Multiclass Speech Emotion Recognition with Neural Networks: Investigations on Aspects of Input Data, Multilingual Modeling, and Data Scarcity

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Michael Neumann

aus Friedrichroda

Hauptberichter	Prof. Dr. Ngoc Thang Vu
Mitberichter	Prof. Dr. Björn Schuller

Tag der mündlichen Prüfung: 19. Juli 2021

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2021

Erklärung (Statement of Authorship)
Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und dabei keine andere als die angegebene Literatur verwendet habe. Alle Zitate und sinngemäßen Entlehnungen sind als solche unter genauer Angabe der Quelle gekennzeichnet.

I hereby declare that this text is the result of my own work and that I have not used sources without declaration in the text. Any thoughts from others or literal quotations are clearly marked.

(Michael Neumann)

Abstract

Emotions are an essential and inherent part of human behavior and communication. In spoken conversations, a lot of information is conveyed beyond the spoken words. This additional, paralinguistic information can for example change the meaning of words and sentences (e.g. by different intonation or word stress), or reveal a speaker's emotional state or other speaker characteristics, such as age, gender or personality traits. In present-day speech based human-computer interaction, a lot if not all of this additional information is not being conveyed. However, it is often argued that in order to achieve a truly effective human-computer interaction, it is inevitable to equip machines with emotional intelligence (Creed and Beale, 2008; Pantic and Rothkrantz, 2003). *Automatic recognition* of emotional states – in the present work from acoustic speech properties – is certainly the first step of any such 'emotion processing pipeline'.

Emotions are an inherently subjective phenomenon and different underlying theories exist of what they are and how they can best be represented in terms of labels or numerical scales. This makes it challenging and expensive to annotate data in order to teach machines to understand emotions. The typical procedure is that several human raters are asked to label a speech sample with some type of emotion representation (e.g. with emotion words like *anger* and *sadness*, or as values on scales of arousal and valence). Due to these challenging prerequisites as well

as the difficulty to collect realistic emotional speech in the first place, the main challenges for speech emotion recognition (SER) are to achieve a high level of accuracy and generalizability, and the problem of data scarcity (that is the lack of large, annotated datasets to train machine learning algorithms on).

One of the main contributions of this thesis is therefore to address these challenges by investigating different aspects of the typical SER pipeline in order to improve performance and to better understand sources of error in the models. For this, we employ a convolutional neural network (CNN) because of its ability to capture local patterns in the input. We present experiments on factors of input data, including the choice of acoustic features, the degree of naturalness of speech, and the duration of the input samples. We show that the type of speech (scripted conversations vs. improvised play) strongly affects the SER performance in terms of accuracy, and that a relatively short segment of a speech utterance can be sufficient for SER, a relevant finding for real-time requirements of applications. Further aspects of investigation pertain to generalizability of the trained models. We present experiments on multi- and cross-lingual SER on English and French speech, and assess the performance under noisy acoustic conditions. We found that arousal level prediction (how calm or excited someone is) is feasible across languages under certain conditions, but valence level prediction (positive vs. negative) does not well generalize.

The second main contribution of this thesis pertains to possible extensions to the typical SER pipeline, addressing performance improvement and data scarcity. We show that unsupervised representation learning on additional unlabeled speech data can be utilized for SER because the learned general-purpose speech representations implicitly contain information about emotional dimensions. Moreover, we show that training data for a

given target emotion label can be artificially generated by means of emotion style transfer using generative adversarial networks. Lastly, as another direction of extension, experiments on audio-visual emotion recognition are presented that demonstrate the benefit of adding visual information (facial expressions), especially to alleviate the effect of decreasing performance in noisy acoustic conditions.

In summary, this thesis shows that neural networks provide an efficient method for speech emotion recognition, and provides valuable findings to better understand and approach the challenges towards robust and generalizable SER.

Deutsche Zusammenfassung

Emotionen sind essentieller und inhärenter Bestandteil menschlicher Kommunikation. Gesprochene Konversationen beinhalten wesentlich mehr Informationen als nur die gesprochenen Worte. Diese zusätzlichen, paralinguistischen Informationen können beispielsweise die Bedeutung von Sätzen beeinflussen (z.B. durch unterschiedliche Intonation oder Betonung) oder Aufschluss geben über den emotionalen Zustand einer Person oder andere Eigenschaften, wie z.B. Alter, Geschlecht oder Persönlichkeitsmerkmale. In der heutigen Mensch-Computer-Interaktion können solche paralinguistischen Informationen nicht oder kaum übermittelt und genutzt werden. Jedoch wird häufig argumentiert, dass es unvermeidbar ist, Maschinen mit emotionaler Intelligenz auszustatten, um eine wirklich effektive Mensch-Computer-Interaktion zu erreichen (Creed and Beale, 2008; Pantic and Rothkrantz, 2003). Die *automatische Erkennung* von Emotionen – im Fall der vorliegenden Arbeit von akustischen Eigenschaften des Sprachsignals – ist der erste notwendige Schritt in jeglicher maschinellen Emotionsverarbeitung.

Emotionen sind grundsätzlich ein subjektives Phänomen und es existieren unterschiedliche Theorien, was Emotionen sind und wie sie am besten repräsentiert werden können, z.B. als nominale Kategorien oder als Werte auf verschiedenen Intervallskalen. Dadurch ist es schwierig und aufwändig Daten zu anno-

tieren, um damit Algorithmen zur Emotionserkennung zu trainieren. Die übliche Vorgehensweise ist, dass mehrere Personen eine Sprachprobe mit einer vorgegebenen Emotionsrepräsentation annotieren (beispielsweise mit Emotionswörtern wie *Wut* oder *Traurigkeit* oder mit Zahlenwerten auf Skalen für die Dimensionen Arousal und Valenz). Aufgrund dieser aufwändigen Vorbedingungen und weil es grundsätzlich schwierig ist, realistische emotionale Sprachdaten in großem Maßstab aufzunehmen, ergeben sich folgende Herausforderungen für die Emotionserkennung in Sprache: Datenknappheit (eng. data scarcity), also ein Mangel an ausreichend annotierten Sprachproben, um robuste Machine-Learning Modelle zu trainieren, und allgemein eine relativ niedrige Genauigkeit der Vorhersagen und mangelnde Generalisierbarkeit in Bezug auf neue, ungesehene Daten.

Basierend auf diesen Herausforderungen ist ein wesentlicher Beitrag der vorliegenden Arbeit die Analyse von verschiedenen Aspekten und Einflussfaktoren der Emotionserkennung mit dem Ziel, die Genauigkeit von Vorhersagen zu verbessern und Fehlerquellen der Modelle besser zu verstehen. Für diese Untersuchungen wurde eine spezielle Art von neuronalem Netzwerk, bezeichnet als Convolutional Neural Network (CNN), eingesetzt, um anhand von akustischen Merkmalen Emotionen (nominale Kategorien) zu klassifizieren. Untersucht wurden verschiedene Aspekte der Eingabedaten: die Auswahl der akustischen Merkmale (die Repräsentation der Daten), der Grad der Natürlichkeit der Sprache, und die Länge des Sprachsignals. Diese Experimente zeigen, dass die Art der Sprache (Drehbuch-basiertes Schauspiel verglichen mit Improvisation) einen erheblichen Einfluss auf die Ergebnisse hat, und dass ein kurzer Auschnitt einer Sprachäußerung ausreichen kann für die Emotionserkennung. Weitere untersuchte Aspekte zielen auf die Generalisierbarkeit des Modells ab. Sprachübergreifende Evaluierungen mit engli-

schen und französischen Sprachdaten demonstrieren, dass die Vorhersage von Arousal (wie ruhig oder erregt eine Person ist) unter bestimmten Bedingungen möglich ist, während die Vorhersage von Valenz (positiv oder negativ) sprachabhängiger erscheint.

Ein weiterer Schwerpunkt dieser Arbeit ist die Untersuchung möglicher Erweiterungen der typischen Machine Learninig Prozedur für Emotionserkennung, um das Problem der Datenknappheit zu addressieren und die Genauigkeit der Modelle zu verbessern. Untersuchungen mit sogenanntem unsupervised representation learning (d.h. dem automatischen Erlernen von abstrakten Repräsentationen des Eingabesignals ohne zusätzliche Annotationen) zeigen, dass solche Repräsentationen, die auf großen Sprachkorpora erlernt wurden, nützlich für die Emotionserkennung sein können, da sie implizit Informationen über bestimmte emotionale Dimensionen enthalten. Weiterhin wird eine Methode vorgestellt, um mittels sogenannter Generative Adversarial Networks (zu deutsch etwa 'erzeugende gegnerische Netzwerke') künstliche Trainingsdaten zu erzeugen, die eine vorgegebene Emotion darstellen. Zuletzt wird eine weitere mögliche Erweiterung in Form von mulimodaler Modellierung untersucht. Experimente unter Hinzunahme von visueller Information (Videoaufnahmen der spechenden Personen) bestätigen einen Mehrwert der zusätzlichen Modalität mit Bezug auf allgemein höhere Genauigkeiten und insbesondere um Verschlechterungen der Modelle bei lauten Umgebungsgeräuschen entgegenzuwirken.

Zusammenfassend zeigt diese Arbeit, dass neuronale Netzwerke eine effiziente Methode zur Emotionserkennung in Sprache darstellen und liefert wertvolle Erkenntnisse, um die Herausforderungen auf dem Weg zu einer robusten Emotionserkennung besser zu verstehen.

Acknowledgements

First and foremost, I would like to thank my advisor, Ngoc Thang Vu, for the great support and advice throughout the last years. Thanks for the trustful relationship and guidance, for the many great ideas and fruitful discussions, and for the space and freedom I was given to develop as a researcher in the best possible way.

I would also like to thank Björn Schuller for being part of the doctoral committee. Throughout the years of my Ph.D. his scientific work was often a source of inspiration for me. I feel honored and proud that he agreed to be on my committee.

During the almost eight years I have spent at the Institute for natural language processing in Stuttgart, many people supported and inspired me – teachers who became colleagues later on, my fellow students, as well as the students I supervised. In a generic way, I want to espress my gratitude to the whole institute for providing such a fruitful and friendly working environment. I want to express special thanks to Antje Schweitzer, who accompanied me during the whole time, as an excellent teacher during my Master studies and then as a colleague and friend. Thanks for always being supportive and encouraging.

Beyond the IMS, I had the chance to get to know many inspiring people, who positively influenced my work during my Ph.D. journey. I want to thank Felix Burkhardt and the team at audEERING GmbH for exchanging ideas, and thanks to Anton Batliner for fruitful conversations about paralinguistics and ethics.

Last but not least, I want to thank my parents for their unconditional support during my studies and my wife, Angela, and our kids, Elisa and Lasse, for their patience and support.

Contents

List of Abbreviations xix

List of Figures xxiii

List of Tables xxv

1 Introduction 1

1.1 Motivation 1

1.2 Research Context 7

1.3 Goals of this Thesis 10

1.4 Overview and Contributions 12

2 Background 17

2.1 Theories of Emotions 18

2.1.1 Basic Emotions 21

2.1.2 Dimensional Models of Emotion 24

2.1.3 Ordinal Representations of Emotions . . . 26

2.2 Emotional Speech Corpora 27

2.2.1 Annotation of Emotional Speech Datasets 28

2.2.2 Degree of Naturalness 31

2.2.3 Overview of Datasets 32

2.3 Acoustic Features 39

2.4 Neural Networks 44

2.4.1 Artifical Neural Networks 44

2.4.2 Convolutional Neural Networks 50

2.4.3 Autoencoders 52
2.4.4 Generative Adversarial Networks 53

3 Speech Emotion Recognition 55
3.1 Related work 57
3.1.1 Neural Networks for Speech Emotion Recognition 57
3.1.2 Cross-corpus and Cross-lingual Emotion Recognition 60
3.2 Method 62
3.3 Data and Acoustic Features 67
3.4 Features, Signal Length & Speech Type 72
3.4.1 Experimental Setup 72
3.4.2 Results 74
3.4.3 Error Analysis 80
3.5 Multilingual and Cross-Lingual SER 83
3.5.1 Experimental Setup 83
3.5.2 Results 85
3.5.3 Analysis of Attention Weights 89
3.6 Summary 92

4 Representation Learning 97
4.1 Unsupervised RL 101
4.1.1 Related Work 101
4.1.2 Methods 102
4.1.3 Data 103
4.1.4 Experimental Setup and Hyper-parameters 106
4.1.5 Results 108
4.2 CycleGAN-based Emotion Style Transfer 117
4.2.1 Related Work 117
4.2.2 Methods 120
4.2.3 Data and Features 124

4.2.4 Experimental Results 125
4.3 Summary 139

5 Audiovisual SER 143
5.1 Related Work 145
5.2 Methods 147
5.3 Data and Features 150
5.3.1 Datasets 150
5.3.2 Acoustic Features 151
5.3.3 Visual Features 153
5.4 Experimental Setup 153
5.5 Results and Analysis 156
5.5.1 Models trained on clean and applied to noisy audio 156
5.5.2 Models trained on single noise levels . . . 163
5.5.3 Data augmentation: Training on all noise levels 164
5.6 Summary 165

6 Conclusion and Future Directions 167
6.1 Summary and Key Findings 168
6.2 Ethical Considerations 170
6.3 Outlook 176

Bibliography 181

List of Abbreviations

ACNN	Attentive convolutional neural network
AE	Autoencoder
AI	Artificial intelligence
ANN	Artifical neural network
ASR	Automatic speech recognition
CNN	Convolutional neural network
CREMA-D	Crowd-sourced emotional multimodal actors dataset
CV	Cross validation
CycleGAN	Cycle consistent generative adversarial network
DBN	Deep belief network
DCT	Discrete cosine transform
DNN	Deep neural network
eGeMAPS	Extended Geneva minimalistic acoustic parameter set
FFT	Fast Fourier transform
FT	Fine-tuning
GAN	Generative adversarial network

GPU	Graphics processing unit
GRU	Gated recurrent unit
HCI	Human-computer interaction
HMM	Hidden Markov model
IEMOCAP	Interactive emotional dyadic motion capture database
LLD	Low level descriptor
lMFB	log Mel filter banks
LSTM	Long short-term memory
MFB	Mel filter banks
MFCC	Mel frequency cepstral coefficients
ML	Machine learning
MSE	Mean squared error
MTL	Multi-task learning
NLP	Natural language processing
NN	Neural network
PCM	Puls-code modulation
RECOLA	Multimodal corpus of remote collaborative and affective interactions
ReLU	Rectified linear unit
RL	Representation learning
SAM	Self-assessment manikins

SER	Speech emotion recognition
SNR	Signal-to-noise ratio
STL	Single-task learning
t-SNE	t-distributed stochastic neighbor embedding
UAR	Unweighted average recall
WA	Weighted accuracy

List of Figures

1.1 Relation between amount of data and machine learning performance 7
1.2 Different modalities for emotion analysis 9

2.1 Plutchik's wheel of emotions 25
2.2 Types of speech data 33
2.3 Self-assessment manikins 34
2.4 Illustration of the convolution operation in CNNs. 51

3.1 Attentive convolutional neural network architecture 63
3.2 Results with decreasing signal length 79
3.3 Error analysis of ACNN with multi-task learning . 81
3.4 Analysis of learned attention weights 90

4.1 ACNN model extended by autoencoder-based representations 104
4.2 Error analysis on IEMCOAP 109
4.3 Error analysis on MSP-IMPROV (cross-corpus evaluation) 110
4.4 t-SNE visualizations of ACNN representations for IEMOCAP 112
4.5 t-SNE visualizations of AE representations for IEMOCAP 113
4.6 t-SNE visualizations of ACNN representations colored by gender 115

4.7 t-SNE visualizations of ACNN and AE representations for MSP-IMPROV 116
4.8 CycleGAN framework for emotion style transfer . 121
4.9 Distributions of source, target, and synthetic features 128
4.10 Overlap of feature distributions between datasets 130
4.11 Overlap of feature distributions between emotion classes 131
4.12 Averaged confusion matriced for IEMOCAP . . . 136

5.1 Unimodal and multimodal feed-forward neural networks 148
5.2 Unimodal and multimodal neural networks with CNN as component 149
5.3 Visual feature processing pipeline 154
5.4 Results for all noise types 157
5.5 Performance of ‘clean’ models applied to noisy data (babble noise) 158
5.6 Confusion matrices for clean and noisy data I . . 160
5.7 Confusion matrices for clean and noisy data II . . 161
5.8 Performance on MSP-IMPROV for matched noise conditions 163
5.9 Performance on MSP-IMPROV with data augmentation 164

List of Tables

2.1 Overview of emotional speech datasets 37

3.1 Label distribution of IEMOCAP 68
3.2 Distribution of binary arousal/valence labels . . . 70
3.3 ACNN results on improvised speech 74
3.4 ACNN results on scripted conversations 74
3.5 ACNN results on IEMOCAP 75
3.6 Results for cross- and multilingual SER 86

4.1 SER Results with additional AE-representations . 108
4.2 Hyper-parameters of feed-forward neural networks 133
4.3 Results with real and generated training data on IEMOCAP 134
4.4 Results for cross-corpus evaluation 138

5.1 Hyper-parameters for unimodal and multimodal networks 155

1 Introduction

> The most important thing in communication is to hear what isn't being said
>
> Peter Drucker, interview in *Bill Moyers A World of Ideas*, 1989

1.1 Motivation

Natural language processing (NLP) as well as automatic speech recognition (ASR) are two of the driving research areas for many applications in the realm of natural language based human-computer interaction (HCI). In the last decades, a lot of efforts were made in those fields to improve the performance and broaden the use and acceptance of (spoken) language technology (Hirschberg and Manning, 2015; Yu and Deng, 2016). The main focus in speech processing has long been on modeling and predicting *what* is being said (i.e. automatic speech recognition), but not *how* something is being said. However, the *how* is crucial for communinication because the meaning of the same sentence can completely change depending on the intonation (or other prosodic events). For example, the sentence *She finished her thesis* with a falling intonation at the end is understood as a statement, whereas the same sentence with a rising intonation at the end is usually understood as a question. Similarly, the way how something is being said can convey certain emotional states. For example, *She finished her thesis* could be expressed joyfully and with excitement (e.g., expressing *It is finally done*), or it could be uttered in a surprised manner (e.g., to express *Wow, she is already done*). Because of such changes in meaning and other things we convey through our voice, like emotions or sarcasm, it is of such great importance – as Peter Drucker said – to understand what is not being said explicitly.

