment, commonly referred to as *infotainment* are enabled. These include the ability to stream movies, play games or make video calls on large high-resolution screens. When driving is highly automated, some users may also want to use their car as a mobile office space, to productively use the time of their office commute.

Infotainment use cases can be enhanced by affective computing. For instance, the emotions of the user can serve as input to a music recommendation system that selects tracks matching the current mood (Ayata, Yaslan, and Kamasak, 2018), and the same concept can be applied to recommending movies or TV shows. The vehicle could also adjust its interior lighting and UI to match or regulate the user mood.

The growing number and complexity of customer functions motivates the inclusion of a vehicle assistant. This assistant could be personalised and responsive to the user's emotions (Braun, Weber, and Alt, 2021). For instance, it could detect that the user enjoys engaging with it and be more proactive, or it could detect that the user is sad or angry and be more restrained in order to not aggravate them further. Displaying emotions in reactions to queries, e. g., by changing its voice, could make the assistant appear more human-like and encourage interaction. Having an assistant

that users enjoy engaging with, and which knows when and how to present information, will also help with building trust into the autonomous driving capabilities of the vehicle (Lee et al., 2019; Alpers et al., 2020).

### 6.2.3 Health and Wellbeing

Another important application of affective computing in the vehicle is for improving functions that seek to promote customer wellbeing. Here the car could sense the passengers' mood and adjust its interior accordingly, e. g., changing the lighting to be more subdued. For relaxation or refreshment programs, affect recognition in combination with physiological sensing would be helpful. Options for placing sensors in the cabin to unobtrusively measure biosignals were investigated by J. Wang, Warnecke, et al., 2020. Besides vehicle-mounted sensors, linking wearable devices to the car is a promising solution.

Various intervention strategies for passengers who are feeling unwell have been proposed, including guided breathing exercises with voice-controlled or haptic stimuli (Paredes, Zhou, et al., 2018), or calming VR simulations of ocean dives (Paredes, Balters, et al., 2018). Mood sensing would be very useful in these features, both for choosing when to suggest them proactively, and for judging their effectiveness.

# Chapter 7

# Conclusion

Automatic affect recognition holds the potential to elevate human-machine interaction by enabling systems that respond to the user's feelings, leading to a more natural and engaging user experience. However, the technology is not yet fully mature, despite considerable recent advances in deep learning, due to the complex nature of emotions and difficulties associated with recognising them in widely varying situations.

Inspired by this problem, this thesis addressed the topic of estimating continuous emotional states based on audiovisual recordings that are collected in-the-wild i. e., in noisy, real-life conditions. Four main research questions were investigated:

RQ–1, which was concerned with choosing suitable features, multi-modal fusion and temporal modelling schemes for recognising value-continuous emotions in-the-wild.

RQ–2, which based on cultural differences of emotional displays and experiences asked how to teach models to recognise emotions in people from cultures other than the ones they were explicitly trained on.

RQ–3, which being motivated by the high cost of emotion annotations contrasted by the availability of large amounts of data containing natural emotional behaviour, concerned itself with finding ways to enhance model performance by using unlabelled or partially labelled data.

RQ–4, which from a desire to develop models that could interpret more than the commonly analysed cues in the face and speech, focused on detecting affect from less frequently studied non-verbal vocalisations e. g., laughter, groans and cries, with occur naturally in emotional displays.

An overview of theories of emotion was given, including both traditional views of universal categories and modern frameworks of diverse, context-dependent and smoothly varying concepts. An introduction of the young and rapidly evolving field of affective computing was given. Commonly used signal modalities and the features extracted from them were explained, and industrial applications of affective computing were presented. For those applications, the focus was placed on the automotive sector, where various manufacturers have presented empathetic intelligent assistants within concept cars, and features related to emotion recognition are being integrated into production vehicles.

The background introduction is followed by the main part of the thesis, containing methodology, experiments description and discussion. Three core contributions are made here:

C–1 focuses on multi-modal sequence-to-sequence prediction of arousal and valence on Aff-Wild2, a challenging, noisy web-sourced dataset. Various feature extractors and fusion methods, as well as recurrent and attention based sequence models, are compared. The effectiveness of end-to-end learning for boosting recognition performance while reducing the number of parameters is demonstrated. A set of lightweight, computationally efficient networks with CNN-based features is developed, which is promising for applications that need to run in resource-constrained or low-latency settings e. g., on smartphones or inside vehicles. An ensemble of these models achieved 5th place in the third Affective Behaviour in the Wild (ABAW) competition.

In C–2, a dataset of short vocal bursts recorded under realistic conditions, spanning four cultures, is used to analyse non-verbal expression of affect. Eight vocal burst categories, arousal and valence, 10 continuously annotated emotions and 40 culture-specific emotions are predicted by models trained with multi-task learning. An audio Transformer pre-trained on English speech is fine-tuned and used as the backbone of multiple architectures with varying degrees of complexity, including parallel or chained task output heads. The resulting models are analysed in detail, and shown to be highly effective. Combination into an ensemble further boosted performance. The models achieved state of the art results on all four tasks, surpassing the winners of the ACII'22 A-VB competition on all three continuous emotion recognition tracks.

For C–3, dyadic conversations between members of six cultures in the SEWA dataset are used to analyse spontaneous displays of emotion in a cross-cultural context. Arousal and valence are predicted in a sequence-to-sequence model based on deep feature extraction from audiovisual data. Domain adaptation via adversarial training on the cultural background is used to improve performance on both the source and unseen target cultures. The target data is completely unlabelled, and the effectiveness of the approach when restricting the amount of labelled source samples by up to 75% is demonstrated. The results are competitive with the state of the art.

Furthermore, suggestions for future work are given. Here, an emphasis is placed on going beyond the challenge of further improving recognition performance in-the-wild. Instead, as the widespread deployment of emotion analysis software raises ethical questions, proposals are made for research into privacy protections, including emotion-preserving video data anonymisation with generative models and edge computing with privacy-preserving federated learning. Research into fairness and explainability of is also recommended, to find hidden biases in deep models and make their decision processes less inscrutable. Taking these steps will be beneficial for establishing trust and promoting acceptance of affective technology.

Finally, an application for an affect recognition and elicitation system is illustrated in the form of a future generation of intelligent vehicle that responds to the feelings of its occupants. Extrapolating from current research, various use cases are presented. These include emotionally responsive autonomous driving behaviour and route selection, mood-based content recommendation systems, an intelligent personal assistant capable of empathetic conversation, and programs for improving wellbeing that can be offered proactively and derive implicit feedback based on the emotional state.

# References


Adadi, A. and M. Berrada (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6, pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

Ajakan, H., P. Germain, H. Larochelle, F. Laviolette, and M. Marchand (2014). *Domain-Adversarial Neural Networks*. DOI: 10.48550/ARXIV.1412.4446. URL: https://arxiv.org/abs/1412.4446.

Allen, J. (2007). "Photoplethysmography and its application in clinical physiological measurement". In: *Physiological Measurement* 28.3, R1. DOI: 10.1088/0967-3334/28/3/R01. URL: https://dx.doi.org/10.1088/0967-3334/28/3/R01.

Alpers, B. S., K. Cornn, L. E. Feitzinger, U. Khaliq, S. Y. Park, B. Beigi, D. Joseph Hills-Bunnell, T. Hyman, K. Deshpande, R. Yajima, L. Leifer, and L. Aquino Shluzas (2020). "Capturing Passenger Experience in a Ride-Sharing Autonomous Vehicle: The Role of Digital Assistants in User Interface Design". In: *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. AutomotiveUI '20. Virtual Event, DC, USA: Association for Computing Machinery, pp. 83–93. DOI: 10.1145/3409120.3410639. URL: https://doi.org/10.1145/3409120.3410639.

Alsaid, A., J. D. Lee, S. I. Noejovich, and A. Chehade (2023). "The Effect of Vehicle Automation Styles on Drivers' Emotional State". In: *IEEE Transactions on Intelligent Transportation Systems* 24.4, pp. 3963–3973. DOI: 10.1109/TITS.2023.3239880.

Amiriparian, S., M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller (Aug. 2017). "Snore sound classification using image-based deep spectrum features". In: *Proceedings of the 18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017*. Ed. by F. Lacerda. Stockholm, Sweden: ISCA, pp. 3512–3516. DOI: 10.21437/interspeech.2017-434. URL: https://www.isca-speech.org/archive/Interspeech%5C_2017/abstracts/0434.html.

Amiriparian, S., T. Hübner, V. Karas, M. Gerczuk, S. Ottl, and B. W. Schuller (2022). "DeepSpectrumLite: A Power-Efficient Transfer Learning Framework for Embedded Speech and Audio Processing From Decentralized Data". In: *Frontiers in Artificial Intelligence* 5. DOI: 10.3389/frai.2022.856232. URL: https://www.frontiersin.org/article/10.3389/frai.2022.856232.

Athanasiadis, C., E. Hortal, and S. Asteriadis (July 2019). "Audio-visual domain adaptation using conditional semi-supervised Generative Adversarial Networks". In: *Neurocomputing* 397, pp. 331–344. DOI: https://doi.org/10.1016/j.neucom.2019.09.106.

Audi (2017). *Audi Elaine*. https://www.audi.com/de/innovation/concept-cars/audi-elaine.html. Retrieved 05 May 2022.

Ayata, D., Y. Yaslan, and M. E. Kamasak (2018). "Emotion Based Music Recommendation System Using Wearable Physiological Sensors". In: *IEEE Transactions on Consumer Electronics* 64.2, pp. 196–203. DOI: 10.1109/TCE.2018.2844736.

Baevski, A., Y. Zhou, A. Mohamed, and M. Auli (2020). "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 12449–12460. URL: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Baird, A., P. Tzirakis, J. A. Brooks, C. B. Gregory, B. Schuller, A. Batliner, D. Keltner, and A. Cowen (2022). "The ACII 2022 Affective Vocal Bursts Workshop & Competition". In: *2022 10th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–5. DOI: 10.1109/ACIIW57231.2022.10086002.

Baird, A., P. Tzirakis, G. Gidel, M. Jiralerspong, E. B. Muller, K. Mathewson, B. Schuller, E. Cambria, D. Keltner, and A. Cowen (2022). *The ICML 2022 Expressive Vocalizations Workshop and Competition: Recognizing, Generating, and Personalizing Vocal Bursts*. arXiv: 2205.01780 [eess.AS].

Baltrusaitis, T., A. Zadeh, Y. C. Lim, and L. Morency (2018). "OpenFace 2.0: Facial Behavior Analysis Toolkit". In: *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66. DOI: 10.1109/FG.2018.00019.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera (2020). "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". In: *Information Fusion* 58, pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012. URL: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

Barrett, L. F. (Oct. 2016). "The theory of constructed emotion: an active inference account of interoception and categorization". In: *Social Cognitive and Affective Neuroscience* 12.1, pp. 1–23. DOI: 10.1093/scan/nsw154. URL: https://doi.org/10.1093/scan/nsw154.

Barrett, L. F., R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak (2019). "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements". In: *Psychological Science in the Public Interest* 20.1, pp. 1–68. DOI: 10.1177/1529100619832930. URL: https://doi.org/10.1177/1529100619832930.

Batliner, A., S. Hantke, and B. W. Schuller (2020). "Ethics and Good Practice in Computational Paralinguistics". In: *IEEE Transactions on Affective Computing* 13.3, pp. 1236–1253. DOI: 10.1109/TAFFC.2020.3021015.

Bengio, Y., A. Courville, and P. Vincent (Aug. 2013). "Representation Learning: A Review and New Perspectives". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8, pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50. URL: http://dx.doi.org/10.1109/TPAMI.2013.50.

BMW (June 2021). *The first ever BMW iX*. https://www.press.bmwgroup.com/global/article/attachment/T0333569EN/483906. Retrieved May 03, 2022.

Braun, M., F. Weber, and F. Alt (Sept. 2021). "Affective Automotive User Interfaces – Reviewing the State of Driver Affect Research and Emotion Regulation in the Car". In: *ACM Comput. Surv.* 54.7. DOI: 10.1145/3460938. URL: https://doi.org/10.1145/3460938.

Breitschaft, S. J., A. Pastukhov, and C. C. Carbon (2021). "Where's My Button? Evaluating the User Experience of Surface Haptics in Featureless Automotive User Interfaces." In: *IEEE Transactions on Haptics*, pp. 1–1. DOI: 10.1109/TOH.2021.3131058.