The automatic analysis of speech-related phenomena beyond pure linguistics has spawned the research field of computational paralinguistics, an area of increasing interest both in academia and industry. The Oxford English dictionary defines paralinguistics as "The branch of linguistics which studies non-phonemic aspects of speech, such as tone of voice, tempo, etc.; non-phonemic characteristics of communication; paralanguage." (Paralinguistics, 2020) Certainly, this is a very broad definition as the study of non-phonemic characteristics of communication includes a large variety of aspects (cf. Schuller et al. (2013) and Schuller and Batliner (2013) for an overview).

The present thesis deals with one of these aspects of computational paralinguistics that is recognizing a person's *emotional state* from acoustic speech properties, apart from the linguistic content, i.e. the meaning of spoken words. This topic, commonly referred to as **speech emotion recognition (SER)** is situated at the intersection of various research areas, namely paralinguistics, psychology, affective computing, and artificial intelligence. Moreover, with increasing availability of multimodal data and machine learning techniques, also NLP and computer vision play a role in a holistic approach to multimodal affective computing. The term affective computing was coined by Rosalind Picard (1995) and it emerged to a research field on its own. Broadly spoken, it comprises the study and development of all kinds of systems that are able to detect, recognize, interpret, or simulate human affects. It is often argued that in order to achieve a truly effective human-computer interaction, it is inevitable to equip machines with such *emotional intelligence*. *Recognition* of affective states (in our case from the speech signal) is certainly the first step of any such 'emotion processing pipeline'.

Possible use cases and applications that benefit from SER already or could potentially incorporate it, include, but are not limited to:

- Call centers: emotion recognition and monitoring is used to improve customer satisfaction (Burkhardt et al., 2009)
- Recruitment: emotion recognition can be part of automated personality tests to find suitable candidates for a job position (however, this is strongly critized, cf. Raghavan et al. (2020))
- Personal assistants: SER is likely going to be integrated in voice-based assistants such as Apple's Siri or Microsoft's Cortana to improve user experience and adapt to the user's affective state (Parmar, 2019)
- Gaming: The plot of a video game or the behavior of game characters could be adapted based on the player's emotional state or stress level (Jones and Sutherland, 2008)
- Health care: SER is used in various areas in digital health care, including diagnosis and monitoring of specific neurological conditions (e.g. autism), and assistive and empathic care robots (Cummins et al., 2018)

Despite being already used to some extent in certain (well-constrained) applications and scenarios, SER is still facing lots of challenges. The basic approach to automatic recognition of emotional states is to use machine learning (ML) techniques to learn from data. Hence, a prerequisite is the availability of *training data*, in our case speech samples. These speech data need to be annotated, i.e. they need to have labels, such as emotion classes (e.g. *anger*, *happiness*, *sadness*). This data labeling process is usually done by human annotators who assign a label to each speech sample they listen to. Because emotion perception is characterized by subjective notions of what emotional states are and how they are expressed, no 'ground-truth' can exist in any labels. Therefore, human 'performance' in this task, or rather agreement in perception between different people is relatively low (cf. Elfenbein and Ambady (2002) for a meta-analysis on human emotion recognition).

Thus, the main challenges in SER are the lack of large, annotated datasets to train such ML algorithms on, and the relatively low accuracy of predictions, which is tightly coupled to the subjective nature of emotions and the resulting low interrater agreements between human annotators. While ASR for example is robust and good enough to be used in many everyday use cases (for certain languages), such as personal voice assistants, voice control in cars and smart homes, call centers and many more, this is far from being the case for emotion recognition.

Recent advancements in machine learning with deep neural networks (NNs) have also impacted the field of SER immensely. However, there are many open questions that have been and still are being actively researched, such as the search for optimal acoustic features, the question of what is actually learned by the employed neural networks (which are often denoted as 'black box' because it is difficult to understand and interpret the internal processing), or the question of how to efficiently collect or create more training data for the task. As illustrated above, (labeled) training data is one pivotal factor of success in ML. One of the reasons why (deep) neural networks work so well in many areas nowadays is the availability of huge amounts of training data. Figure 1.1 illustrates this relation between the amount of training data and model performance with regard to traditional ML approaches vs. NNs. For small datasets, the relative order of the algorithms is not clearly defined, but as the figure suggests, it is likely to achieve higher performance on very small datasets with traditional approaches; and it is certainly the case that with an increasing amount of training data, NNs outperform these methods.

This thesis aims to provide anwsers to some of these open questions and challenges. Throughout this work, we explore various aspects of SER, including the use of convolutional neural networks for modeling, methods to make use of additional unlabeled speech data, and audiovisual ML to improve performance in noisy conditions – always with a focus on thorough analyses of the results to gain insights about the used models, beyond of just comparing performance metrics. In the following we describe the broader research context and the goals of this work.

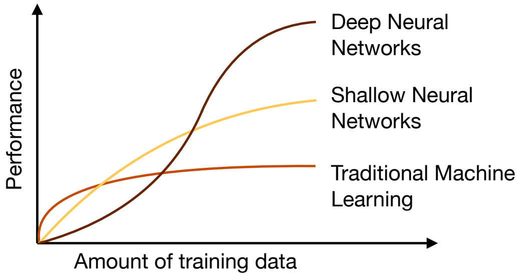


Figure 1.1: Illustration of the relation between the amount of training data and performance with respect to different machine learning paradigms.[1]

1.2 Research Context

Human emotions are an inherently multimodal phenomenon. This means, several modalities or information channels are involved when a person expresses a certain emotion. Examples for these different channels are physiological signals like an increased blood pressure, facial expressions, a change in the tone of voice, or the choice of words to express emotions. Focusing on methods that can be employed remotely and in a scalable fashion (e.g. with the use of smartphone microphones and cameras), speech and language as well as facial expressions are the commonly exploited modalities for automatic analysis of human

[1] Graphic inspired by Andrew Ng's lecture 'Neural Networks and Deep Learning', `https://coursera.org/share/0a61ee951fa8f1edee71a8bdf764c145` [Accessed March 04, 2021]

emotions. Figure 1.2 illustrates the research fields related to these modalities, which are also closely related to each other. As already mentioned above, we approach the task of emotion recognition as a paralinguistic task, hence we focus only on speech acoustics, and do not take into account the lexical information (which could be retrieved from the speech signal by using ASR). In contrast to SER, where speech prosody plays the most important role, the related disciplines of sentiment and (textual) emotion analysis are dealing with the content and meaning of written text. The difference between these two is that sentiment analysis is about binary classification into positive and negative sentiment, as opposed to a more fine-grained spectrum of emotional states; a well known use case of sentiment analysis is the classification of product reviews. Finally, although not directly linked to speech and language, facial expression analysis is depicted in this overview because the interplay between these modalities and their combined use in multimodal machine learning has gained a lot of attention in recent years. We used this visual channel as additional information to complement SER in noisy acoustic conditions.

Regarding the methodology for SER, it is fair to say that neural networks have replaced traditional ML methods like decision trees, support vector machines or hidden Markov models – at

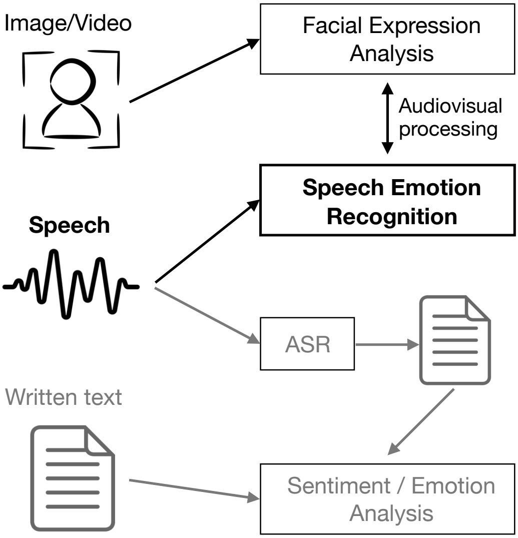

Image/Video
Facial Expression Analysis
Audiovisual processing
Speech Emotion Recognition
Speech
ASR
Written text
Sentiment / Emotion Analysis

Figure 1.2: Different modalities that are commonly exploited in affective computing and closely related to each other. The focus of this thesis is **speech emotion recognition** (parts in boldface). Facial expression analysis plays a marginal role in this work, while written text and lexical information (parts in gray) are not taken into account.[2]

least in an academic context. Deep learning has gained traction in the field roughly since 2013, and since then many varieties of neural networks have been explored. The majority of these

[2] The following icons from the Noun Project (`https://thenounproject.com`) are used in Figure 1.2: *portrait* by ffabio44, *Soundwave* by Maxim Kulikov, and *Document* by Rediffusion

approaches can be assigned to one of the following two general types of NNs: *feedforward* or *recurrent* NNs (see section 2.4 for details). Other types exist, as well as many variants and extensions, but for the sake of putting this work into the broader research context, we retain this division. While both types of NNs are well suitable for SER and can also be combined (Zhao et al., 2017), the approach taken in this thesis focuses on CNNs. We model the task of emotion recognition on utterance level, under the assumption that in a relatively short speaker utterance only one emotional state is expressed (and therefore temporal dependencies within an utterance are not as crucial as in other sequence modeling tasks).

1.3 Goals of this Thesis

The goals of this thesis can be split up into two areas: **systematic investigations** of various aspects of SER, and **extensions** to the basic approach of training and evaluating a machine learning model on a given dataset. Approaching the two challenges outlined before – data scarcity and low performance due to the complex nature of emotions – we aim at investigating ways to improve performance and to better understand sources of error, as well as proposing extensions that address data scarcity and consequently also improvements in accuracy.

Systematic investigations: Since SER is a relatively young and fast evolving research field, there are many aspects that have not yet been investigated. The work presented in this thesis aims to shed light on some of these. Specifically, we investigate the choice of acoustic features as input to a CNN model, the impact of utterance length on SER performance, the feasibility of cross-lingual and multilingual SER, and the performance under noisy acoustic conditions. In doing so, we focus on comprehensive analyses that go beyond reporting and comparing accuracy numbers; among others, confusion matrices are one helpful tool used throughout this work to uncover and interpret error patterns.

Extensions: The basic ML approach to a given task is to train a model on some labeled dataset and evaluate its performance (either on a separate test set or by means of cross validation). In order to improve the performance and robustness of the models, the second goal is to investigate possible extensions to this basic approach. Concerning the CNN model itself, we propose and analyze an attention mechanism, which is a method to make a NN learn to focus on specific parts of the input data. Furthermore, we incorporate a multi-task learning paradigm where two different types of output labels are predicted (a secondary or auxiliary task is added) in order to improve the accuracy on the main task.

Another extension pertains to data (and the problem of data scarcity). As illustrated above with Figure 1.1, having a reasonably large amount of training data available is a key factor of success for deep learning. Because of this and because existing emotional speech datasets are relatively small, we explore two directions of making use of additional, unlabeled speech data: representation learning and generative modeling to create additional training data. The last type of extension concerns multimodal processing: we investigate how audiovisual emotion recognition can improve the performance, specifically in noisy conditions.

1.4 Overview and Contributions

In the following, we outline the structure of this thesis and state the main contributions presented in each chapter.

Chapter 2 provides the necessary background for the presented work. The chapter is divided into two parts. First, we provide an introduction to emotion modelling and commonly utilized representations of emotions, along with a description of the used speech datasets and their annotation. Second, the technical background is introduced, including speech processing and acoustic feature extraction as well as background on neural networks, with a focus on convolutional neural networks.

Chapter 3 presents SER experiments with an attentive convolutional neural network (ACNN) framework. We first describe the neural network architecture and detail the acoustic feature extraction. Then, experimental results are presented and analyzed.

The ACNN was first presented in Neumann and Vu (2017), along with a comparison of different acoustic feature sets as input, a comparison between two types of speech (scripted conversations vs. improvised play), and an investigation of the impact of signal length on the SER performance. These systematic investigations of different aspects of SER using a CNN model are one of the main contributions of this thesis, and the proposed model also constitutes the basis for further experiments. The results of this first series of experiments are discussed in section 3.4. The main findings were that the recognition performance strongly depends on whether acted (based on a script) or freely improvised conversations are the object of study, and that it can be sufficient to analysize only a short audio snippet from the beginning of a user utterance to make a reasonable prediction of the emotional state.

Utilizing the same neural network architecture, we presented cross-lingual and multilingual experiments with a French and an English dataset in Neumann and Vu (2018). We showed that arousal prediction is possible in a multilingual setup as well as with cross-lingual training followed by fine-tuning on the target data, whereas valence prediction is more sensitive to cross-lingual training. Further, this publication presented an analysis of the learned attention weights of the ACNN. These experiments and their results are detailed in section 3.5.

Chapter 4 addresses one fundamental problem of speech emotion research, that is the lack of large scale, annotated datasets to train ML models on (known as data scarcity). We present experiments on unsupervised representation learning with autoencoders and on generating synthetic training data (data augmentation) by using a cycle consistent generative adversarial network (CycleGAN).

The first part of this chapter (section 4.1) builds up on the ACNN model proposed in Neumann and Vu (2017) and presents an extension to it by incorporating speech representations that are learned in an unsupervised fashion on a large, unlabeled speech corpus. This work was reported in Neumann and Vu (2019). One important finding was that the autoencoder representations, which are learned without any emotion labels, are notably discriminative for the distinction between low and high arousal and that they can improve the SER performance when integrated into the ACNN model.

A different way of approaching the data scarcity problem is to create additional (labeled) training data using a generative ML model. In the second part of this chapter (section 4.2) we present such a method that is based on CycleGANs, which employ adversarial training to generate synthetic feature vectors representing a certain target emotion. This work emerged from the Master thesis by Fang Bao, which was supervised by Ngoc Thang Vu and myself. We published the results in a joint effort in Bao et al. (2019) and the results are extended in the present thesis by the addition of a second feature set and by more in-depth analyses. The main contributions of this work is that the generated additional data can be beneficially used for data augmentation, and that we could achieve relatively high performance when training *only* on these synthetic feature vectors, compared to related studies.

Chapter 5 presents an analytical investigation of *audiovisual* emotion recognition in noisy acoustic conditions. With the addition of the visual modality we bring in yet another important aspect of emotion recognition research, since multimodal data processing and machine learning are becoming increasingly popular and useful due to the ubiquitous availability of smart devices equipped with cameras and microphones. Based on the analyses in Neumann and Vu (2021) we present results on how the addition of visual information can, to a large extent, alleviate performance declines of SER when applied in noisy acoustic environments.

Chapter 6 summarizes the key findings of this thesis and provides a discussion on ethical considerations with respect to the work presented here. Finally, some ideas for future directions are outlined.

2 Background

> Almost everyone except the psychologist knows what an emotion is. [...] The trouble with the psychologist is that emotional processes and states are complex and can be analyzed from so many points of view that a complete picture is virtually impossible
>
> Paul T. Young, *Feeling and Emotion* in Handbook of General Psychology, 1973

This chapter introduces concepts and definitions that build the foundation for this thesis. The necessary background knowledge can be divided into two big parts: (I) a conceptional foundation concerning emotion theories and representations and (II) technical background concerning speech signal processing and machine learning methods.

In the first part, an overview of the theory of emotions is provided (section 2.1). This shall provide the necessary background to understand why different theories and representations of emotions exist and to be able to compare and discuss different representations and their simplifications with respect to speech emotion processing, or more general computational processing of human emotions. Further, the first part contains an overview of emotional speech corpora (section 2.2) that are subject of the experiments in this thesis, since the decision for a certain emotion representation is tightly coupled with available data and its annotation. Hence, different annotation schemes and related datasets are described. Specific details of the data with respect to different experimental settings are contained in the respective experimental chapters.

The second part of this chapter provides technical background knowledge that lays the foundation for all experiments presented in this thesis. This is divided into two main topics: speech signal processing (describing speech representations for emotion recognition; section 2.3) and machine learning (introducing the utilized neural network types; sections 2.4).

2.1 Theories of Emotions and their Representations

What is an emotion? This central question has occupied researchers from many fields, such as philosophy, evolutionary biology, psychology, and neuroscience for centuries – and there is no easy answer to it. In fact, in consequence of being such a multidisciplinary research subject, there are many definitions of what emotions are and how they are elicited (Gendron and Feldman Barrett, 2009; Kleinginna and Kleinginna, 1981).