Busso, C., M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan (Nov. 2008). "IEMOCAP: interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42.4, p. 335. DOI: 10.1007/s10579-008-9076-6. URL: https://doi.org/10.1007/s10579-008-9076-6.

Canales, L., W. Daelemans, E. Boldrini, and P. Martínez-Barco (2022). "EmoLabel: Semi-Automatic Methodology for Emotion Annotation of Social Media Text". In: *IEEE Transactions on Affective Computing* 13.2, pp. 579–591. DOI: 10.1109/TAFFC.2019.2927564.

Cao, Q., L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman (2018). "VGGFace2: A Dataset for Recognising Faces across Pose and Age". In: *2018 13th IEEE International Conference on Automatic Face  Gesture Recognition (FG 2018)*, pp. 67–74. DOI: 10.1109/FG.2018.00020.

Chatterjee, R., S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri (Feb. 2021). "Real-Time Speech Emotion Analysis for Smart Home Assistants". In: *IEEE Transactions on Consumer Electronics* 67.1, pp. 68–76. DOI: 10.1109/TCE.2021.3056421.

Chen, H., Y. Deng, S. Cheng, Y. Wang, D. Jiang, and H. Sahli (2019). "Efficient Spatial Temporal Convolutional Features for Audiovisual Continuous Affect Recognition". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. AVEC '19. Nice, France: Association for Computing Machinery, pp. 19–26. DOI: 10.1145/3347320.3357690. URL: https://doi.org/10.1145/3347320.3357690.

Chen, H., Y. Deng, and D. Jiang (2021). "Temporal Attentive Adversarial Domain Adaption for Cross Cultural Affect Recognition". In: *Companion Publication of the 2021 International Conference on Multimodal Interaction*. ICMI '21 Companion. Montreal, QC, Canada: Association for Computing Machinery, pp. 97–103. DOI: 10.1145/3461615.3491110. URL: https://doi.org/10.1145/3461615.3491110.

Chen, H., D. Jiang, and H. Sahli (2020). "Transformer Encoder with Multi-modal Multi-head Attention for Continuous Affect Recognition". In: *IEEE Transactions on Multimedia*, pp. 1–1. DOI: 10.1109/TMM.2020.3037496.

Chen, S., Y. Liu, X. Gao, and Z. Han (2018). "MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices". In: *Biometric Recognition*. Ed. by J. Zhou, Y. Wang, Z. Sun, Z. Jia, J. Feng, S. Shan, K. Ubul, and Z. Guo. Cham: Springer International Publishing, pp. 428–438.

Cowen, A. S. and D. Keltner (2017). "Self-report captures 27 distinct categories of emotion bridged by continuous gradients". In: *Proceedings of the National Academy of Sciences* 114.38, E7900–E7909. DOI: 10.1073/pnas.1702247114. eprint: https://www.pnas.org/content/114/38/E7900.full.pdf. URL: https://www.pnas.org/content/114/38/E7900.

Cowie, R., M. Sawey, C. Doherty, J. Jaimovich, C. Fyans, and P. Stapleton (2013). "Gtrace: General Trace Program Compatible with EmotionML". In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 709–710. DOI: 10.1109/ACII.2013.126.

Deng, D., Z. Chen, and B. E. Shi (2020). *Multitask Emotion Recognition with Incomplete Labels*. arXiv: 2002.03557 [cs.CV].

Deng, D., L. Wu, and B. E. Shi (Oct. 2021). "Iterative Distillation for Better Uncertainty Estimates in Multitask Emotion Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3557–3566.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings*

*of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Vol. abs/1810.04805. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *CoRR*. URL: arXiv%20preprint%20arXiv:2010.11929.

Ekman, P. and D. Cordaro (Sept. 2011). "What is Meant by Calling Emotions Basic". In: *Emotion Review* 3.4, pp. 364–370. DOI: 10.1177/1754073911410740. URL: https://doi.org/10.1177/1754073911410740.

Ekman, P. and W. V. Friesen (Feb. 1971). "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17 (2), pp. 124–9.

Ekman, P. and W. V. Friesen (1978). "Facial action coding system: A technique for the measurement of facial movement." In: *Environmental Psychology & Nonverbal Behavior*.

Ekman, P., W. V. Friesen, and J. C. Hager (2002). "Facial Action Coding System: Facial action coding system: the manual: on CD-ROM". In: *Network Research Information, Salt Lake City, UT*.

Euro NCAP (2017). *Roadmap 2025*. https://cdn.euroncap.com/media/30700/euroncap-roadmap-2025-v4.pdf.

Eyben, F., K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong (Apr. 2016). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7.2, pp. 190–202. DOI: 10.1109/TAFFC.2015.2457417.

Eyben, F., M. Wöllmer, and B. Schuller (Mar. 2012). "A Multitask Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech". In: *ACM Trans. Interact. Intell. Syst.* 2.1. DOI: 10.1145/2133366.2133372.

Freitag, M., S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller (2018). "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks". In: *Journal of Machine Learning Research* 18.173, pp. 1–5. URL: http://jmlr.org/papers/v18/17-406.html.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky (2016). "Domain-adversarial training of neural networks". In: *The journal of machine learning research* 17.1, pp. 2096–2030.

Ghaleb, E., M. Popa, E. Hortal, and S. Asteriadis (Sept. 2017). "Multimodal fusion based on information gain for emotion recognition in the wild". In: *2017 Intelligent Systems Conference (IntelliSys)*, pp. 814–823. DOI: 10.1109/IntelliSys.2017.8324224.

Girard, J. M. (2014). "CARMA: Software for continuous affect rating and media annotation." In: *Journal of open research software* 2 (1).

Girard, J. M. and A. G. C. Wright (2018). "DARMA: Software for dual axis rating and media annotation". In: *Behavior Research Methods* 50.3, pp. 902–909. DOI: 10.3758/s13428-017-0915-5. URL: https://doi.org/10.3758/s13428-017-0915-5.