Having a closer look at the history of psychological theories of emotions, Gendron and Barrett identify three fundamental approaches: (a) “basic emotion” concepts, (b) “appraisal” theories, and (c) “psychological constructionist” approaches (Gendron and Feldman Barrett, 2009). These concepts strongly influenced how emotions are viewed and represented with respect to the technical task of emotion recognition that we are concerned with. However, it is important to note that the field of SER and the work presented in this thesis in particular is not aiming at exactly modeling psychological phenomena, nor do we contribute to open questions in these areas. Instead, simplifying assumptions need to be made and the term *emotion* is used rather in its broad, ‘everyday-language’ sense. As the introductory quote states, almost everyone has a strong intuition of what happiness is or feels like, and how people might sound when they are angry. Still, it is inevitable to define the concept ’emotion’ and have a look at how (and why) emotional states can possibly be detected from a person’s speech and facical expressions.[1]

Paul and Anne Kleinginna reviewed over 90 definitions from the diverse literature on emotions and proposed the following definition as a working model (Kleinginna and Kleinginna, 1981):

[1] While this thesis is mainly about *speech* emotion recognition, we also have a look at facial expressions and a multimodal approach to emotion recognition in chapter 5.

> Emotion is a complex set of interactions among subjective and objective factors, mediated by neural-hormonal systems, which can (a) give rise to affective experience such as feelings of arousal, pleasure/displeasure; (b) generate cognitive processes such as emotionally relevant perceptual effects, appraisals, labeling processes; (c) activate widespread physiological adjustments to the arousing conditions; and (d) lead to behavior that is often, but not always, expressive, goal-directed, and adaptive.

In the realm of automatic emotion recognition, we are mostly concerned with the third point of this definition, (c) physiological adjustments to the arousing condition. These physiological phenomena are what can be perceived from the outside, such as facial expressions, changes in speech prosody, changes in heart rate, blood pressure, and skin conductance, to name a few phenomena from different modalities. In turn, these observations of physiological changes are used to infer an underlying emotional state.

This section summarizes the most frequently utilized approaches to emotion representation in the field of speech emotion recognition along with some historical context on the development of emotion theories. However, this overview is by no means a complete picture of the various theories that exist. The reader might refer to Gendron and Feldman Barrett (2009) for a more detailed historic summary. Here, we focus on the types of emotion representations, not the underlying psychological phenomena. In addition to the two prevailing approaches – *basic emotions* and *dimensional models* – we present more recent (and therefore less frequently used) representations at the end of this section.

2.1.1 Basic Emotions

Historically, emotion research was foremost concerned with facial expressions. In his publication *The Expression of the Emotions in Man and Animals*, Charles Darwin laid out the foundation of the basic emotions theory. About the universal nature of emotional expressions Darwin stated, "We can thus also understand the fact that the young and the old of widely different races, both with man and animals, express the same state of mind by the same movements." (Darwin and Prodger, 1872, p. 352)

Contrary to Darwin who believed that emotions are mental states that trigger physiological reactions ("we cry because we feel sorry"), William James was the first one to hypothesize that such bodily changes follow directly an exciting event and that our feeling of these reactions is the emotion ("we feel sorry because we cry"). This view on emotion became known as the James-Lange theory. James characterized what he called *standard emotions* as having a "distinct bodily expression" (James, 1884, p.189). He identified fear, grief, love, and rage as such.

Nowadays, one of the most cited proponents of the basic emotions theory in the field of SER is Paul Ekman. In his early work, he identified and studied six basic emotions and their associated facial expressions, which are universally found in humans across cultures: *anger, disgust, fear, happiness, sadness* and *surprise* (Ekman, 1970). It should be noted that Ekman's basic emotions framework does explicitly allow for additional 'non-basic' emotions ("All the emotions which share the characteristics I have described are basic. If *all* emotions are basic, what then is the value of using that term?" (Ekman, 2000, p. 57)).

The basic emotions representation is frequently used as annotation scheme for emotional speech data (and facial expressions). Especially in the early days of automatic facial/vocal expression recognition, those basic emotion categories have been the most prevelant way to represent emotions (Sebe et al., 2005). Speech corpora are typically segmented into utterances (sometimes referred to as speaker *turns* in the case of dialogs) and the most common annotation with basic emotions is to assign one emotion class from a given set of classes to each utterance (cf. section 2.2 on page 27 for details). This annotation scheme is convenient to implement (labeling each utterance with one class is fast) and resembles what people intuitively think of emotion classification (most people have an intuitive notion of how sad speech or angry speech sounds like). However, the approach also has problems. First, emotions and emotion words are highly subjective; different people have different notions of sadness or happiness. To attenuate the effect of subjectiveness, corpora are usually annotated by at least three raters (the more the better). Second, a fixed set of discrete emotion classes does not allow for blended emotional states or for other emotion words that are not contained in the given set. One way to collect more fine-grained annotations in this regard is to assign a primary class and optionally a secondary class to an utterance (see for example MSP-IMPROV (Busso et al., 2017)). However, the interpretation of these secondary classes remains difficult.

Throughout this thesis, we mainly use the basic emotions representation, and mainly four classes commonly found in available datasets: *anger, happiness, neutral state, sadness.* Although this is a simplification of real, complex emotional states, using these classes serves as a reasonable representation to implement into many applications. For example, in a call center application, the distinction between *happiness/neutral state* on the one hand and *anger* on the other hand is the crucial classification to make in order to estimate customer satisfaction. Of course, depending on the application, a different set of categories might be appropriate; for example, in mental health monitoring it is conceivable that states of fear/anxiety need to be detected. Another simplification is that an emotional state is the same throughout an utterance of a speaker. Again, in many application contexts, especially within dialog systems (where user turns are often short), this is a valid assumption. Hence, it is important to note that we are not aiming at developing a realistic psychological model of emotional states (in their complexity and with temporal variations), but rather contibute to the development of application-oriented systems, which can recognize a constraint (and simplified) set of emotion classes.

2.1.2 Dimensional Models of Emotion

Besides the notion of universal, basic emotions (represented as distinct categories), another well-known and widely utilized model is the representation of emotional states on independent, continuous bi-polar dimensions. Wilhelm Wundt was the first to propose the three dimensions *valence, arousal* and *intensity* to describe emotions (Wundt, 1897). Williams and Sundene (1965) studied facial and vocal expressions, suggesting that the two dimensions 'general evaluation' (valence) and 'activity' (arousal) have generality between both modalities across different emotions. James Russell hypothesized that the following three dimensions are both necessary and sufficient to define emotional states: pleasure ↔ displeasure, degree of arousal, and dominance ↔ submissiveness (Russell and Mehrabian, 1977).

A hybrid model that combines basic emotion categories and dimensional theories is Robert Plutchik's wheel of emotions, depicted in figure 2.1 on the next page. In his psycho-evolutionary theory of emotions (Plutchik, 1980), he identified eight basic

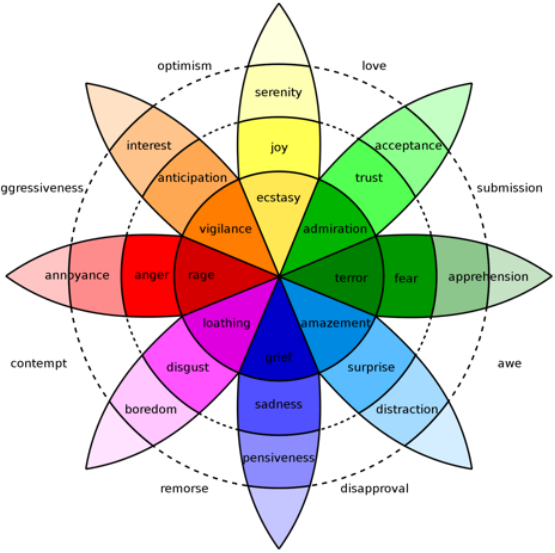


Figure 2.1: Plutchik's wheel of emotions. Opposite emotions form bipolar emotion pairs, e.g. *joy* vs. *sadness*. Color intensity represents the intensity of the emotions, along with different emotion words. Image source: (Machine Elf 1735, 2011)

emotions (Ekman's six plus *trust* and *anticipation*), which are – from an evolutionary perspective – distinctive triggers of behavior with high survival value (e.g. *trust* inspires behaviors like sharing and grooming). The wheel of emotions arranges emotions in concentric circles where outer circles represent more complex emotions and decreasing intensity.

2.1.3 Ordinal Representations of Emotions

If we distinguish emotion representations by the *level of measurement* that is used for annotation, we have so far mainly considered *nominal* labels (basic emotions) and *interval* scales, such as continuous values of arousal and valence. A different perspective that is endorsed in Yannakakis et al. (2018) is the *ordinal* nature of emotions, that is a rank-based representation. The authors argue that ordinal information is what people can deliver most reliably, i.e. it is easier and more reliable to compare two data samples and rank them on a given scale (e.g. based on intensity or on proximity to a prototypical state), instead of assigning specific values or labels to an individual sample in isolation. They provide an extensive background from multiple disciplines (behavioral economics, neuroscience, machine learning, psychology, affective computing, philosophy, marketing) to demonstrate why an ordinal view is appropriate and might even be the most natural representation. Therefore, Yannakakis et al. advocate for a holistic *ordinal* affective computing pipeline, including data annotation, data processing, and modeling, for which statistical analysis or preference learning (Fürnkranz and Hüllermeier, 2010; Yannakakis, 2009) can be used.

Although this type of representation is not in the scope of this thesis, we included it in this overview for the sake of completeness and because it seems to be a promising alternative modelling approach to the aforementioned basic emotions and dimensional models. Note, that a mapping between certain representations is feasible (cf. Yannakakis et al. (2018) for details). For instance, an interval scale can be treated as ordinal representation without any information loss.

2.2 Emotional Speech Corpora

The essential foundation for every machine learning problem is *data* to learn from. In fact, the success of deep learning in many areas in NLP and speech processing can be attributed to two main driving forces: increasing computational power and algorithmic efficiency on the one hand and the availability of large (annotated) datasets for training these algorithms on the other hand. While for certain tasks, such as ASR, the amount of available training data is enormous[2] and the ground-truth transcriptions are mostly objective and undisputable, for SER that is not the case. The size of available annotated emotional speech datasets is orders of magnitude smaller than the available data for ASR training (cf. Table 2.1). The main reasons for this are the difficulty to collect naturalistic emotional speech and the expensive annotation of emotion labels, being highly subjective and debatable (Devillers et al., 2005; Schuller et al., 2011a).

[2] To give an example of the large available datasets for ASR, the English dataset from Mozilla's freely available *common voice* project contains 1,469 hours of verified, transcribed speech (as of June 2020), cf. `https://commonvoice.mozilla.org/de/datasets` [accessed Dec. 15, 2020]

In this section we briefly discuss two of the main challenges in emotional speech data collection, namely different *annotation schemes* and the *degree of naturalness* of speech. We then provide an overview of the datasets that are used for experiments throughout the thesis and describe commonalities and differences between them.

2.2.1 Annotation of Emotional Speech Datasets

The labeling procedure of emotional states (and therefore, the underlying representation model) is one of the main differences between datasets. The basic emotions model is reflected in distinct *emotion classes*, which we will also call *categorical* annotations. The most common categorical annotation scheme is to assign one label for a whole utterance. While emotion classes are convenient to use for the annotation procedure and for training classifiers, these hard boundaries also introduce problems: a closed set of emotion words might exclude certain states; different raters might have different notions of words like *anger*, *happiness* or *sadness*; and it is difficult to represent mixed emotions.

The second major representation of emotions, namely continous dimensional representations, has been used to annotate data more and more recently. The most common approach here is to use the two dimensions arousal and valence and annotate speech data on a certain scale for each dimension (e.g. continuous values from -1 to 1). Sometimes, dominance is added as a third dimension. When using dimensional models as foundation, the task of predicting an emotional state can either be formulated as a regression (predicting a value on the continuous scale) or a classification problem (by grouping values into discrete classes[3]).

Different from the utterance-level annotation used with the basic emotions approach, dimensional labels are often annotated time-continuously. For this, annotation tools like Feeltrace (Cowie et al., 2000), Anvil (Kipp, 2001) or Annemo[4] are utilized, where a human annotator watches or listens to the stimulus and simultaneously annotates the sample, either by moving the mouse pointer in the two-dimensional space of arousal and valence as in Feeltrace or by moving a slider on one dimension at a time as in Annemo. The annotations are then the recorded positions of the mouse pointer or slider at a certain sampling rate, e.g. one value every 40ms between -1 and 1 on the arousal dimension. One main advantage of continuous annotations is that temporal dynamics are modeled, which is especially relevant for longer stimuli and can provide a more detailed picture of a speaker's emotional state throughout a complete conversation.

[3] Such grouping or any other transformation of labels can, however, lead to problems. This is discussed in chapter 3

[4] `https://diuf.unifr.ch/main/diva/recola/annemo` [Accessed March 06, 2021]

However, due to the more complex annotation process, there are also challenges to overcome and more sources for variation in the annotations compared to the basic emotion approach. Kessler et al. (2015) investigated the effects of different annotation tools on the labeling process and found that the resulting labels vary depending on the used software. One main challenge is the dynamically varying time delay between an expression and the rater's reaction. Yet, methods have been established for temporal alignment of annotations in order to create a reliable 'Gold Standard'[5] (Zhou and De la Torre, 2015).

There are also dimensional annotations that are done on utterance level and therefore do not pose the problem of time delay. In this case, labelers are asked to rate the arousal or valence level of the whole utterance, e.g. on a Likert scale or with self-assessment manikins (cf. Figure 2.3 on page 34).

For the work presented in this thesis, we followed an utterance-level modeling approach (as described in section 1.2 in the Introduction), and therefore used categorical annotations and framed the task of SER as a classification task.

[5] Along the lines of Kossaifi et al. (2019), we use the term 'Gold Standard' to refer to emotion annotations instead of 'ground-truth', because no objective truth exists for inherently subjective phenomena like emotions

2.2.2 Degree of Naturalness

The second major difference between datasets is the type of speech, and closely related the eliciation of emotional states. Speech type is often broadly divided into *acted* vs. *spontaneous* interactions. In the early days of SER, available data was usually always acted, that is, speakers are given a script or sentence, which they shall utter with a certain emotional expression. While it is quite straightforward to obtain data in such a way, the crucial problem is that such speech data is not naturalistic and realistic. Therefore, researchers and developers need to be careful with any conclusions drawn from acted datasets with respect to real-world applications; or, as Schuller et al. put it:

> In retrospect, the concentration on a few acted emotions at the beginning of the whole endeavour resulted in a sort of reality shock when non-prompted, realistic, and sparse events were addressed. (Schuller et al., 2011a, p.1065)

However, this does not ultimately render acted speech useless. These datasets are still useful as benchmark data to evaluate and compare models on, in order to advance the development of systems that then can be transferred to more realistic settings. One simply needs to be aware of the differences in speech and of the fact that models trained on acted or read speech will not generalize well on spontaneous speech 'in the wild'. Furthermore, another legitimate reason to use acted (prototypical) emotional speech can be research on acoustic parameters in a controlled setting (e.g. to compare different expressions of the same lexical content in identical, clean recording conditions).

In recent years, emotion recognition 'in the wild' has received increasing attention (see for example Dhall et al. (2013); Pandit et al. (2018); Kossaifi et al. (2019)). As the term suggests, the goal is to leave clean and controlled laporatory conditions and tackle the much more challenging data found in the real world.

For the present work, we distinguish types of speech data more fine-grained than only acted vs. spontaneous. In the datasets presented below, we identified four different types: read speech, scripted conversations (actors perform a given script), improvised play (actors improvise based on hypothetical scenarios), and free speech (i.e. natural, spontaneous interactions). Figure 2.2 on the next page illustrates that these speech types are of increasing naturalness in the presented order.

2.2.3 Overview of Datasets

Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)

IEMOCAP is a widely used multimodal and multispeaker corpus that has been created at the *Signal Analysis and Interpretation Laboratory (SAIL)* at the University of Southern California

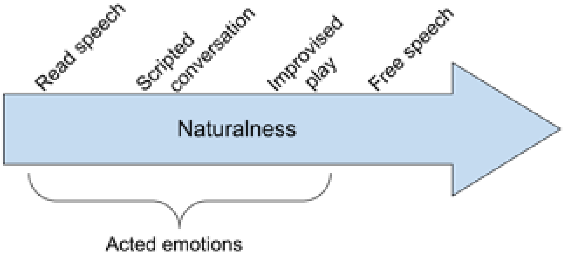


Figure 2.2: Four different types of speech with respect to naturalness of emotions.

(Busso et al., 2008). It consists of approximately 12 hours of audiovisual data (speech, video, facial motion capture markers) from two recording scenarios: scripted and improvised scenarios. The corpus is divided into five sessions, each of which contains English dyadic interactions between a female and a male speaker (ten speakers total, seven professional actors and three senior acting students). The conversations have been manually segmented into dialog turns, resulting in 10,039 utterances in total.