Gomez-Donoso, F., A. Mishra, J. Cha, and S. Kim (Jan. 2022). "Privacy-Preserved In-Cabin Monitoring System for Autonomous Vehicles". In: *Computational Intelligence and Neuroscience* 2022. DOI: 10.1155/2022/5389359. URL: https://doi.org/10.1155/2022/5389359.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative adversarial nets". In: *Advances in neural information processing systems*, pp. 2672–2680.

Gou, J., B. Yu, S. J. Maybank, and D. Tao (2021). "Knowledge Distillation: A Survey". In: *International Journal of Computer Vision* 129.6, pp. 1789–1819. DOI: 10.1007/s11263-021-01453-z. URL: https://doi.org/10.1007/s11263-021-01453-z.

Guo, Z., Y. Pan, G. Zhao, S. Cao, and J. Zhang (2018). "Detection of Driver Vigilance Level Using EEG Signals and Driving Contexts". In: *IEEE Transactions on Reliability* 67.1, pp. 370–380. DOI: 10.1109/TR.2017.2778754.

Hamieh, S., V. Heiries, H. Al Osman, and C. Godin (2021). "Multi-Modal Fusion for Continuous Emotion Recognition by Using Auto-Encoders". In: *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*. MuSe '21. Virtual Event, China: Association for Computing Machinery, pp. 21–27. DOI: 10.1145/3475957.3484455. URL: https://doi.org/10.1145/3475957.3484455.

Han, J., Z. Zhang, Z. Ren, and B. W. Schuller (2021). "EmoBed: Strengthening Monomodal Emotion Recognition via Training with Crossmodal Emotion Embeddings". In: *IEEE Transactions on Affective Computing* 12.3, pp. 553–564. DOI: 10.1109/TAFFC.2019.2928297.

Hantke, S., A. Abstreiter, N. Cummins, and B. W. Schuller (2018). "Trustability-Based Dynamic Active Learning for Crowdsourced Labelling of Emotional Audio Data". In: *IEEE Access* 6, pp. 42142–42155. DOI: 10.1109/ACCESS.2018.2858931.

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Hershey, S., S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson (Mar. 2017). "CNN architectures for large-scale audio classification". In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135. DOI: 10.1109/ICASSP.2017.7952132.

Hinrichs, H., M. Scholz, A. K. Baum, J. W. Y. Kam, R. T. Knight, and H.-J. Heinze (2020). "Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG system for clinical applications". In: *Scientific Reports* 10.1, p. 5218. DOI: 10.1038/s41598-020-62154-0. URL: https://doi.org/10.1038/s41598-020-62154-0.

Hu, J., L. Shen, S. Albanie, G. Sun, and E. Wu (Aug. 2020). "Squeeze-and-Excitation Networks". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 42.8, pp. 2011–2023. DOI: 10.1109/TPAMI.2019.2913372. URL: https://doi.org/10.1109/TPAMI.2019.2913372.

Huang, H., S. Klettner, M. Schmidt, G. Gartner, S. Leitinger, A. Wagner, and R. Steinmann (Dec. 2014). "AffectRoute - considering people's affective responses to environments for enhancing route-planning services". In: *International Journal of Geographical Information Science* 28.12, pp. 2456–2473. DOI: 10.1080/13658816.2014.931585. URL: https://doi.org/10.1080/13658816.2014.931585.

Huang, J., J. Tao, B. Liu, Z. Lian, and M. Niu (May 2020). "Multimodal Transformer Fusion for Continuous Emotion Recognition". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3507–3511. DOI: 10.1109/ICASSP40776.2020.9053762.

Hukkelås, H., R. Mester, and F. Lindseth (2019). "Deepprivacy: A generative adversarial network for face anonymization". In: Springer, pp. 565–578.

Hume AI (Oct. 2022). *The ACII Affective Vocal Bursts (A-VB) Workshop Competition*. https://www.competitions.hume.ai/avb2022. accessed 14.08.2024.

Izard, C. E. (2011). "Forms and functions of emotions: Matters of emotion-cognition interactions". In: *Emotion review* 3.4, pp. 371–378.

Jeon, M. (2015). "Towards affect-integrated driving behaviour research". In: *Theoretical Issues in Ergonomics Science* 16.6, pp. 553–585. DOI: 10.1080/1463922X.2015.1067934. eprint: https://doi.org/10.1080/1463922X.2015.1067934. URL: https://doi.org/10.1080/1463922X.2015.1067934.

Karas, V., D. M. Schuller, and B. W. Schuller (2024). "Audiovisual Affect Recognition for Autonomous Vehicles: Applications and Future Agendas". In: *IEEE Transactions on Intelligent Transportation Systems* 25.6, pp. 4918–4932. DOI: 10.1109/TITS.2023.3333749.

Karas, V., M. K. Tellamekala, A. Mallol-Ragolta, M. Valstar, and B. W. Schuller (2022). "Time-Continuous Audiovisual Fusion with Recurrence vs Attention for In-The-Wild Affect Recognition". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2381–2390. DOI: 10.1109/CVPRW56347.2022.00266.

Karas, V., A. Triantafyllopoulos, M. Song, and B. Schuller (July 2022). "Self-Supervised Attention Networks and Uncertainty Loss Weighting for Multi-Task Emotion Recognition on Vocal Bursts". In: *Proceedings of the ACII Affective Vocal Bursts Workshop Competition 2022 (A-VB)*, pp. 1–5. DOI: https://doi.org/10.48550/arXiv.2210.15754.

Karnewar, A. and O. Wang (2020). "Msg-gan: Multi-scale gradients for generative adversarial networks". In: pp. 7799–7808.

KIA (2019). *Real Time Emotion Adaptive Driving*. https://www.kia.com/in/discover-kia/innovation/future-tech.html. Retrieved 05 May 2022.

Kim, H., S. Kim, H. Kim, Y. Ji, and C.-H. Im (2022). "Modulation of Driver's Emotional States by Manipulating In-Vehicle Environment: Validation With Biosignals Recorded in An Actual Car Environment". In: *IEEE Transactions on Affective Computing* 13.4, pp. 1783–1792. DOI: 10.1109/TAFFC.2022.3206222.

Koesdwiady, A., R. Soua, F. Karray, and M. S. Kamel (2017). "Recent Trends in Driver Safety Monitoring Systems: State of the Art and Challenges". In: *IEEE Transactions on Vehicular Technology* 66.6, pp. 4550–4563. DOI: 10.1109/TVT.2016.2631604.

Kollias, D. (2022). "Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges". In: pp. 2328–2336.

– (2023). "ABAW: Learning from Synthetic Data & Multi-task Learning Challenges". In: *Computer Vision - ECCV 2022 Workshops*. Ed. by L. Karlinsky, T. Michaeli, and K. Nishino. Cham: Springer Nature Switzerland, pp. 157–172.

Kollias, D., A. Schulc, E. Hajiyev, and S. Zafeiriou (2020). "Analysing Affective Behavior in the First ABAW 2020 Competition". In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 637–643. DOI: 10.1109/FG47880.2020.00126.

Kollias, D., P. Tzirakis, A. Baird, A. Cowen, and S. Zafeiriou (June 2023). "ABAW: Valence-Arousal Estimation, Expression Recognition, Action Unit Detection & Emotional Reaction Intensity Estimation Challenges". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5888–5897.

Kollias, D., P. Tzirakis, A. Cowen, S. Zafeiriou, I. Kotsia, A. Baird, C. Gagne, C. Shao, and G. Hu (June 2024). "The 6th Affective Behavior Analysis In-the-wild (ABAW) Competition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4587–4598.

Kollias, D., P. Tzirakis, M. . Nicolaou, A. Papaioannou, G. Zhao, B. W. Schuller, I. Kotsia, and S. Zafeiriou (June 2019). "Deep Affect Prediction in-the-Wild: Aff-Wild

Database and Challenge, Deep Architectures, and Beyond". In: *International Journal of Computer Vision* 127.6-7, 907–929. DOI: 10.1007/s11263-019-01158-4.

Kollias, D. and S. Zafeiriou (2019). "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace". In: *CoRR* abs/1910.04855. arXiv: 1910.04855. URL: http://arxiv.org/abs/1910.04855.

– (2021). "Analysing affective behavior in the second abaw2 competition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3652–3660.

Kossaifi, J., R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, et al. (2019). "SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild". In: *arXiv preprint arXiv:1901.02839*.

Kuhnke, F., L. Rumberg, and J. Ostermann (2020). "Two-Stream Aural-Visual Affect Analysis in the Wild". In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 600–605. DOI: 10.1109/FG47880.2020.00056.

Lanatà, A., G. Valenza, A. Greco, C. Gentili, R. Bartolozzi, F. Bucchi, F. Frendo, and E. P. Scilingo (2015). "How the Autonomic Nervous System and Driving Style Change With Incremental Stressing Conditions During Simulated Driving". In: *IEEE Transactions on Intelligent Transportation Systems* 16.3, pp. 1505–1517. DOI: 10.1109/TITS.2014.2365681.

Latif, S., S. Khalifa, R. Rana, and R. Jurdak (2020). "Poster Abstract: Federated Learning for Speech Emotion Recognition Applications". In: *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, pp. 341–342. DOI: 10.1109/IPSN48710.2020.00-16.

Lee, S. C., H. Sanghavi, S. Ko, and M. Jeon (2019). "Autonomous Driving with an Agent: Speech Style and Embodiment". In: *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*. AutomotiveUI '19. Utrecht, Netherlands: Association for Computing Machinery, pp. 209–214. DOI: 10.1145/3349263.3351515. URL: https://doi.org/10.1145/3349263.3351515.

Levenson, R. W. (Sept. 2011). "Basic Emotion Questions". In: *Emotion Review* 3.4, pp. 379–386. DOI: 10.1177/1754073911410743. URL: https://doi.org/10.1177/1754073911410743.

Liang, J., S. Chen, J. Zhao, Q. Jin, H. Liu, and L. Lu (2019). "Cross-culture Multimodal Emotion Recognition with Adversarial Learning". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4000–4004. DOI: 10.1109/ICASSP.2019.8683725.

Lin, L. I.-K. (1989). "A Concordance Correlation Coefficient to Evaluate Reproducibility". In: *Biometrics* 45.1, pp. 255–268. DOI: 10.2307/2532051. URL: http://www.jstor.org/stable/2532051.

Ling, J., J. Li, K. Tei, and S. Honiden (2021). "Towards Personalized Autonomous Driving: An Emotion Preference Style Adaptation Framework". In: *Proceedings of the 2021 IEEE International Conference on Agents (ICA)*, pp. 47–52. DOI: 10.1109/ICA54137.2021.00015.

Liu, S., E. Johns, and A. J. Davison (June 2019). "End-To-End Multi-Task Learning With Attention". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lv, C., Y. Li, Y. Xing, C. Huang, D. Cao, Y. Zhao, and Y. Liu (Apr. 2021). "Human-Machine Collaboration for Automated Driving Using an Intelligent Two-Phase Haptic Interface". In: *Adv. Intell. Syst.* 3.4, p. 2000229. DOI: 10.1002/aisy.202000229. URL: https://doi.org/10.1002/aisy.202000229.

Macary, M., M. Tahon, Y. Estève, and A. Rousseau (2021). "On the Use of Self-Supervised Pre-Trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition". In: *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 373–380. DOI: 10.1109/SLT48900.2021.9383456.

Martin, O., I. Kotsia, B. Macq, and I. Pitas (2006). "The eNTERFACE'05 audio-visual emotion database". In: *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, pp. 8–8.

McDuff, D. (Jan. 2023). "Camera Measurement of Physiological Vital Signs". In: *ACM Comput. Surv.* 55.9. DOI: 10.1145/3558518. URL: https://doi.org/10.1145/3558518.

McKeown, G., M. Valstar, R. Cowie, M. Pantic, and M. Schroder (Jan. 2012). "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent". In: *IEEE Transactions on Affective Computing* 3.1, pp. 5–17. DOI: 10.1109/T-AFFC.2011.20.

McStay, A. and L. Urquhart (Sept. 2022). "In cars (are we really safest of all?): interior sensing and emotional opacity". In: *International Review of Law, Computers & Technology* 36.3, pp. 470–493. DOI: 10.1080/13600869.2021.2009181. URL: https://doi.org/10.1080/13600869.2021.2009181.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (July 2021). "A Survey on Bias and Fairness in Machine Learning". In: *ACM Comput. Surv.* 54.6. DOI: 10.1145/3457607. URL: https://doi.org/10.1145/3457607.