Annotation of emotional states was done in two ways: (a) with categorical emotion descriptors, and (b) with continuous emotion descriptors on the three dimensions arousal, valence, dominance. For (a) categorical descriptors, each turn was annotated by three raters (plus a self-report by the actors themselves) with one of the following emotion labels: *anger, sadness, happi-*

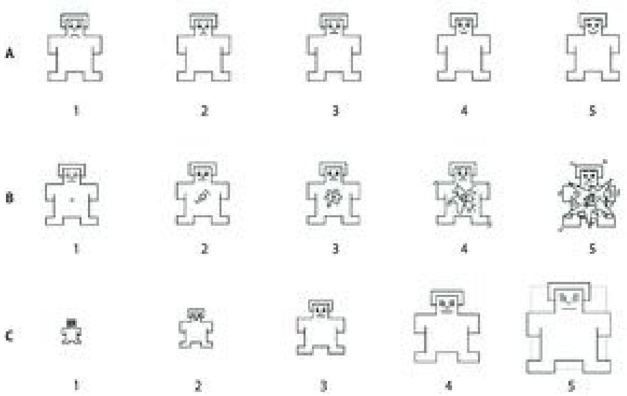


Figure 2.3: Self-assessment manikins (SAM) for the attributes A) valence, B) arousal, and C) dominance. Source: Galindo-Aldana et al. (2017)

ness, disgust, fear and surprise (Ekman's basic emotions), plus *frustration, excitement and neutral state.* Raters could additionally select *other* if none of the labels was adequate and they were allowed to select two labels to account for blended emotions. For (b) continuous descriptors, each turn was annotated by two raters (plus self-report) with the use of self-assessment manikins (Bradley and Lang, 1994) for the dimensions *valence* (1-negative, 5-positive), *arousal* (1-calm, 5-excited), and *dominance* (1-weak, 5-strong), depicted in Figure 2.3.

In the present thesis, the IEMOCAP data was used for the experiments presented in chapters 3 and 4.

MSP-IMPROV

The MSP-IMPROV dataset (Busso et al., 2017), a multimodal corpus of dyadic interactions between six speaker pairs, was collected at the Multimodal Signal Processing Laboratory (MSP) at the University of Texas at Dallas. The corpus is similar to IEMOCAP, in that specific emotion-eliciting scenarios were designed, which have been improvised by acting students. Different from IEMOCAP, this dataset contains specific target sentences that had to be uttered in each scenario for each of the four emotions *happiness, sadness, anger, neutral state.* In addition to the improvised scenarios, the target sentences were recorded as isolated read speech, and the corpus also contains annotated interactions between the speakers during preparation and breaks. As for the videos, speakers were recorded frontally in front of a green screen, which makes the videos of MSP-IMPROV more suitable for facial expression analysis than those of IEMOCAP (where the focus is on the motion capture markers).

Annotation of MSP-IMPROV was done similar to IEMOCAP, but only for the four targeted emotion classes, plus a class *other.* Dimensional annotations have been obtained for arousal, valence, dominance with the five-item SAMs (Figure 2.3). For MSP-IMPROV, crowdsourcing via Amazon Mechanical Turk was used for annotation, instead of a small group of trained evaluators; each utterance is annotated by at least five raters.

We used MSP-IMPROV in the studies presented in chapters 4 and 5.

Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)

The CREMA-D dataset (Cao et al., 2014) contains crowdsourced video clips with English read speech. 91 speakers were asked to read twelve target sentences in six different emotions (*happiness, sadness, anger, fear, disgust*, and *neutral state*). The videos show the speakers in front of a green screen. Annotations have also been obtained through crowdsourcing. Each sample is annotated for the emotion class and intensity (on a continous slider ranging from ‘mildly’ to ‘extremely’) by at least six raters. There are individual annotations for audio-only, video-only (without sound) and audiovisual data, from which we use the latter in the experiments presented in chapter 5.

Multimodal Corpus of Remote Collaborative and Affective Interactions (RECOLA)

RECOLA (Ringeval et al., 2013) is a multimodal dataset of French speech that consists of dyadic conversations during a video conference in which participants had to solve a collaborative task. From a total of 46 speakers, a subset of 23 speakers is publically available, comprising 1,308 annotated utterances.

RECOLA is annotated with continuous values for arousal and valence in the range [-1, 1] at a rate of 40ms. Annotation was done by six raters with the web-based tool ANNEMO, which had been developed in conjunction with the corpus creation.

RECOLA is the only non-English dataset that has been used for this work, namely in the multilingual study presented in chaper 3.

	Lang	Speakers	Speech type	Samples	Annotation	Hours of speech
IEMOCAP	en	10	acted & improvised	10,039	utterance level 10 classes A/V/D scores	12
MSP-IMPROV	en	12	improvised & free & read	8,438	utterance level 5 classes A/V/D scores	9.6
CREMA-D	en	91	read	7,442	utterance level 6 classes	5.25
RECOLA	fr	23	free	1,308	continuous A/V values in [-1,1]	2

Table 2.1: Overview of emotional speech datasets used in this thesis.
Lang: language, A: arousal, V: valence, D: dominance.

General Remarks on Evaluation and Comparability of Results in Speech Emotion Recognition

Reseach in SER is facing the problem that reported results are often not directly comparable to each other because no standardized test conditions are defined. To approach this issue, various challenges have been carried out in recent years to set benchmarks on pre-defined test sets. The first of its kind was the Interspeech 2009 Emotion Challenge (Schuller et al., 2009b), which used the FAU Aibo dataset (Batliner et al., 2008) as benchmark dataset. Since then, the 'Interspeech Computational Paralinguistics Challenge' (ComParE) was established as annual event and it has covered many paralinguistic topics beyond emotion.[6] Other challenges followed like the 'Audio/Visual Emotion Challenge' (AVEC) since 2011 (Schuller et al., 2011b) or the 'Emotion Recognition in the Wild Challenge' (EmotiW) since 2013 (Dhall et al., 2013).

However, for the datasets used in this thesis (with the exception of RECOLA), no such challenge baselines or other standardized benchmarks exist. Hence, fair comparison under exactly the same test conditions remains almost impossible. Reasons for this lie in different evaluation methods (e.g. held-out test sets vs. cross validation), in different underlying modeling approaches (e.g. basic emotions vs. continuous dimensions), in different reported performance metrics (e.g. weighted accuracy vs. unweighted average recall), and last but not least in possible variations of results due to non-determinism in certain GPU operations and to random parameter initialization. With our

[6] The challenge name ComParE was established in 2013, after the 'Paralinguistics Challenge', the 'Speaker State Challenge' and the 'Speaker Trait Challenge' in the years before. `http://www.compare.openaudio.eu` [Accessed March 08, 2021]

experiments, we followed established practices from the literature, such as speaker-independent cross validation and reporting the mean of several runs of experiments in terms of unweighted average recall. Still, interpretation of results and comparison to related studies should be done cautiously because of these many possible factors of variation. Generally, when reporting experimental outcomes by means of comparison, there are different possible comparisons to bring up: results can be better than a) chance level, b) some challenge baseline (if existing), c) related works, or d) own previous experiments/baselines. Comparison with others in terms of state-of-the-art results is frequently done, but can be difficult and/or misleading due to the aforementioned reasons. Throughout this thesis, we confine comparisons to our own baselines in order to report improvements of certain methods and model extensions.

2.3 Acoustic Features for Emotion Recognition

From the search for optimal acoustic features for SER (which has been an ongoing endeavor for more than two decades by now), many approaches arose. In the early days of SER, small sets of acoustic variables consisting of mainly *prosodic features* had been used. Among others, these often included the fundamental frequency (F0) (Dellaert et al., 1996), energy (Schuller et al., 2003), temporal measures (speech rate and pausing), and formants (Petrushin, 1999). Since then, a variety of acoustic feature types and modeling approaches (dynamic vs static) emerged. In the following we briefly present the different types of acoustic features that have been utilized for our experiments.

Spectral features. Despite the fact that spectral characteristics of speech strongly depend on the phonetic content of a speech utterance (and are therefore desirable features for ASR), they have also been shown to contain useful information to discriminate emotional states (Wang and Guan, 2004; Koolagudi and Rao, 2012). Fayek et al. (2016) have demonstrated the transferability of filter bank features between the two tasks of ASR and SER. *Mel frequency cepstral coefficients (MFCCs)* are a well known feature type in speech processing that is widely used for ASR and music information retrieval, among others. With the advent of deep learning, a closely related representation, log Mel filter banks (lMFBs), has become another widely used feature type.

The procedure to obtain MFCCs involves the computation of Mel filter banks (MFBs), which are highly correlated due to the overlapping filters. Such correlated input poses a problem to traditional machine learning algorithms, such as Gaussian Mixture Models - Hidden Markov Models (GMMs-HMMs). Therefore, to decorrelate the features, a discrete cosine transform (DCT) can be applied to the filter banks, resulting in MFCCs. Because neural networks are less susceptible to highly correlated input (and can potentially make use of additional information in the MFBs), the DCT as an extra transformation step appears not to be necessary when employing deep learning for speech processing tasks (Hinton et al., 2012; Mohamed, 2014, Chapter 4). Log Mel filter banks were used as features throughout all investigations in this thesis with the exception of the work presented in section 4.2. MFCCs were used for experiments described in chapter 3 to compare the results between lMFBs and MFCCs.

Low-level descriptors and funcationals. Besides these 'general-purpose' features, there are also affect-specific acoustic feature set, such as the Geneva minimalistic acoustic parameter set (GeMAPS) and its extended version eGeMAPS (Eyben et al., 2016), or the 'emobase2010' reference feature set, which is based on the Interspeech 2010 Paralinguistic Challenge feature set (Schuller et al., 2010a). These feature sets are based on the principle to first extract *low-level descriptors* (LLDs) from the speech signal (e.g. the fundamental frequency contour), and then apply *functionals* to them (such as the arithmetic mean, standard deviation and certain percentiles) to arrive at a more high-level representation of the input signal.

The emobase2010 reference feature set for SER is a large set of 1,582 features. When it was proposed within the openSMILE toolkit[7] (Eyben et al., 2010d, 2013), the so-called feature brute-forcing was widely used in SER, that is extracting hundreds or even thousands of features and let the machine learning algorithm figure out which are useful (Schuller et al., 2008; Eyben et al., 2010a). The 1,582 features in the emobase2010 set result from a set of 34 LLDs and their delta coefficients, to which 21 functionals are applied to (1,428 features). Additionally, 19 functionals are applied to four pitch-based LLDs and their deltas (152 features); finally, two features (number of pitch onsets and total duration) are appended.

GeMAPS represents quite the opposite of feature brute-forcing, as it resulted from a joint effort of several research groups to create a minimalistic set of voice parameters for affective computing and voice research. In this thesis, the extended version, eGeMAPS, is used. It consists of 25 LLDs (containing frequency-related, energy-related and spectral parameters, such as pitch, jitter, loudness, shimmer, frequency and relative energy of formants), to which the arithmetic mean and the coefficient of variation are applied as functionals. To loudness and pitch, additional eight functionals are applied. Finally, specific features are computed in addition on only voiced or only unvoiced segments, and certain temporal features are added. This results in 88 parameters total. For the experiments in this thesis, we used the 25 LLDs from eGeMAPS without functionals as time-

[7] `https://audeering.github.io/opensmile/about.html` [Accessed March 22, 2021]

preserving input (chapter 3) as well as the complete set of 88 utterance-level features in chapters 4 and 5.

Data-learned and model-based representations. In contrast to extracting specific 'hand-crafted' features, another recently emerging approach is to automatically *learn* appropriate features from data. *Representation learning* is essentially concerned with learning a useful abstract representation for a given task form the raw input signal. In section 4.1 we present work on the combination of representation learning and traditional data pre-processing.

Yet another related approach are model-based representations, that is using a pre-trained ML model to generate abstract representations of the input. An example for this, used in chapter 5 of this thesis, are 'deep spectrum features' (Amiriparian et al., 2017b). They are generated by first extracting spectrograms from the speech signal and then feedings these into a pre-trained deep neural network for image classification. The activation values at a certain hidden layer of the network are then taken as feature vector to represent the speech signal.

2.4 Machine Learning Background: Neural Networks

This section provides the technical background on machine learning and introduces methods and concepts that are used in the experiments presented in this thesis. Note, that this is not meant to be an exhaustive introduction to machine learning and we assume that the reader is familiar with basic concepts of ML and neural networks. The reader might refer to (Jurafsky and Martin, 2020, chapters 5,7) for an introduction to logistic regression and a concise overview of neural networks for language and speech processing, to (Goldberg, 2017) for more advanced ML methods for NLP, or to (Goodfellow et al., 2016) for an extensive textbook on deep learning in general. In the following, we present the necessary background on artificial neural networks (ANNs), and then introduce three specific types of neural networks that were used in the experiments presented in this thesis: convolutional neural networks (CNNs), autoencoders (AEs), and generative adversarial networks (GANs).

2.4.1 Artifical Neural Networks

In a typical *supervised* ML problem, given training samples x_i and the respective labels y_i, the objective is to learn the function $\hat{y} = f(x; \Theta)$ by learning appropriate values for the parameters Θ, so that the difference between predicted labels $\hat{y}_i$ and real labels

y_i becomes minimal. This difference is defined by a so-called cost function (also known as loss function) and the goal of training a neural network is to reduce this cost (of misclassified samples). The simplest case of such a function f is *linear regression*, which can be defined as:

$$\hat{y} = w \cdot x + b \tag{2.1}$$

where x denotes a feature vector of the input sample, w a weight vector, and b a bias term; the learnable parameters Θ in this case comprise w and b. Although it is a powerful machine learning method, the *linear* property is a strong limitation, which is illustrated by the famous example of the logical XOR expression, which cannot be modeled with a linear function. In order to overcome this problem, an *activation function* $g(x)$ can be applied to the input to perform a *non-linear* transformation. One possibility for such a non-linearity is the sigmoid function, which has a return value between 0 and 1:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

Applying the sigmoid function brings us from linear regression to *logistic regression*, which can be used for binary classification:

$$\hat{y} = \sigma(w \cdot x + b) \tag{2.3}$$

Now, to move from logistic regression to neural networks, we can view logistic regression as one single neuron in a network;

thus, each neuron computes an activation $g(x)$ given certain inputs to it and outputs the result to other neurons. ANNs are loosely inspired by biological neural networks in the brain: over their connections the neurons 'transmit a signal' to other neurons (like synapses do in the brain), and the learnable weights for these connections affect how strong the transmitted signal is. ANNs consists of one or more *hidden layers*, denoted as h_j. Each hidden layer consists of multiple neurons (or units). Mathmatically, a two-layer feed-forward neural network can be expressed in a vectorized notation, where W represents a weight matrix, as follows:

$$h_1 = g_1(W_1 x + b_1) \tag{2.4}$$

$$h_2 = g_2(W_2 h_1 + b_2) \tag{2.5}$$

$$\hat{y} = W_3 h_2 \tag{2.6}$$

Typically, all neurons in one hidden layer apply the same activation function $g(x)$ (to enable this computationally efficient vectorization), and that is usually also kept consistent for all hidden layers in a network. Essentially, a neural network is a chain of matrix-vector multiplications, or as Ronan Collobert put it in one of his lectures:

> Deep learning is just a buzzword for neural nets, and neural nets are just a stack of matrix-vector multiplications, interleaved with some non-linearities. No magic there. (Collobert, 2011)

As already mentioned above, the power of NNs comes from the non-linear activation functions, for which the sigmoid function is only one example. There are others, such as hyperbolic tangent (tanh) or the rectified linear unit (ReLU), which is frequently employed.

While the hidden layers in a NN apply non-linear feature transformations of their inputs, the last layer, referred to as output layer, makes the actual class predictions. For this, typically the softmax function is used to normalize the output to a probability distribution over the predicted classes. The softmax function (equation 2.7) is a generalization of the sigmoid function for multiple dimensions. The output for each class label y_j is a value between 0 and 1, and these probabilities for all classes sum to 1.

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_j e^{y_j}} \tag{2.7}$$

Training and optimization How a neural networks 'learns' to fulfill its task is usually referred to as *training* and is an iterative procedure. It consists of a forward pass and a backward pass through the network and is repeated either for a certain number of iterations (or epochs) or until a predefined stopping criterion is reached. The forward pass is simply the calculation of the network's output given a training sample as input, as shown in equations 2.4-2.6. Then the predicted output is compared to the original label of the sample (by means of a loss function) and in the backward pass the networks weights are updated to improve its predictions. This is done with an algorithm called *backpropagation*, which is the layer-wise computation of the gradient of the loss function with respect to the network weights.

The procedure of finding optimal parameters, i.e. network weights, is called *optimization* and is solved with gradient-based optimization algorithms, such as gradient descent. The basic principle of gradient descent and its variants is to update the parameters in the direction of the negative gradient of the loss function in order to find the global minimum of this loss surface and therefore minimize the prediction error on a given training dataset. These parameter updates are normally done on small batches (so called minibatches) of the training data at once, in order to reduce the computational cost of calculating the loss function over the entire dataset. The *learning rate* of gradient descent determines the magnitude of the updates – the smaller the learning rate, the longer it takes for the algorithm to converge to the minimum; but if the learning rate is too large, it can cause gradient descent to 'overshoot' and might even increase the loss and cause instability. There are optimization algorithms that are based on adaptive learning rates, for example the *Adam* algorithm (Kingma and Ba, 2015), which was used for neural network training throughout the experiments presented in this thesis.

One problem that ML algorithms often face due to their large number of parameters is *overfitting*, that is the effect that a model learns to perfectly predict the training data at the cost of poor *generalization*. Consequently, the model is 'overfitted' to the training data and performs poorly on unseen test samples. To prevent this from happening, *regularization* needs to be applied. A traditional approach for this is to add a so called regularization term R to the loss function, which puts certain constraints on the learned parameters. One commonly used example of this is the squared L_2 norm of the weight matrix, $R_{L_2} = ||W||_2^2 = \sum(W_{i,j})^2$. Adding this to the loss function regularizes the parameters in W towards zero (with the effect that unnecessary connections in the neural network might even become canceled out effectively). Another widely used approach to regularization, which was also applied in the models described in this thesis, is *dropout* (Srivastava et al., 2014). With this method, a certain number of neurons is 'droppped out', i.e. set to zero, during the forward and backward pass at random. More technically, at each training iteration, individual neurons are either dropped out with probability p or kept unchanged with probability $1 - p$, where p is referred to as dropout rate (or sometimes the terms are reversed and p is called keep rate). The higher the dropout rate, the more constrained or regularized the network will be. With dropout, a neural network learns more robust and generalizable feature representations that do not rely too much on individual neurons.