Meng, L., Y. Liu, X. Liu, Z. Huang, Y. Cheng, M. Wang, C. Liu, and Q. Jin (2022). *Multi-modal Emotion Estimation for in-the-wild Videos*. arXiv: 2203.13032 [cs.CV].

Mercedes-Benz (2022a). *ENERGIZING*. https://www.mercedes-benz.de/passengercars/mercedes-benz-cars/energizing/energizing-comfort.module.html. Retrieved 09 May 2022.

– (2022b). *MBUX Interior Assistant*. https://www.mercedes-benz.de/passengercars/mercedes-benz-cars/models/eqs/saloon-v297/specifications/digital-interior.module.html. Retrieved 05 May 2022.

Metallinou, A. and S. Narayanan (2013). "Annotation and processing of continuous emotional attributes: Challenges and opportunities". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. DOI: 10.1109/FG.2013.6553804.

Mollahosseini, A., B. Hasani, and M. H. Mahoor (2019). "AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild". In: *IEEE Transactions on Affective Computing* 10.1. cited By 4, pp. 18–31. DOI: 10.1109/TAFFC.2017.2740923.

Murali, P. K., M. Kaboli, and R. Dahiya (Feb. 2022). "Intelligent In-Vehicle Interaction Technologies". In: *Adv. Intell. Syst.* 4.2, p. 2100122. DOI: 10.1002/aisy.202100122. URL: https://doi.org/10.1002/aisy.202100122.

Nio (Feb. 2020). *NOMI – World's first in-vehicle artificial intelligence*. https://www.nio.com/blog/nomi-worlds-first-vehicle-artificial-intelligence. Retrieved 13 May 2022.

Ortony, A. (July 2021). "Are All "Basic Emotions" Emotions? A Problem for the (Basic) Emotions Construct". In: *Perspectives on Psychological Science* 17.1, pp. 41–61. DOI: 10.1177/1745691620985415.

Pan, S. J. and Q. Yang (2010). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10, pp. 1345–1359.

Pandit, V. and B. Schuller (2020). *The Many-to-Many Mapping Between the Concordance Correlation Coefficient and the Mean Square Error*. arXiv: 1902.05180 [cs.LG].

Panlima, A. and K. Sukvichai (2023). "Investigation on MLP, CNNs and Vision Transformer models performance for Extracting a Human Emotions via Facial Expressions". In: *2023 Third International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP)*, pp. 127–130. DOI: 10.1109/ICA-SYMP56348.2023.10044742.

Paredes, P. E., S. Balters, K. Qian, E. L. Murnane, F. Ordóñez, W. Ju, and J. A. Landay (Dec. 2018). "Driving with the Fishes: Towards Calming and Mindful Virtual Reality Experiences for the Car". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.4. DOI: 10.1145/3287062. URL: https://doi.org/10.1145/3287062.

Paredes, P. E., F. Ordonez, W. Ju, and J. A. Landay (2018). "Fast & Furious: Detecting Stress with a Car Steering Wheel". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, pp. 1–12. DOI: 10.1145/3173574.3174239. URL: https://doi.org/10.1145/3173574.3174239.

Paredes, P. E., Y. Zhou, N. A.-H. Hamdan, S. Balters, E. Murnane, W. Ju, and J. A. Landay (Mar. 2018). "Just Breathe: In-Car Interventions for Guided Slow Breathing". In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2.1. DOI: 10.1145/3191760. URL: https://doi.org/10.1145/3191760.

Park, C., S. Shahrdar, and M. Nojoumian (2018). "EEG-Based Classification of Emotional State Using an Autonomous Vehicle Simulator". In: *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 297–300. DOI: 10.1109/SAM.2018.8448945.

Park, D. S., W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le (Sept. 2019). "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition". In: *Proceedings of 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*. ISCA. Graz, Austria: ISCA, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680.

Picard, R. W. (2010). "Affective computing: from laughter to IEEE". In: *IEEE Transactions on Affective Computing* 1.1, pp. 11–17. DOI: 10.1109/T-AFFC.2010.10.

Plutchik, R. and H. Kellerman (1980). "Chapter 1 - A General Psychoevolutionary Theory of Emotion". In: *Theories of Emotion*. Academic Press, pp. 3–33. DOI: 10.1016/B978-0-12-558701-3.50007-7. URL: https://www.sciencedirect.com/science/article/pii/B9780125587013500077.

Poria, S., E. Cambria, R. Bajpai, and A. Hussain (2017). "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information Fusion* 37, pp. 98–125.

Poria, S., D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea (2018). "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *CoRR* abs/1810.02508.

Raamkumar, A. S. and Y. Yang (2023). "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities". In: *IEEE Transactions on Affective Computing* 14.4, pp. 2722–2739. DOI: 10.1109/TAFFC.2022.3226693.

Read, J., B. Pfahringer, G. Holmes, and E. Frank (May 2021). "Classifier Chains: A Review and Perspectives". In: *J. Artif. Int. Res.* 70, pp. 683–718. DOI: 10.1613/jair.1.12376. URL: https://doi.org/10.1613/jair.1.12376.

Ringeval, F., B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, et al. (2018). "AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition". In: *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, pp. 3–13.

Ringeval, F., B. Schuller, M. Valstar, N. Cummins, R. Cowie, and M. Pantic (Oct. 2019). "AVEC'19: Audio/Visual Emotion Challenge and Workshop". In: pp. 2718–2719. DOI: 10.1145/3343031.3350550.

Ringeval, F., B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, S. Song, S. Liu, Z. Zhao, A. Mallol-Ragolta, Z. Ren, M. Soleymani, and M. Pantic (2019). "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. AVEC '19. Nice, France: Association for Computing Machinery, pp. 3–12. DOI: 10.1145/3347320.3357688. URL: https://doi.org/10.1145/3347320.3357688.

Ringeval, F., A. Sonderegger, J. Sauer, and D. Lalanne (2013). "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. DOI: 10.1109/FG.2013.6553805.

Rizos, G., A. Baird, M. Elliott, and B. Schuller (May 2020). "Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition". In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3502–3506. DOI: 10.1109/ICASSP40776.2020.9054579.

Roka, S. and D. B. Rawat (2023). "Fine Tuning Vision Transformer Model for Facial Emotion Recognition: Performance Analysis for Human-Machine Teaming". In: *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 134–139. DOI: 10.1109/IRI58017.2023.00030.

Rovetta, S., Z. Mnasri, F. Masulli, and A. Cabri (2020). "Emotion Recognition from Speech: An Unsupervised Learning Approach". In: *International Journal of Computational Intelligence Systems*. DOI: 10.2991/ijcis.d.201019.002. URL: https://doi.org/10.2991/ijcis.d.201019.002.

Russell, J. A. and A. Mehrabian (1977). "Evidence for a three-factor theory of emotions". In: *Journal of Research in Personality* 11.3, pp. 273–294. DOI: https://doi.org/10.1016/0092-6566(77)90037-X. URL: https://www.sciencedirect.com/science/article/pii/009265667790037X.

Russell, J. A. (1980). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161.

Russell, J. A. and L. F. Barrett (May 1999). "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." In: *Journal of personality and social psychology* 76 (5), pp. 805–19. DOI: 10.1037//0022-3514.76.5.805.

Schmitt, M., N. Cummins, and B. Schuller (Sept. 2019). "Continuous emotion recognition in speech: do we need recurrence?" In: *Proceedings of 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019*. Graz, Austria, pp. 2808–2812. DOI: 10.21437/Interspeech.2019-2710.

Schroff, F., D. Kalenichenko, and J. Philbin (2015). "FaceNet: A unified embedding for face recognition and clustering". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

Schuller, B. W. (Apr. 2018). "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends". In: *Communications of the ACM* 61, pp. 90–99. DOI: 10.1145/3129340.

Schuller, B. W., A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, and M. Gerczuk (2021). "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech,

Escalation & Primates". In: *Proceedings of the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. Vol. 6. Brno, Czech Republic: ISCA, pp. 4291–4295.

Schuller, B. W., R. Picard, E. André, J. Gratch, and J. Tao (2021). "Intelligent Signal Processing for Affective Computing [From the Guest Editors]". In: *IEEE Signal Processing Magazine* 38.6, pp. 9–11. DOI: 10.1109/MSP.2021.3096415.

Sharma, K., C. Castellini, F. Stulp, and E. L. van den Broek (2020). "Continuous, Real-Time Emotion Annotation: A Novel Joystick-Based Analysis Framework". In: *IEEE Transactions on Affective Computing* 11.1, pp. 78–84. DOI: 10.1109/TAFFC.2017.2772882.

Shukla, A., S. Petridis, and M. Pantic (2021). "Does Visual Self-Supervision Improve Learning of Speech Representations for Emotion Recognition". In: *IEEE Transactions on Affective Computing*, pp. 1–1. DOI: 10.1109/TAFFC.2021.3062406.

Sini, J., A. C. Marceddu, M. Violante, and R. Dessì (2021). "Passengers' Emotions Recognition to Improve Social Acceptance of Autonomous Driving Vehicles". In: *Progresses in Artificial Intelligence and Neural Systems*. Ed. by A. Esposito, M. Faundez-Zanuy, F. C. Morabito, and E. Pasero. Singapore: Springer Singapore, pp. 25–32. DOI: 10.1007/978-981-15-5093-5_3. URL: https://doi.org/10.1007/978-981-15-5093-5_3.

Song, M., Z. Yang, A. Triantafyllopoulos, X. Jing, V. Karas, X. Jiangjian, Z. Zhang, Y. Yoshiharu, and B. W. Schuller (2022). *Dynamic Restrained Uncertainty Weighting Loss for Multitask Learning of Vocal Expression*. arXiv: 2206.11049 [cs.SD].

Srinivasan, R. and A. M. Martinez (2018). "Cross-Cultural and Cultural-Specific Production and Perception of Facial Expressions of Emotion in the Wild". In: *IEEE Transactions on Affective Computing*. cited By 0; Article in Press. DOI: 10.1109/TAFFC.2018.2887267.

Stappen, L., A. Baird, L. Schumann, and B. W. Schuller (2021). "The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements". In: *IEEE Transactions on Affective Computing*, pp. 1–1. DOI: 10.1109/TAFFC.2021.3097002.

Stappen, L., E.-M. Meßner, E. Cambria, G. Zhao, and B. W. Schuller (2021). "MuSe 2021 Challenge: Multimodal Emotion, Sentiment, Physiological-Emotion, and Stress Detection". In: *Proceedings of the 29th ACM International Conference on Multimedia*. MM '21. Virtual Event, China: Association for Computing Machinery, pp. 5706–5707. DOI: 10.1145/3474085.3478582. URL: https://doi.org/10.1145/3474085.3478582.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. A. Alemi (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the Thirty-first AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, pp. 4278–4284.

Tan, Z., N. Dai, Y. Su, R. Zhang, Y. Li, D. Wu, and S. Li (2022). "Human-Machine Interaction in Intelligent and Connected Vehicles: A Review of Status Quo, Issues, and Opportunities". In: *IEEE Transactions on Intelligent Transportation Systems* 23.9, pp. 13954–13975. DOI: 10.1109/TITS.2021.3127217.

Toisoul, A., J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic (2021). "Estimation of continuous valence and arousal levels from faces in naturalistic conditions". In: *Nature Machine Intelligence* 3.1, pp. 42–50. DOI: 10.1038/s42256-020-00280-0. URL: https://doi.org/10.1038/s42256-020-00280-0.

Tomkins, S. (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company.

Toyota (Oct. 2019). *LQ*. https://global.toyota/en/newsroom/corporate/30063126.html. Retrieved 10 May, 2022.

Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou (2016). "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204. DOI: 10.1109/ICASSP.2016.7472669.

Tsai, Y.-H. H., S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov (July 2019). "Multimodal Transformer for Unaligned Multimodal Language Sequences". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6558–6569. DOI: 10.18653/v1/P19-1656. URL: https://www.aclweb.org/anthology/P19-1656.

Tzirakis, P., J. Zhang, and B. W. Schuller (2018). "End-to-End Speech Emotion Recognition Using Deep Neural Networks". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5093. DOI: 10.1109/ICASSP.2018.8462677.