2.4.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are a special kind of NN that use an operation called *convolution* instead of general matrix multiplication in at least one layer. CNNs operate on multidimensional input and are well known for their successful applications in computer vision, but are also widely used within NLP and speech processing (cf. section 3.1).

The convolution operation in a CNN is sometimes called a 'sliding dot product' because it involves taking the dot product between the input I and the kernel K, while the kernel is sliding through the input. For a two-dimensional convolution, this operation can be formally expressed as in Goodfellow et al. (2016), chapter 9:

$$S(i,j) = (K * I)(i,j) = \sum_{m} \sum_{n} I(i+m, j+n) \cdot K(m,n) \quad (2.8)$$

where S denotes the output (feature map), i and j are iterators over the input matrix I, and m and n iterate over the 2D-kernel K. Note, that equation 2.8 strictly represents a function called *cross-correlation*, which is essentially convolution with a reversed kernel. For more details on the relation between convolution and cross-correlation and on the differences in notation, see Goodfellow et al. (2016), chapter 9. We use the formal representation in equation 2.8 because we find it is most intuitive to understand (in conjunction with the example in Figure 2.4). In the machine learning context, the cross-correlation function is ususally also called convolution, and the fact that the kernel is 'flipped' is not important for practical purposes (because the kernel weights are learned during training anyway).

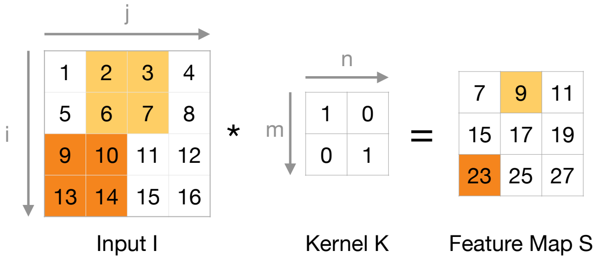


Figure 2.4: Illustration of the convolution operation in CNNs.

Figure 2.4 illustrates the convolution with an example; the colored areas in the input correspond to the convolved outputs in the resulting feature map with the same color. The distance with which the kernels are moved is called *stride* (set to 1 in the example). Note, that one convolutional layer in a CNN contains not only one, but many kernels, which learn to represent different aspects of the input.

Due to their design, CNNs have a number of useful and advantageous properties, especially that they are able to learn patterns in the input independent of their particular position, scale and possible rotation.

2.4.3 Autoencoders

Autoencoders (AEs) are a type of neural networks that is suitable for *unsupervised* representation learning (RL). AEs are trained to reconstruct (or attempt to copy) their input, given certain constraints. A basic autoencoder consists of two parts, the encoder and the decoder. The encoder is trained to learn a mapping function $h = f(x)$, where x is the input and h denotes a *code* that is used to represent the input. Then, the decoder produces the reconstruction $r = g(h)$ from this code. The overall objective is to minimize the loss function

$$L(x, g(f(x))), \tag{2.9}$$

which penalizes $g(f(x))$ from being dissimilar from x. This could for example be the mean squared error (MSE) between the two terms.

While simply copying the input to the output seems useless, the *constraints* that are imposed to the AE are what makes it useful. Normally, we are not interested in the output $r = g(h)$ itself, but in the code h. One classical variant is the *undercomplete* autoencoder, which was also used for the experiments presented in chapter 4. Here, the constraint is that h has a (much) smaller dimension than the input x (also called *bottleneck* layer). By that, the model is forced to capture the most salient features of the input, and therefore to learn a compact representation. This learned representation can then be utilized for other downstream tasks.

2.4.4 Generative Adversarial Networks

Generative adversarial networks (GANs), introduced by Goodfellow et al. (2014), are not one specific type of NN, but rather describe a class of *adversarial* machine learning frameworks, which can be implemented in many ways with a variety of NNs. The basic principle of GANs is that two neural networks play a minimax game, where one network's gain is the other one's loss. The idea of adversarial networks is not new (Schmidhuber, 2020), but arguably has become very popular and sucessful with the introduction of GANs and their many variants in recent years.

GANs represent a different type of modeling approach from what we have looked at so far; as their name suggests, they belong to the class of *generative* models, as opposed to *discriminative* models. While discriminative models are designed to learn the conditional probability $p(y|x)$, i.e. what is the probability of label y given features x, generative models learn the joint probability $p(x, y)$ (cf. Ng and Jordan (2002)). Hence, generative models can not only be used for classification (by using Bayes' theorem), but also for generating samples, i.e. draw samples from this learned distribution.

The intuition behind GANs is the following: a generator network G samples from some random distribution and attempts to output (generate) samples that are indistinguhisable from a certain target data distribution. On the other hand, the discriminator network D is trained to tell apart the real samples from the ones generated by G. Both networks are trained jointly. An illustrative metaphor, which is frequently used to explain this principle (Marr, 2019), is that G is a blind art forger, who does not know how real master pieces look like, but still wants to produce paintings that look similar – D is a detective who judges whether G's paintings are real or fake, and all what G knows is the detective's assessment. Iteratively, by producing thousands of paintings and receiving D's judgement, the art forger learns to fool the detective, who in turn also becomes better and better at telling apart real from fake paintings.

The technical details of the GAN variant that we have used for experiments in chapter 4, such as the loss functions and training procedure, are provided in section 4.2.2.

3 Speech Emotion Recognition with Convolutional Neural Networks

> Thus, I may speak of "recognizing emotions" but this should be interpreted as "measuring observations of motor system behavior that correspond with high probability to an underlying emotion or combination of emotions."
>
> Rosalind Picard, *Affective Computing*, Technical Report, 1995

In this chapter we present an attentive convolutional neural network (ACNN) with multi-task learning objective function, proposed in Neumann and Vu (2017). This model builds the foundation for several experiments presented in this thesis.

First, in section 3.1 the most relevant related work is summarized, in whose context our approach and the experiments came into being. Section 3.2 contains the details of the ACNN model, including its attention mechanism and the multi-task learning approach. The datasets for model training and evaluation, including preprocessing and acoustic feature extraction, are described in section 3.3.

The first series of experiments, presented in section 3.4, is concerned with three different aspects of the input data to the network. We compare and analyze the classification accuracy using: (a) different lengths of the input signal, (b) different types of acoustic features and (c) different types of emotional speech (improvised play vs. scripted conversations, cf. Figure 2.2 on page 33). The experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database show that the recognition performance strongly depends on the type of speech data, independent of the choice of input features. Further, emotion recognition appears to be possible with only a short snippet of an utterance (e.g. 2 seconds) because the accuracy decreases only slightly compared to the full input length.

The second series of experiments, presented in section 3.5, is concerned with cross- and multilingual SER. We investigated cross-lingual and multilingual experimental setups, as a step towards language-independent emotion recognition in natural speech. Experiments on English and French speech data with similar characteristics in terms of interaction (human-human conversations) are presented. Besides pure cross-lingual (train on language A, apply to language B) and multilingual (train and test on the union of both corpora) settings, we explored the possibility of fine-tuning a pre-trained cross-lingual model with a small number of samples from the target language, a relevant scenario for low-resource languages. The results show that multilingual as well as fine-tuned cross-lingual models yield reasonable results for arousal prediction, whereas valence prediction appears to be more sensitive to language differences. An analysis of the learned attention weights in the ACNN's attention layer shows that the highest attention lays mostly on the beginning of a speech signal.

3.1 Related work

3.1.1 Neural Networks for Speech Emotion Recognition

As for many tasks in speech and natural language processing, neural networks have become the state-of-the-art method for SER in the research community during the last decade. First work on recurrent neural networks for SER was published already in 2008 (Wöllmer et al., 2008), but looking at the numbers of publications and the growing variety of research directions related to neural networks and SER, these various approaches really started to gain traction around 2013. At that time, neural networks have been shown to significantly boost emotion recognition performance, including Deep Belief Networks (DBN), hybrid deep neural network–hidden Markov model (DNN-HMM) frameworks, and deep autoencoders (Cibau et al., 2013; Li et al., 2013; Huang et al., 2014; Han et al., 2014; Xia and Liu, 2015).

Convolutional neural networks (CNN) proposed by Waibel et al. (1989) and Le Cun et al. (1990) are a special kind of neural networks that have been successfully used not only for computer vision but also for speech processing (Abdel-Hamid et al., 2012; Sainath et al., 2013, 2015) and NLP (Collobert et al., 2011; Kim, 2014; Kalchbrenner et al., 2014). For speech recognition, CNNs proved to be effective in modeling correlations in time and frequency in spectral representations (Sainath et al., 2015) and robust against noise compared to other neural network models (Palaz et al., 2015). Furthermore, Sainath and Parada (2015) showed that CNNs are suitable for small memory footprint keyword spotting due to the parameter sharing mechanism. Therefore, because of a smaller number of parameters, CNN models can be trained with a smaller amount of training data than conventional DNN models.

First work on speech emotion recognition with CNNs was presented in 2014. Mao et al. (2014) proposed feature learning using a sparse autoencoder to learn CNN kernels in an unsupervised fashion, which are then applied on spectrograms to extract local invariant features. In the following years, the number of publications about CNNs for SER increased steadily (cf. for example Zheng et al. (2015); Anand and Verma (2015); Xue et al. (2015); Bertero et al. (2016); Fayek et al. (2016)). Several studies (Keren and Schuller, 2016; Trigeorgis et al., 2016; Lim et al., 2016) presented CNNs in combination with Long Short-Term Memory models (LSTM) to improve speech emotion recognition based on log Mel filter banks (lMFBs) or raw audio signal. Trigeorgis et al. (2016) demonstrated an end-to-end training from raw signal, dropping the step of manual feature extraction completely.

Attention mechanisms were first employed primarily in recurrent neural networks that had been successfully applied to a wide range of tasks such as handwriting generation (Graves, 2013), machine translation (Bahdanau et al., 2015), image caption generation (Xu et al., 2015) and speech recognition (Chorowski et al., 2015). Inspired by this, attention mechanisms had also been proposed for CNNs in NLP tasks (Adel and Schütze, 2017; Meng et al., 2015; Yin et al., 2016). These mechanisms are especially helpful when the input signal is rather long or complex. To combine the strengths of CNNs with an attention mechanism, we proposed the attentive convolutional neural network for speech emotion recognition.

3.1.2 Cross-corpus and Cross-lingual Emotion Recognition

A common approach to automatic emotion recognition is to train and test a classifier on one annotated (mostly mono-lingual) corpus, either by subdividing the data into train, validation and test sets or by means of cross validation. This way, the system is highly specialized with respect to multiple factors, such as the speaker group, the recording situation, the language, and the type of speech (spontaneous or acted). Further, no conclusions can be drawn to what extend such a system can generalize across different interaction scenarios and languages. One relatively simple method to assess this generalization ability is to evaluate a machine learning model on a separate dataset, which has different properties than the training data. One case of such cross-corpus evaluation, which we investigate in the last part of this chapter, is the *cross-lingual* setting, meaning that the datasets differ in language.

Various cross-corpus analyses have been conducted with regard to SER (Schuller et al., 2010b; Lefter et al., 2010; Eyben et al., 2010b; Polzehl et al., 2010; Schuller et al., 2011d,c). It has been shown that cross-corpus classification is feasible in general, but leads to performance drops in many cases compared to within-corpus evaluation (Lefter et al., 2010; Jeon et al., 2013). Eyben et al. (2010b) pointed out that for cross-corpus training it is crucial to find out which corpora are generic enough to be included and which ones are too specific. In an extensive study with six corpora Schuller et al. (2010b) examined many different combinations of corpora as training set, emphasizing the variety in conditions and therefore the transferability of conclusions to real-world applications. Closely related to that, Feraru et al. (2015) presented a comprehensive overview using data from eight languages and showed that cross-lingual emotion recognition is feasible, but with notably lower accuracy than mono-lingual recognition, especially for valence prediction. These studies give an overall impression on the performance of cross-corpus emotion recognition, however, they make the interpretation of results difficult because of these many factors of variation in the data. In contrast, the herein presented experiments focus on a more narrow comparison between two corpora that differ in language but match the interaction type (human-human) and are close to each other in the degree of natural, spontaneous speech.

A closely related task is *multilingual* emotion recognition, that is merging data from different languages into one dataset for training a model. It has been shown previously that the performance of a multilingual system is inferior compared to monolingual models (Hozjan and Kačič, 2003). In this early investigation, authors used the InterFace datasets (Hozjan et al., 2002), which contain data from four languages. Since these datasets consist of acted speech from only two to three speakers per language, the general validity of these findings is questionable. Kim et al. (2017a) combined seven emotional speech datasets; as mentioned before, this means that there are many factors of variation, e.g. acted vs. spontaneous speech, different annotation schemes, or different types of interaction (human-computer vs. human-human). There are also approaches to multilingual emotion classification that involve multiple models. For example, Sagha et al. (2016) presented a two-step procedure of first identifying the language and then selecting a language-specific model. In contrast to this, we examined the performance of *one* model trained on multiple languages.

3.2 Method

The attentive convolutional neural network is depicted in Figure 3.1 on the next page. It consists of two main parts: a CNN with one convolutional and one pooling layer, and an attention layer. The CNN learns a representation of the audio signal, while the attention layer computes the weighted sum of all the information extracted from different parts of the input. The output from the pooling layer and the attention vector are then fed into a fully connected softmax layer.

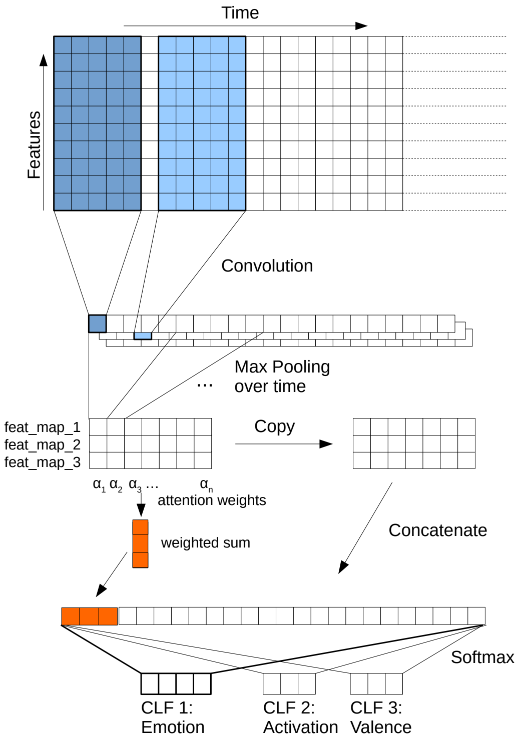


Figure 3.1: ACNN for speech emotion recognition, Classifier (CLF) 1 predicts emotional categories, CLF 2 and 3 predict arousal and valence categories (multi-task learning).

Convolutional Neural Network

To form the input matrix for the CNN, the audio signals are divided into s overlapping segments, each represented by a d-dimensional feature vector. Thus, each utterance is represented as an input matrix $I \in R^{d \times s}$. The number of segments s depends on the signal length as well as on the window size and overlap between frames. The particular details also vary for the implemented feature sets and are found in section 3.3. For the convolution operation (equation 3.1), 2D kernels K (with width $|K|$), spanning all d features, are applied. As a result, the output of the convolutional layer is a vector that represents features on the temoporal dimension. The reason for this 'one-dimensional' approach is that we use different types of pre-processed features (such as MFCCs and the LLDs of the eGeMAPS feature set) for which the 'spatial' relations in the feature dimension are not necessarily clear. Hence, the intuition is that more meaningful information can be extraced when looking at all features within a certain time window.

$$(K * I)(i, j) = \sum_{m=0}^{d-1} \sum_{n=0}^{|K|-1} I(i+m, j+n) \cdot K(m, n) \qquad (3.1)$$

After the convolution, a max pooling layer is applied to further reduce the output the most salient features. Then, all feature maps are concatenated to form one final feature vector.

Attention Mechanism

The attention layer in the ACNN model is based on the attention mechanism introduced in Bahdanau et al. (2015) and inspired by the work presented in Adel and Schütze (2017).

As depicted in Figure 3.1 on page 63, the output of the max pooling layer is formed by stacking all feature maps (the number of feature maps equals the number of kernels that are employed in the convolutional layer). We now view each column of this matrix as the feature vector x_i at time step i. For each vector x_i in this sequence x, the so-called attention weights α_i can be computed with equation (3.2), where $f(x)$ is the scoring function.

$$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_j))} \tag{3.2}$$

In this work, $f(x)$ is the linear function $f(x) = W^T x$, where W is a weight matrix, i.e. a parameter that is learned during model training. Following the terminology in Adel and Schütze (2017), the feature map matrix x in this case is both the *focus* and *source* of the attention mechanism, i.e. the attention weights are learned based on the information in the feature maps and are then applied to them. The output of the attention layer, *attentive_x*, is the weighted sum of the input sequence (equation (3.3)).

$$attentive\_x = \sum_i \alpha_i x_i \tag{3.3}$$

The intuitions behind using this attention mechanism for emotion recognition are two-fold: a) speech emotion recognition is related to sentence classification with emotional content being differently distributed over the signal and b) the emotion of the whole signal is a composition of emotions from different parts of the signal. Therefore, attention mechanisms are suitable to first weight the information extracted from different pieces of the input and then combine them in a weighted sum. However, because the input signal is noisy with regard to expressed emotions and can be very long, a max pooling layer is still helpful to only select the most salient features and filter noise. Therefore, we combine the CNN output vector after max pooling with the attention vector for the final softmax layer.

Multi-task Learning

As introduced in section 2.1, two types of emotion representation are predominantly used for SER, namely categorical labels (e.g. *anger, happiness*) and continuous labels in the 2-dimensional arousal/valence space. In Xia and Liu (2015), it was shown that multi-task learning (MTL) with both categorical and continuous labels for training can improve prediction results. Similar to this approach, we train the model with categorical labels as primary (target) classes, and use arousal and valence labels as secondary, auxiliary target.[1] Each training sample s is represented as $[x_s, (y_{e,s}, y_{a,s}; y_{v,s})]$, where x_s is the feature representation, and $y_{e,s}, y_{a,s}, y_{v,s}$ are the associated category, arousal and valence labels. Equation 3.4 shows the overall objective function of the model, where N is the number of samples and $h_{l,s}$ the output of the last hidden layer for sample s.

$$\begin{aligned} J = \frac{1}{N}(1-\alpha-\beta) \cdot \sum_{i}^{N} -log(P(y_{e,s}|h_{l,s})) \\ +\alpha \cdot \sum_{s}^{N} -log(P(y_{a,s}|h_{l,s})) + \beta \cdot \sum_{s}^{N} -log(P(y_{v,s}|h_{l,s})) \end{aligned} \tag{3.4}$$

Two parameters α and β are used to control the relative impact of the auxiliary arousal and valence labels separately.

3.3 Data and Acoustic Features

Datasets The first series of experiments (feature sets and signal length analysis) was conducted and evaluated on the IEMOCAP database (cf. section 2.2 for details). In order to be com-

[1] In Neumann and Vu (2017) we used the term *multi-view learning* instead of *multi-task learning*. Both can be somehow misleading in this context, because it is about mulitple emotion representations (labels) of the same data. So, while being essentially the same task of emotion classification, the term multi-task learning describes it correctly, because two different sets of labels are learned from a shared data representation.

parable to related work and because some of the ten emotion classes in IEMOCAP contain only very few samples, the same four classes as in Xia and Liu (2015); Jin et al. (2015); Ghosh et al. (2016b); Rozgic et al. (2012); Gideon et al. (2017) are used: *anger, happiness, sadness*, and *neutral state*, where *happiness* and *excitement* were merged into one class. The dataset contains two types of speech: *improvised* and *scripted.* To investigate performance differences between these two, we report results for three subsets from the data: only *improvised* (2,943 utterances), only *scripted* (2,588 utterances), and *all* sessions (5,531 utterances). Table 3.1 presents the class label distribution of each subset.

	Anger	Happiness	Neutral	Sadness
improvised	289	947	1,099	608
sctipted	814	689	609	476
$\sum$	1,103	1,636	1,708	1,084

Table 3.1: Label distribution of IEMOCAP and its subsets.

One restriction of using a CNN is that the sample length has to be equal for all samples. For this reason, we set the sample length to 7.5s, which corresponds approximately to the mean utterance duration plus standard deviation.[2] Longer turns are cut at 7.5s and shorter ones are padded with zeros at the end. Using this threshold, most speech material is included

[2] The mean length of all turns is 4.46s (max.: 34.1s, min.: 0.6s)

without having too much zero padding (which would slow down the training considerably).

As auxilliary labels for the MTL approach arousal and valence labels are used. They are represented as Likert-scale scores ranging from 1 to 5, and because the mean across annotators is taken, real numbers are possible. To avoid having too many classes, we group them into three categories each by applying the same range mapping as in Metallinou et al. (2012): low: [1,2]; medium: (2,4); high: [4,5]. The class distribution for arousal is: 13.1% low, 68.1% medium, and 18.8% high. The distribution for valence is: 28.3% low (negative), 47.2% medium (neutral), and 24.5% high (positive).

For the second series of experiments (cross- and multilingual), the RECOLA corpus (cf. section 2.2 for details) is used as second dataset in addition to IEMOCAP. The main criterion for selecting the data was that both corpora contain the same type of speech in terms of conversation type (human-human) and naturalness.

RECOLA is annotated with continuous labels for arousal and valence in the range [-1, 1] on a 40ms rate. Since we are interested in recognition of emotions on utterance level, we calculated the mean of all values for one turn, and then took the average across all annotators as the final label.

To be able to train a model on several corpora, the different annotation schemes have to be mapped to a unified form. We de-

	Arousal		Valence	
	Low	High	Negative	Positive
IEMOCAP	3,121	6,918	5,356	4,683
RECOLA	520	788	241	1,067
$\sum$	3,641	7,706	5,597	5,750

Table 3.2: Distribution of binary arousal and valence labels in IEMOCAP and RECOLA.

cide to focus on a binary classification task of arousal (low/high) and valence (negative/positive). The mapping of original annotations to a binary scheme is as follows: IEMOCAP annotations in the range [1, 2.5] are mapped to low/negative and scores within (2.5, 5] to high/positive; for RECOLA the respective ranges are [-1, 0] and (0, 1]. The resulting distribution of these binary labels is shown in Table 3.2.

Acoustic Features In order to analyse and compare the emotion recogntion performance of a CNN with different feature sets, we assessed the ACNN's performance using the following four feature sets: (a) 26 lMFBs, (b) 13 MFCCs, (c) a small prosody feature set, and (d) eGeMAPS: the extended Geneva minimalistic acoustic parameter set. For all feature sets standardization (also known as z-score normalization) was applied for each speaker independently.

The openSMILE toolkit (Eyben et al., 2013) was used to extract all features. For filter banks, MFCCs, and prosody features, the audio signal is segmented into 25ms long frames with a 10ms shift. To extract filter banks, a Hamming window is applied and the fast Fourier transform (FFT) with 512 points is computed. Then, the logarithmic power of 26 Mel-frequency filter banks over a range from 0 to 6.5kHz is computed. Finally, to retrieve MFCCs from these lMFBs, a discrete cosine transform is applied and the first 13 MFCCs are kept as features.

The prosody feature set is motivated by the question, how well SER can be performed with the proposed model using a minimalistic set of only seven prosody-related features (compared to the other, more informative features). We extracted the following features: PCM loudness, envelope of F0 contour, voicing probability, smoothed F0 contour, local (frame-to-frame) jitter, differential jitter, and local shimmer. This choice was based on the Interspeech 2010 Paralinguistic Challenge feature set (Schuller et al., 2010a), from which we selected only the prosody-related LLDs and removed all other feature types (e.g. MFCCs) and functionals.

The eGeMAPS feature set is an 88-dimensional utterance-level representation (cf. section 2.3 for details) As input the the ACNN model, we did not apply functionals, but extracted only the 25 frame-by-frame LLDs, which are the basis for eGeMAPS.

3.4 Analysis of Input Features, Signal Length, and Speech Type

3.4.1 Experimental Setup

The IEMOCAP dataset consists of five sessions with one male and one female speaker each. To train the models in a speaker-independent manner, we use leave-one-session-out cross validation. Data from 8 speakers is taken to construct training and development sets and the remaining two speakers' data constitute the test set.

To investigate the impact of signal length on the performance, models were trained and tested with decreasing utterance length by cutting the speech signals at 7, 6, 5, 4, 3, 2, and 1 seconds, respectively.

The CNN models were implemented with the Theano library (Bergstra et al., 2010; Bastien et al., 2012). For optimization we used stochastic gradient descent with an adaptive learning rate, known as Adam (Kingma and Ba, 2015). For regularization dropout is applied to the last hidden layer (Srivastava et al., 2014). The system's hyper-parameters were tuned in a grid search, assessing the accuracy on the validation set. The final hyper-parameters are the following: 100 kernels of width 5 (spanning 5 input frames) and 100 kernels of width 10, resulting in a total of 200 feature maps; a mini-batch size of 30 for lMFBs, prosody and eGeMAPS, and 50 for MFCC features; a dropout rate of 0.8; a pool size of 30, and stride of the convolution operation of 3 for all configurations. The reason for the seemingly large size of the pooling window for the max-pooling layer is the large amount of overlap in the feature maps, especially with a kernel width of 10 and stride of 3. Multi-task learning can be switched off by setting $\alpha = \beta = 0.0$ (i.e. arousal and valence information is not considered) without changing anything in the model structure. We will refer to this setting as single-task learning (STL) in the remainder of this chapter. For MTL we set $\alpha = \beta = 0.3$. All models were trained for 100 epochs, and we selected the model parameters and results from the epoch with the highest accuracy on the validation set. This essentially corresponds to early stopping (Morgan and Bourlard, 1989) with infinite patience, i.e. there is no specific stopping criterion, but training runs for the complete number of determined epochs. The main reason for this approach was a considerable amount of loss oszilation throughout the training procedure. However, while this can easily be done with small neural networks on small datasets, it is not advisable for larger problems where the absence of a reasonable stopping criterion potentially leads to a big increase in training time and use of computational resources.

3.4.2 Results

Differences between Scripted and Improvised Conversations and Comparison of Input Features

	CNN		Attentive CNN	
Features (dim.)	STL	MTL	STL	MTL
lMFB (26)	58.37 (61.71)	**61.27** (62.06)	60.20 (61.95)	58.76 (62.11)
MFCC (13)	58.22 (61.31)	**59.98** (61.35)	58.30 (60.85)	58.16 (61.35)
eGeMAPS (25)	58.27 (60.25)	58.96 (60.28)	59.53 (60.26)	**59.87** (61.27)
Prosody (7)	51.27 (56.34)	51.12 (56.33)	51.27 (57.11)	**51.33** (57.12)

Table 3.3: Results on improvised sessions; unweighted average recall (weighted accuracy in parentheses).

	CNN		Attentive CNN	
Features (dim.)	STL	MTL	STL	MTL
lMFB (26)	48.22 (51.07)	50.12 (51.64)	50.44 (52.64)	**50.93** (51.70)
MFCC (13)	50.78 (52.35)	52.21 (53.01)	**52.39** (53.19)	51.94 (52.72)
eGeMAPS (25)	50.70 (51.84)	50.81 (52.82)	50.10 (52.31)	**52.81** (53.19)
Prosody (7)	48.12 (49.17)	46.64 (48.76)	**48.36** (48.69)	46.80 (49.02)

Table 3.4: Results on scripted sessions; unweighted average recall (weighted accuracy in parentheses).

For all experiments, the results are presented as unweighted average recall (UAR) and as weighted accuracy (WA). WA represents the accuracy on the whole test set, i.e. the ratio of correctly predicted samples to the total number of samples. It is referred to as *weighted* accuracy because the class distribution in the data affects it, meaning that the accuracy is weighted by class size. For example, misclassifying all samples of a class

	CNN		Attentive CNN	
Features (dim.)	STL	MTL	STL	MTL
lMFB (26)	55.75 (55.38)	58.08 (55.92)	57.54 (54.86)	**58.98** (56.10)
MFCC (13)	57.12 (55.33)	**57.82** (55.74)	57.54 (55.12)	57.32 (55.40)
eGeMAPS (25)	56.52 (54.73)	56.09 (54.71)	**57.03** (54.93)	56.98 (54.78)
Prosody (7)	49.81 (48.90)	49.33 (48.79)	50.54 (48.99)	**50.73** (49.13)

Table 3.5: Results on the complete dataset; unweighted average recall (weighted accuracy in parentheses).

that has only few samples does have only little impact on the overall result. In contrast to that, UAR represents the average of the individual class recalls. Because the class recalls (number of correct samples within a class divided by total number of samples within that class) are computed first and then averaged, every class has the same impact on the result. Contrary to the publication in which these experiments were first presented (Neumann and Vu, 2017), we want to focus on UAR here, as this metric is known to be better suited when working with unbalanced data. For the sake of comparison with the original results and as basis for further discussion on performance metrics, both UAR and WA are presented in Tables 3.3-3.5.

Improvised speech: The results for the improvised subset of the data are shown in Table 3.3 on the preceding page. The best performance is reached with lMFBs. In terms of UAR, the CNN with MTL performs best with 61.27%; in terms of WA, the ACNN with MTL performs best with 62.11%.

Scripted conversations: The results for scripted conversations (Table 3.4 on page 74) are overall notably lower than for improvised speech. For this subset of the data, MFCC and eGeMAPS features lead to higher accuracies than lMFBs. However, all results are in a narrow range. In terms of UAR, the best result of 52.81% is achieved with the ACNN with MTL using eGeMAPS features; the highest WA (53.19%) is achieved with the ACNN (MFCC with STL and eGeMAPS with MTL).

Complete dataset: The results for the full IEMOCAP dataset (Table 3.5 on the previous page) lie inbetween those of the subsets, as one would expect. Using MFCC and lMFB features yields similar results, the accuracy with eGeMAPS is slightly lower, whereas prosody features perform notably worse. The best result is achieved with lMFB features using the ACNN with MTL (58.98% UAR, 56.10% WA).

General findings: The following summary of the results refers to the UAR results, as these are more meaningful in the context of unbalanced data. We will discuss differences between UAR and WA results further below. All results show that the prosody feature set yields notably lower results than cepstral features like lMFB and MFCC. We assume that the prosody feature set simply contains too little information (only seven features) to compete with the others. The performance differences between lMFB, MFCC and eGeMAPS are generally small. This suggests that a CNN model is able to learn high-level features equally well from these different input representations.

The proposed multi-task learning setup yields small improvements in certain constellations, for example with lMFB features on the complete dataset (Table 3.5), but at the same time it deteriorates the results in other cases. One potential reason for these results is that the secondary task of predicting arousal/valence levels is so closely relatded to the main task, that it cannot generate beneficial additional information in the learning process.

Similarly to MTL, the results with and without attention mechanism do not show significant differences. There are small improvements in some cases, in particular for the STL setup.

Looking at the results through the lens of the two performance metrics UAR and WA, the importance of considering a *suitable* metric becomes evident. Depending on which one is chosen, the results are viewed in a different light, particularly when we compare Tables 3.3-3.5 to each other. While for the complete dataset (Table 3.5), UAR is higher than WA in all configurations, the opposite is the case for the two subsets of improvised and scripted dialogs.[3] The reason for this is the strongly unbalanced class distribution in these subsets (shown in Table 3.1 on page 68).

[3] Average difference across all results per table (as $WA - UAR$): improvised: 2.92%, scripted: 1.42%, all: -1.77%

Overall, our proposed model performs better on improvised play than on the scripted conversations, independent of the feature choice. Above all, these findings show that SER can be very sensitive to the type of speech data. This is in line with findings by Tian et al. (2015), who concluded that 'the performance of features and models is largely influenced by the dialogue type and the size of the data set'. Hence, it is important to carefully select suitable training data for a particular application and – even more so – to develop machine learning models that are robust against varying kinds of speech.

Model Performance for Decreasing Signal Length

In this experiment, we used the ACNN model with MTL to perform emotion recognition on data of different sample length. lMFB and MFCC features were used because they provided the best results in the previous investigation. Now we attempt to answer the question how long a system should wait to make a prediction. In other words, is it possible to predict the emotional state of a speaker before she finished the utterance? This can be beneficial or even critical in real-time applications. The results are presented in Figure 3.2 on the next page.

Generally, the accuracy decreases with shorter sample length – as one would expect because less information is available. Comparing lMFB and MFCC features, there are no large deviations, only smaller variations. However, when comparing the

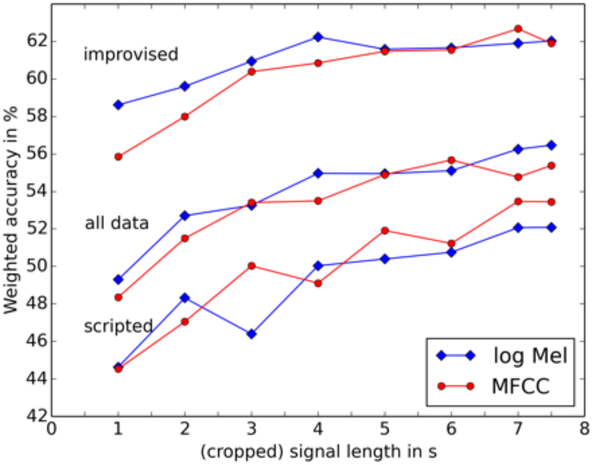


Figure 3.2: System performance with decreasing signal length.