Tzirakis, P. (2020). "End2You: Multimodal Profiling by End-to-End Learning and Applications". In: *Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*. MuSe'20. Seattle, WA, USA: Association for Computing Machinery, p. 9. DOI: 10.1145/3423327.3423513. URL: https://doi.org/10.1145/3423327.3423513.

Tzirakis, P., G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou (Dec. 2017). "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks". In: *IEEE Journal of Selected Topics in Signal Processing* 11.8, pp. 1301–1309. DOI: 10.1109/JSTSP.2017.2764438.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention is All you Need". In: ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Vögel, H.-J., C. Süß, T. Hubregtsen, V. Ghaderi, R. Chadowitz, E. André, N. Cummins, B. Schuller, J. Härri, R. Troncy, B. Huet, M. Önen, A. Ksentini, J. Conradt, A. Adi, A. Zadorojniy, J. Terken, J. Beskow, A. Morrison, K. Eng, F. Eyben, S. A. Moubayed, and S. Müller (2018). "Emotion-Awareness for Intelligent Vehicle Assistants: A Research Agenda". In: *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*. SEFAIS '18. Gothenburg, Sweden: Association for Computing Machinery, pp. 11–15. DOI: 10.1145/3194085.3194094.

Wagner, J., A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller (2023). "Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.9, pp. 10745–10759. DOI: 10.1109/TPAMI.2023.3263585.

Wang, J., W. Chai, A. Venkatachalapathy, K. L. Tan, A. Haghighat, S. Velipasalar, Y. Adu-Gyamfi, and A. Sharma (2022). "A Survey on Driver Behavior Analysis From In-Vehicle Cameras". In: *IEEE Transactions on Intelligent Transportation Systems* 23.8, pp. 10186–10209. DOI: 10.1109/TITS.2021.3126231.

Wang, J., J. M. Warnecke, M. Haghi, and T. M. Deserno (2020). "Unobtrusive Health Monitoring in Private Spaces: The Smart Vehicle". In: *Sensors* 20.9. DOI: 10.3390/s20092442.

Wang, K., X. Zeng, J. Yang, D. Meng, K. Zhang, X. Peng, and Y. Qiao (2018). "Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues". In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. ICMI '18. Boulder, CO, USA: Association for Computing Machinery, pp. 640–645. DOI: 10.1145/3242969.3264991. URL: https://doi.org/10.1145/3242969.3264991.

Wang, Y., L. Hespanhol, and M. Tomitsch (2021). "How Can Autonomous Vehicles Convey Emotions to Pedestrians? A Review of Emotionally Expressive Non-Humanoid Robots". In: *Multimodal Technologies and Interaction* 5.12. DOI: 10.3390/mti5120084.

Waytz, A., J. Heafner, and N. Epley (2014). "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle". In: *Journal of Experimental Social Psychology* 52, pp. 113–117. DOI: 10.1016/j.jesp.2014.01.005.

Wei, K., J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor (2020). "Federated Learning With Differential Privacy: Algorithms and Performance Analysis". In: *IEEE Transactions on Information Forensics and Security* 15, pp. 3454–3469. DOI: 10.1109/TIFS.2020.2988575.

Wei, W., Q. Jia, and Y. Feng (Dec. 2017). "Emotion recognition based on feedback weighted fusion of multimodal emotion data". In: *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1682–1687. DOI: 10.1109/ROBIO.2017.8324660.

Weiss, K., T. M. Khoshgoftaar, and D. Wang (May 2016). "A survey of transfer learning". In: *Journal of Big Data* 3.1, p. 9. DOI: 10.1186/s40537-016-0043-6. URL: https://doi.org/10.1186/s40537-016-0043-6.

Wiegand, G., M. Eiband, M. Haubelt, and H. Hussmann (2020). ""I'd like an Explanation for That!"Exploring Reactions to Unexpected Autonomous Driving". In: *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '20. Oldenburg, Germany: Association for Computing Machinery. DOI: 10.1145/3379503.3403554. URL: https://doi.org/10.1145/3379503.3403554.

Xin, D., S. Takamichi, and H. Saruwatari (2022). *Exploring the Effectiveness of Self-supervised Learning and Classifier Chains in Emotion Recognition of Nonverbal Vocalizations*. DOI: 10.48550/ARXIV.2206.10695. URL: https://arxiv.org/abs/2206.10695.

Zadeh, A., M. Chen, S. Poria, E. Cambria, and L.-P. Morency (2017). "Tensor fusion network for multimodal sentiment analysis". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by M. Palmer, R. Hwa, and S. Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1103–1114. DOI: 10.18653/v1/D17-1115. URL: https://aclanthology.org/D17-1115.

Zepf, S., J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard (June 2020). "Driver Emotion Recognition for Intelligent Vehicles: A Survey". In: *ACM Comput. Surv.* 53.3. DOI: 10.1145/3388790. URL: https://doi.org/10.1145/3388790.

Zhang, T., A. El Ali, C. Wang, A. Hanjalic, and P. Cesar (2020). "RCEA: Real-Time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, pp. 1–15. DOI: 10.1145/3313831.3376808. URL: https://doi.org/10.1145/3313831.3376808.

Zhao, J., X. Mao, and L. Chen (2019). "Speech emotion recognition using deep 1D & 2D CNN LSTM networks". In: *Biomedical signal processing and control* 47, pp. 312–323. DOI: https://doi.org/10.1016/j.bspc.2018.08.035.

Zhao, J., R. Li, J. Liang, S. Chen, and Q. Jin (2019). "Adversarial Domain Adaption for Multi-Cultural Dimensional Emotion Recognition in Dyadic Interactions". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. AVEC '19. Nice, France: Association for Computing Machinery, pp. 37–45. DOI: 10.1145/3347320.3357692.

Zhao, Y., Z. Liang, J. Du, L. Zhang, C. Liu, and L. Zhao (2021). "Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech." In: *Frontiers in neurorobotics* 15, p. 684037.

Zhu, M., J. Chen, H. Li, F. Liang, L. Han, and Z. Zhang (2021). "Vehicle driver drowsiness detection method using wearable EEG based on convolution neural network". In: *Neural Computing and Applications* 33.20, pp. 13965–13980. DOI: 10.1007/s00521-021-06038-y. URL: https://doi.org/10.1007/s00521-021-06038-y.