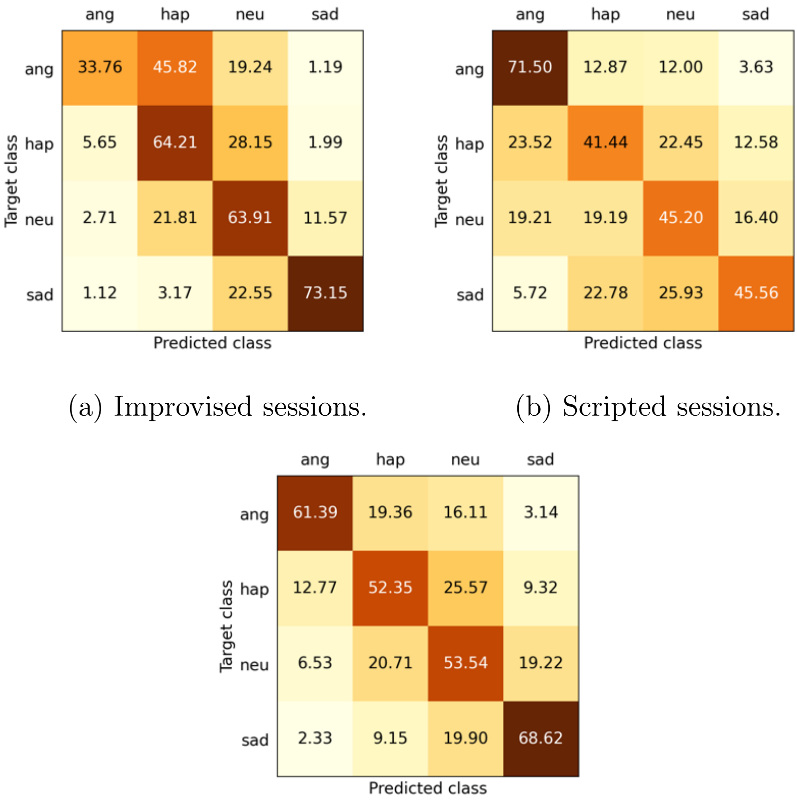
speech types (i.e. the data subsets), we observe a notable difference in the performance decline between *improvised* and *scripted* speech, especially with lMFBs. The absolute difference in accuracy between longest (7.5s) and shortest (1s) input is 3.4% on improvised and 7.5% on scripted data. These results suggest that the improvised speech utterances are more likely to carry emotional content in the first seconds already, compared to the scripted speech. In general, these are promising findings, showing that a relatively short snippet of a speech signal can be sufficient to perform emotion recognition with only a small loss in accuracy.

3.4.3 Error Analysis

In this section we take a closer look at the results in terms of error patterns by analyzing the predictions of the ACNN with multi-task learning and lMFBs as input. Figures 3.3a-3.3c show the corresponding confusion matrices; the results are averaged across the five cross validation folds.[4]

For *improvised* speech (Figure 3.3a) one considerable observation is that the model predicts *happiness* for 45.82% *anger* samples. First and foremost, this is due to the skewed class distribution in the improvised and scripted subsets of IEMOCAP (cf. Table 3.1); the class *anger* is strongly underrepresented in improvised sessions, which leads to a bias towards predicting the overrepresented classes *happiness* and *neutral state*. Furthermore, the seemingly counter-intuitive mistake of confusing *anger* and *happiness* becomes more plausible when considering the arousal-valence space. Both classes exhibit a high arousal level. Hence, the system's frequent confusion is also due to the fact that valence is harder to predict than arousal (Schuller et al., 2009a; Trigeorgis et al., 2016; Eyben et al., 2016). The class *sadness* is predicted best (73.15%). This observation is in line with findings in Busso et al. (2008). Further, the neutral class is frequently confused with other classes. This seems

[4] The numbers in the presented confusion matrices differ slightly from those in Neumann and Vu (2017) because some experiments were rerun for the present analysis. The overall findings and conclusions of this error analysis are not affected by this.



(a) Improvised sessions.

(b) Scripted sessions.

(c) All sessions.

Figure 3.3: Error distribution of the ACNN predictions (lMFB features, multi-task learning).

plausible because the neutral state is located in the center of the arousal-valence space, which makes the discrimination from other classes more difficult.

In contrast, for *scripted* sessions the recall for *anger* is notably higher (*anger* is the majority class in this case), and relatively low for *sadness* and *happiness.* In general, there are more errors in almost all classes. The main reason for the high discrepancy in the class *anger* is the different class distribution as described above. However, this does not entirely explain the differences in predictions of the other classes. The analysis suggests that the *improvised* utterances might generally be more variable and therefore make it easier to discriminate affective states. This conclusion is also supported to some extend by the results in Figure 3.2, which show that the performance loss is relatively small for shorter input signals on the improvised speech subset. Another observation worth noting is the high percentage of *sadness* samples predicted as *happiness* (22.78%). To find out the reason for this frequent confusion, further analysis is necessary.

The error distribution on the complete dataset (Figure 3.3c) lies between those seen in Figures 3.3a and 3.3b. There are similar patterns as for *improvised* data, but the confusion between *anger* and *happiness* is not as severe.

3.5 Multilingual and Cross-Lingual Speech Emotion Recognition

3.5.1 Experimental Setup

For the multi- and cross-lingual experiments the following four settings are investigated: (a) mono-lingual baseline models, (b) multilingual models (merge the datasets RECOLA and IEMOCAP for training), (c) cross-lingual models (trained on one corpus and tested on the other one), and (d) fine-tuning of a cross-lingual model on a small number of samples from the target dataset.

Mono-lingual and multilingual models were evaluated in a cross validation (CV) scheme because there are no predefined train and test splits for these datasets. The IEMOCAP data consists of five sessions with one male and one female speaker each. As in the previous experiments, data from four sessions is used to construct training and development sets and the remaining session is used for testing, resulting in 5-fold CV. For RECOLA, we manually construct five splits so that they are balanced with respect to number of speakers and gender. This way, we ensure speaker-independent training (in contrast to random sampling). This segmentation into five splits (matching the number of sessions in IEMOCAP) is done to facilitate multilingual training with cross validation.

The evaluation of cross-lingual training is straightforward: one dataset is taken entirely as training set and the respective other one as test set. For fine-tuning (FT) in a simulated low-resource setup we take trained models from the cross-lingual setting as starting point. The model is then refined using 100 randomly selected samples from the target language for each CV split. Concretely, for the training procedure this means that the stored model parameters from the cross-lingual model are loaded as initial weights. Then, with these 100 target language samples, the network is trained again for a certain number of iterations, i.e. the network weights are fine-tuned towards the target language. Log Mel filter banks were used as input features for all models because they yielded the best overall results in the previous experiment. To be consistent with the experiments in section 3.4, the input length was kept at 7.5s; longer utterances are cut and shorter ones are padded with zeros, as described in Section 3.3 on page 67. Since the average utterance length in the RECOLA dataset (mean=2.2s, standard deviation=1.8s) is considerably shorter than for IEMOCAP, we additionally repeated experiments (a)-(c) with an input signal length of 4s to see if this makes any notable difference in the results.

For these experiments, we re-implemented the ACNN model with the Tensorflow library (Abadi et al., 2016) because the development of Theano was stopped in late 2017. For binary classification, the output layer was adapted accordingly to yield two outputs (no multi-task learning is involved here). The model's hyper-parameters for this study were similar to the setup described in Section 3.4 on page 72: 200 kernels with a size of 26x10 in the convolutional layer (spanning all 26 lMFBs); a mini-batch size of 32; and a pool size of 30 for max-pooling. For dropout regularization in the last hidden layer a dropout rate of 0.5 was applied. We ran training for 50 epochs in all experiments except for fine-tuning where the pre-trained models were refined with only 10 epochs. All experiments were run five times and the means across these five runs are reported.

3.5.2 Results

The performance measure used throughout all experiments is UAR. The results are presented in Table 3.6 on the following page.[5] The mono-lingual baseline results for both IEMOCAP and RECOLA reflect the well-known circumstance that the prediction of valence levels from acoustic properties is more difficult than for arousal (cf. Schuller et al. (2009a); Eyben et al. (2010b); Feraru et al. (2015); Trigeorgis et al. (2016); Ghosh et al. (2016a); Abdelwahab and Busso (2018)). The perfor-

[5] The results presented here are slightly different from those in Neumann and Vu (2018). The reason for this is an error in the paper regarding the valence label distribution, which was due to a mistake in data preprocessing. The experiments were conducted again with the correct labels. However, the differences are marginal and do not change the implications drawn from the results.

	IEMOCAP (English)		RECOLA (French)	
	Arousal	Valence	Arousal	Valence
mono-lingual	68.09 (68.98)	**61.84** (61.62)	60.77 (60.45)	**52.30** (49.00)
multilingual	**70.06** (71.16)	60.43 (60.63)	62.51 (60.94)	49.23 (48.58)
cross-lingual	59.32 (61.38)	49.38 (49.84)	61.27 (60.46)	48.26 (42.68)
CL + FT	67.03 (64.66)	50.94 (50.72)	**63.07** (60.15)	50.98 (51.01)

Table 3.6: Results as unweighted average recall (UAR); results with shorter input (4s) in parentheses. Cross-lingual: only trained on source language, CL + FT: pre-trained on source language and fine-tuned on 100 samples from target lanugage (CL - cross-lingual, FT - fine-tuning).

mance for RECOLA is notably lower than for IEMOCAP. This is partially due to the small size of the dataset, containing only 1,308 samples. However, this does not fully explain the differences. To test this, we trained a model on IEMOCAP using only 1,308 randomly selected samples, which still yielded better results (68.20% arousal and 58.77% valence).

Another relevant factor is again the highly imbalanced class distribution of valence labels in the French dataset, with only 241 negative samples and 1,067 positive ones (similar to what we observed with imbalanced classes on the improvised and acted subsets of IEMOCAP, cf. section 3.4.3). The UAR of 52.30% is only marginally above chance level, which means the model does not really learn anything from the underlying data.[6] One

[6] Inspection of the model predictions for valence on RECOLA has shown that: (a) a monolingual model predicts almost all samples as positive,

factor that assumingly contributes to this problem – and potentially to the skewed class distribution in the first place – is the transformation of continuous labels into binary categories. This can introduce and/or strenghten a measurement bias in the annotations, which is one of the four technical biases discussed in Dobbe et al. (2018).

With multilingual training we wanted to investigate the effect of merging the two corpora and find out whether multilingual speech emotion recognition is possible without performance loss. The results show that we are able to use a system trained on both languages and achieve similar performance compared to the baselines. For arousal prediction, the additional training data even improves performance, whereas we observe a decrease in performance for valence. These findings demonstrate that multilingual SER might be viable without further adaptation.

Cross-lingual training is useful in cases where no or only little training data in the target language is available. We therefore examined the performance of the system when trained on one language and tested on the other (and vice versa), given the same type of speech (human-human dyadic interaction). The results in Table 3.6 suggest that cross-lingual training could potentially work to some extend for arousal prediction, achieving

(b) a multilingual model predicts around 90% of all samples as positive, (c) even in cross-lingual testing around 60-70% of samples in each class are predicted as positive (alhough trained completely on IEMOCAP), (d) fine-tuning again strengthens the bias towards positive valence.

an UAR above chance level. However, it does not produce sensible results for valence prediction. A first cautious conclusion is that valence prediction could be more language-dependent than predicting arousal. However, the results from a model trained on RECOLA and tested on IEMOCAP can not be taken as basis for solid conclusions as we already described that the binary positive/negative discrimination does not work in this setup. For arousal, the performance drops notably for IEMOCAP (trained on RECOLA) compared to the mono-lingual baseline, achieving 59.32% UAR. For RECOLA (trained on IEMOCAP) it remains stable (60.77% mono-lingual, 61.27% cross-lingual). Again, these results have to be interpreted cautiously, because performance differences cannot solely be attributed to the different languages English and French. Other factors come into play, such as the very small dataset size of RECOLA, the class distributions, and the fundamentally different annotations schemes of the two datasets.

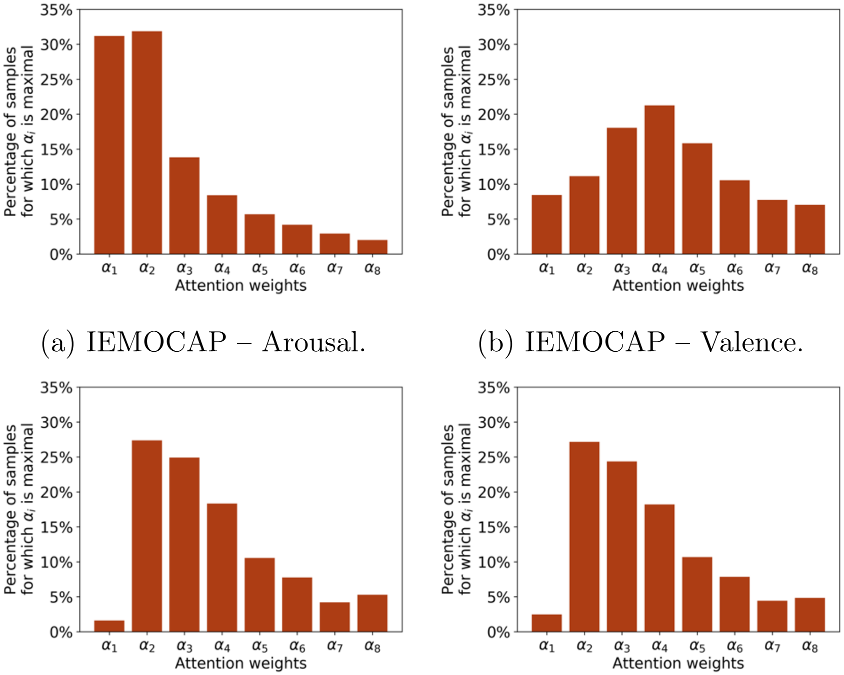
Fine-tuning the cross-lingual model with 10 training epochs on 100 samples from the target language produces promising results for arousal prediction. For IEMOCAP, the performance comes close to the baseline and for RECOLA, it is notably higher than the baseline (only for the setup with 7.5s long input data). Again, the performance for valence remains approximately at chance level. In summary, these results show that cross-lingual training can set a useful baseline. Especially for a target language with a small amount of annotated data, training a cross-lingual model and then fine-tuning it on the available target data can be a useful approach.

3.5.3 Analysis of Attention Weights

To gain more insights about which parts of the input are important for classification, we analyzed the learned attention weights α_i from the attention layer after the last training epoch. For this analysis, we focus on the mono-lingual baseline experiments. For each training sample, we output its attention weights vector and identify the maximum weight and therefore the segment on the temporal dimension that the model judges to be most salient for this sample. Figures 3.4a to 3.4d show for every index in the attention weight vector (α_1 to α_8) the respective proportion of training samples for which this α_i yielded the maximum weight.[7] For example, in Figure 3.4a, for 31.2% of training samples, α_1 was highest, which means the first segment of the input to the attention layer is considered most salient for these samples.

In Figure 3.4a we see that for arousal prediction on IEMOCAP for a large majority of samples the highest attention lies at the beginning of the input. This finding aligns with the obser-

[7] The number of attention weights corresponds to the output vector of the max-pooling layer and therefore depends on input signal length, kernel size and pool size. In the present configuration, one attention weight α_i corresponds to roughly 1 second of the input utterance.



(a) IEMOCAP – Arousal.

(b) IEMOCAP – Valence.

(c) RECOLA – Arousal.

(d) RECOLA – Valence.

Figure 3.4: Distribution of maximum attention over time.

vation in section 3.4.2 that a short snippet from the beginning of an utterance can be sufficient for a reasonable prediction. In addition to depicting the maximum attention weights, we took a closer look at the actual values of the maximum and the second highest weight to find out more about the weight distribution. Note, that the weights α_1 to α_8 sum up to 1.0 for every sample. For the English data we found that for 73.1% of all samples the difference between highest and second highest attention weight is greater than 0.5. This means, for the majority of data one single segment is weighted much higher than all others.

The distribution for valence prediction on IEMOCAP (Figure 3.4b) looks quite different: high attention is most frequently put on the middle of the signal. Further, the distribution of the attention weights themselves is a lot flatter; only for 8.8% of the samples the difference between highest and second highest α is greater than 0.5. These insights demonstrate that the attention mechanism – despite only small contributions to performance gains – can also serve as a useful analysis tool.

For the French data, Figure 3.4c and 3.4d for arousal and valence show similar characteristics: α_2 to α_4 yield the maximum weight for a large proportion of the data. Apart from α_1, the distribution exhibits a similar trend as for arousal prediction on IEMOCAP, that is the beginning of the input is much more often considered important than the end. Our first hypothesis to explain the low rate for α_1 was that RECOLA samples might contain a certain amount of silence at the beginning more often. However, using voice activity detection, we found that most signals contain speech straight from the beginning. Hence, further analysis is necessary to explain this difference. For the RECOLA dataset the distribution of attention weights is relatively flat (both arousal and valence). Furthermore, in contrast to IEMOCAP, we observed notable variations in this analysis across multiple runs of the experiment. These findings suggest that it is more difficult to learn useful attention weights for the French data compared to English. However, as mentioned already in the previous section, it is difficult to infer valid conclusions about differences in language because of other varying factors. Although the maxim for selecting the datasets was similarity in the type of conversations, they still differ considerably in the given scenarios for the recordings. One limitation of this attention weight analysis lies in the zero-padding method to create fixed-length input features. It must be assumed that this has an effect on the attention mechanism and contributes to the fact that high attention is rarely put on the last pieces of the signal; however this effect cannot be quantified or filtered out.

3.6 Summary

This chapter presented a series of investigations on speech emotion recognition with convolutional neural networks. First, a comparison of different input features was presented. We showed that the recognition performance with lMFB, MFCC, and eGeMAPS features is on a similar level, ranging from 56 to 59% UAR (54 to 56% weighted accuracy) for a four-class classification task on the complete IEMOCAP dataset. However, it is notably lower with the smaller prosody feature set. We hypothesize that the markedly smaller size (7 features) is the main reason for this difference. The similar results show that a CNN is able to learn high-level representations for the task equally from these different features. Therefore, we conclude that the particular choice of features might not be as important as the model architecture and especially the amount and type of training data. We found significant differences between *improvised* and *scripted* speech, obtaining better results on the first. The two extensions to the model, namely multi-task learning and an attention mechanism, turned out to have only little impact on the results without a completely clear pattern in favor of one or the other. We showed that MTL improves the results particularly for the CNN without attention layer, whereas it is not as clear for the ACNN. Attention itself slightly improved some results for certain combinations of features and MTL or STL, but also impaired the performance for other configurations. In any case, the attention mechansim can be useful for analyzing certain aspects of the model and the learning procedure, as shown in the second part of the chapter.

Experiments with decreasing signal length showed that the performance deteriorates slightly, but remains at a relatively stable level even for short signals down to two seconds. The analysis of error patterns revealed – unsurprisingly – that the class label distribution of the relatively small training dataset has a strong effect on the model predictions. This was particularly well observable for the class *anger*, which is strongly underrepresented in the improvised sessions of IEMOCAP and conversely overrepresented in the scripted sessions.

The second series of experiments presented in this chapter was concerned with binary arousal/valence classification in cross-lingual and multilingual settings. We have shown that multilingual classification of emotions in speech is possible, especially for arousal prediction – a valuable finding for research on code-switching speech. Further, we have shown that a model trained on a source language and fine-tuned with only a small number of samples from the target language can produce reasonable results for arousal prediction, whereas valence prediction appears to be more sensitive to cross-lingual training. These findings are useful for emotion research on low-resource languages.

The analysis of attention weights, which are learned during model training, revealed that for arousal prediction the focus lies on the beginning of the utterance in most cases (with the exception of the very first attention weight α_1 for the RECOLA data). For valence prediction the distribution of maximum attention for RECOLA is similar to arousal, but the distribution of the weight values themselves is much flatter and there is more variability across different runs of the experiment. For valence prediction on IEMOCAP the distribution of maximum attention is overall flatter with a peak in the middle and more stable compared to RECOLA.

It is important to note that the presented results have to be interpreted cautiously. In particular, for the RECOLA dataset it turned out that the binary classification of positive vs negative valence does not work because of the strongly skewed class distribution in the data. This leads us to an important aspect, that has often been unconsidered in SER research publications: the side-effects that annotations and their transformations entail. Since various annotation schemes exist for labeling emotional states in speech (cf. chapter 2), these annotations are frequently transformed in some way in practice, for example to enable cross-corpus validation or to adapt representations to certain application use cases. These transformations – in the case of RECOLA from interval scores to nominal data – are likely to introduce unwanted biases and shift the data away from the underlying Gold Standard annotation. For these reasons, Yannakakis et al. (2018) advocate the use of *ordinal* annotations and models for SER. They presented a detailed account of the various problems that such transformations can cause and described compatible and incompatible transformations between interval, nominal and ordinal ratings. These issues can be seen as technical biases, which are artifacts of *implementation* (opposed to *pre-existing* biases in the data) (Dobbe et al., 2018).

To avoid or at least reduce these aforementioned issues, the focus of the remainder of this thesis is on emotion classes (i.e. nominal data). Although one might argue that this approach is also sub-optimal (because it is questionable whether basic emotion categories are the ideal representation of the underlying phenomenon of emotional states), it is still used widely in research and application development. We are, however, aware that continuous representations (such as the arousal/valence space) as well as ordinal approaches are emerging as preferred models (cf. Gunes and Schuller (2013); Yannakakis et al. (2018)) and this needs to be considered for any future work.

4 Representation Learning and Synthetic Features

> "But it is a pipe."
> "No, it's not," I said. "It's a drawing of a pipe. Get it? All representations of a thing are inherently abstract. It's very clever."
>
> John Green, *The Fault in Our Stars*, 2012

One of the main obstacles for the development of automatic speech emotion recognition systems has been and still is the lack of large, naturalistic, annotated datasets. Compared to other speech processing tasks such as automatic speech recognition, for which thousands of hours of annotated data are available, emotional speech datasets are tiny (in the realm of a few hours to tens of hours).

This chapter deals with this problem of data scarcity and presents two different methods to appraoch it. The first part presented in section 4.1 is concerned with unsupervised representation learning (RL) from unlabeled speech using autoencoders (AE), and the second part in section 4.2 is about data augmentation with synthetically generated feature vectors.

Representation learning is concerned with automatically learning useful representations of input data for a given task. Note, that in this context the terms *features* and *representation* are often used interchangeably (because they are in fact the same: a certain form of more or less abstract representation of the data). However, to avoid confusion, in this work we use the term features to refer to *pre-defined extracted features* and the term representation to refer to an *automatically learned* depiction of some kind (e.g. activations of a neural network's hidden layer). Autoencoders are a well known and often used method for RL. They are a special kind of network architecture with the aim to reconstruct the original input given certain contraints, for instance a so called bottleneck layer that reduces the dimension of the representation. A classical AE consists of two parts, the encoder and the decoder. The first part of the network, the encoder, transforms the input features into a latent space (bottleneck layer), while the second part, the decoder, aims to reconstruct the input from this latent space. The encoder's learned transformation can then be used to transfer input data to a representation that is compact and encodes the most important information. Autoencoders, being a form of unsupervised learning (it requires no labels or other annotations), can be applied to any available data. Hence, the question arose whether such learned latent representation from an AE trained on a large speech corpus can be used to gain any additional useful information for SER.

We present experimental evidence on how representation learning can be beneficially utilized for SER. We trained unsupervised AEs on large unlabeled speech corpora and then used the encoder to generate a compact latent representation of the emotional speech samples to be classified. These representations were then integrated as an extension into the ACNN model described in the previous chapter. Experimental results show that this additional information improves the recognition performance of the classifier. To gain insights on the learned representations with respect to affective information, visualizations of the representations are presented. Evaluation was conducted on the datasets IEMOCAP and MSP-IMPROV by means of within- and cross-corpus testing.

In the second part in section 4.2, we present an appraoch to data augmentation. Since annotation is expensive and time-consuming – and in the field of emotion recognition especially difficult due to the complex and subjective nature of emotions – it is desirable to find ways of generating training data artificially. Generative adversarial networks (GANs), introduced by Goodfellow et al. (2014), have proven to be a powerful method for generating realistic synthetic data, especially in the realm of computer vision. Inspired by the successful applications of GANs and their variants for tasks like image translation and image style transfer, we transferred the idea towards generating synthetic emotional speech using this technique. As first step towards synthetic emotional speech generation, we applied a cycle consistent adversarial network (CycleGAN) on feature vectors instead of raw audio signals because this simplifies the learning process. Taking the Tedlium corpus as a large source speech resource and the IEMOCAP dataset as emotional target samples, we aimed at generating feature vectors that are close to certain target emotions in feature space (*happiness, sadness, anger, neutral state*). Further, an extension to the cycleGAN framework was introduced, which improves the discriminability of the generated data. Experimental results show that adding those synthetic features to the training set improves recognition performance in both within-corpus and cross-corpus evaluation. These experiments and findings are based on the joint work with Fang Bao (Bao et al., 2019).

4.1 Unsupervised Representation Learning for Speech Emotion Recognition

4.1.1 Related Work

A variety of different approaches to RL exists, many of them using variants of autoencoders to learn suitable representations from the data in an unsupervised manner (Ghosh et al., 2016a; Sahu et al., 2017; Ghosh et al., 2016b; Parthasarathy and Busso, 2018; Latif et al., 2018). Latif et al. (2018) used variational AEs for RL and fed the learned representation into a long short-term memory (LSTM) network for emotion recognition. A similar approach was presented by Ghosh et al. (2016a), which closely relates to the experiments presented here. The authors compared different types of AEs and input features. In contrast to the present work, these studies have not used any additional unlabeled speech resources. Potential ways to incorporate additional data have been presented in (Eskimez et al., 2018; Lakomkin et al., 2017). Eskimez et al. (2018) used four different types of AEs, trained on Librispeech data, and employed the encoders to generate representations of labeled emotional speech as input into a convolutional neural network (CNN) for SER. While this study used solely the AE representations as input, Lakomkin et al. (2017) experimented with a combination of emotion-specific and ASR-specific representations in a progressive neural network.

4.1.2 Methods

For the acutal task of SER, the ACNN model with multi-task learning objective, presented in the previous chapter, was employed. As input features, 26 lMFBs in the range 0 to 6.5kHz were extracted for 25ms long frames with a 10ms shift (cf. section 3.3 for details).

For learning a compact latent representation from unlabeled speech as additional information source we trained a time-recurrent sequence-to-sequence autoencoder on spectrograms. We used the auDeep toolkit (Freitag et al., 2017; Amiriparian et al., 2017a) for spectrogram extraction, autoencoder training and for generating representations with the learned model. For spectrogram extraction, 128 Mel frequency bands were extracted for 80ms long FFT windows with a window overlap of 40ms, following the authors' recommendations.[1] The auDeep toolkit provides a variety of options to train sequence-to-sequence autoencoders on two-dimensional data (in this case spectrograms). We employed a time-recurrent AE that processes spectrograms (of possibly varying length) along the time-axis and produces a fixed-length hidden representation. The encoder and decoder

[1] `https://github.com/auDeep/auDeep#extracting-spectrograms` [Accessed: Feb. 24, 2021]

for our experiments consist of two layers each, with 256 gated recurrent units (GRU) in each layer (cf. section 4.1.4 for details). The latent representation that is learned is a 1,024-dimensional vector.

Figure 4.1 on the next page presents an overview of the architecture and shows how the representation generated by the encoder is integrated into the ACNN training. The two networks are trained consecutively, as depicted in Figure 4.1. First, the AE is trained on some large speech corpus, e.g. Tedlium. After this training step is done, the encoder is used to generate representations for the emotional speech samples (step 2). The third step is then to train the ACNN with the additional representations as extension to the last hidden layer. Note that the encoder representations generated in step (2) are not changed anymore by the ACNN, they are just ‘plugged in’ to supply additional information about the speech sample. We applied dropout for regularization on the whole concatenated feature vector before the final softmax classification. The ACNN model architecture as well as the multi-task objective function were identical to the model described in section 3.2.

4.1.3 Data

For the experiments the two emotional speech datasets IEMOCAP and MSP-IMPROV were used. Both corpora have been created and annotated in a similar way (cf. section 2.2). Identi-

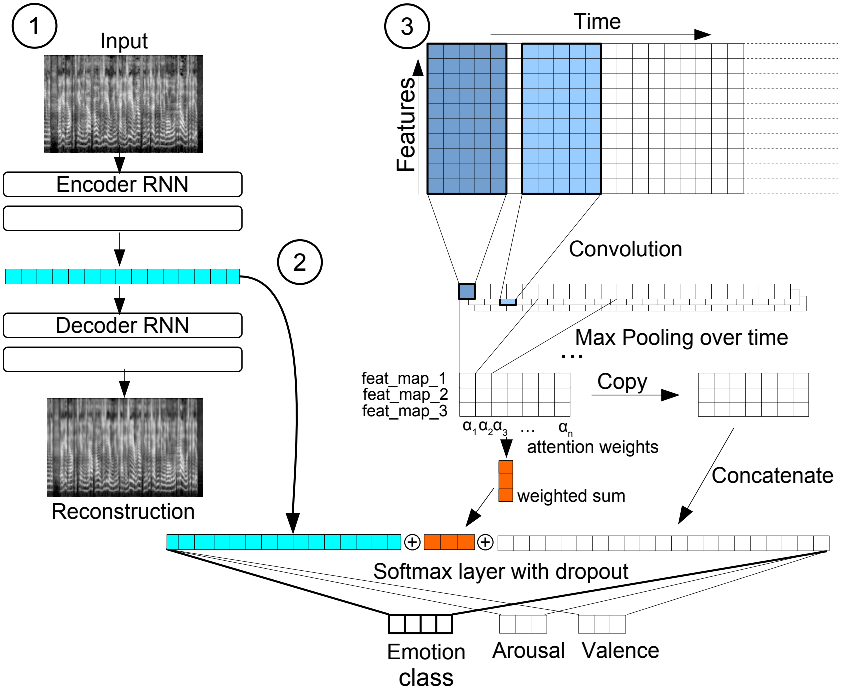


Figure 4.1: Overview of the model architecture. The training procedure follows these consecutive steps: (1) autoencoder training on a large speech corpus, (2) generation of latent representations for the emotional speech samples, (3) ACNN training with those representations as additional feature vector.

cal to the experiments in the previous chapter, the four emotion classes *anger, happiness, sadness*, and *neutral state* were taken into account (for IEMOCAP, samples from the classes *excitement* and *happiness* are merged to form one class). As in previous experiments, we set the maximal length for each sample to 7.5s. Longer turns were cut at 7.5s and shorter ones were padded with zeros. lMFBs were used as features. Arousal and valence labels were grouped into three classes each for multitask learning, following the same mapping as in section 3.4: low: [1,2]; medium: (2,4); high: [4,5] (cf. Neumann and Vu (2017); Metallinou et al. (2012)).

As additional unlabeled data for AE training we used two well-known corpora from the field of automatic speech recognition: Tedlium (release 2) (Rousseau et al., 2014) and Librispeech (Panayotov et al., 2015). Tedlium 2 is a collection of 1,495 Ted talks comprising 207 hours of transcribed English speech. We segmented the talks according to the timing information in the provided transcripts, resulting in 92,973 utterances. We have trained two models, one with the full dataset and one with a smaller subset consisting of 400 talks, or 25,303 utterances respectively. Librispeech contains 1,000 hours of read English speech from audiobooks. Due to computational limitations, we have used a subset of 100 hours, respectively 28,539 utterances.

4.1.4 Experimental Setup and Hyper-parameters

The baseline for this study is the ACNN model without any additional representation data (right-hand side of Figure 4.1 on page 104). We conducted 5-fold cross validation on IEMOCAP, taking samples from eight speakers as train and development sets and the ones from the remaining two speakers as test set. Results are averaged over the five folds.

For generating additional feature representations, we trained autoencoders on four datasets with the following motivations:

(a) The main research question is whether additional unlabeled data can be utilized to improve the accuracy of SER. For that purpose, we trained an AE on the full Tedlium 2 corpus as the main experiment.

(b) As control condition, we trained an AE only on IEMOCAP itself (respectively MSP-IMPROV for cross-corpus evaluation). In doing so, we can verify that any observed effects are in fact related to *additional* speech data compared to just adding an AE representation of the test corpus itself.

(c) To investigate a potential effect of the amount of additional data, we trained a model on a small subset of Tedlium.

(d) To confirm our findings, we used another kind of additional data in form of a subset of Librispeech.

Another research question we investigated is the effect of our approach on cross-corpus evaluation. For that, we used IEMOCAP as training set and MSP-IMPROV as test set (in the same four conditions as described above). All experiments were run ten times with different random seeds to be able to report on any variations in the results due to random parameter initialization.

Hyper-parameters The encoder and decoder of the AE consist both of 2 layers with 256 GRUs each. After testing several combinations of uni- and bidirectional encoders and decoders with regard to the reconstruction loss, we found that using a unidirectional encoder and a bidirectional decoder is a good choice. The auDeep toolkit employs Adam optimization, for which we used the default initial learning rate of 0.001; dropout was used as regulaization at a rate of 0.2 and we trained the models for 64 epochs (except for the full Tedlium dataset, for which training was done for 32 epochs).

The hyper-parameters for the ACNN model are similar to the setup in section 3.4, however, slight changes were introduced because hyper-parameter tuning was done again with the new Tensorflow implementation. The final hyper-parameters are: 200 convolutional filters of size 26x10 (spanning all 26 lMFBs), convolutional stride of 3, pooling size of 30, and a Glorot uniform initialization (Glorot and Bengio, 2010) of kernel weights. The model was trained for 100 epochs and dropout was applied at a rate of 0.8 for IEMOCAP and 0.7 for MSP-IMPROV to the last layer. For multi-task learning, the influence of arousal/valence predictions was set to a weight of 0.2 for each.

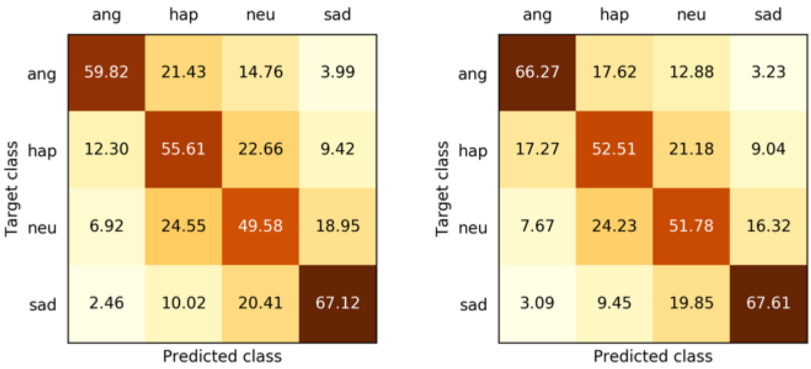
4.1.5 Results

Table 4.1 presents the results as mean UAR across all runs for each experimental configuration described in section 4.1.4. The left-hand side of the table shows the performance on IEMOCAP (5-fold cross validation) and the right-hand side the results of cross-corpus evaluation (trained on IEMOCAP and tested on MSP-IMPROV).

	IEMOCAP	MSP-IMPROV (cross-corpus)
Baseline	58.03 ± 0.76	42.99 ± 0.66
Control	58.07 ± 1.02	42.37 ± 0.77
Small Tedlium	58.85 ± 0.83†	45.21 ± 0.89‡
Librispeech	59.05 ± 0.75†	44.82 ± 1.09‡
Full Tedlium	**59.54** ± 0.63‡	**45.76** ± 0.62‡

Table 4.1: Results in UAR. Baseline: no additional represenation, Control condition: AE trained on IEMOCAP/MSP-IMPROV.
† and ‡ indicate statistically significant difference compared to the respective baseline ($^{\dagger}p < 0.05$, $^{\ddagger}p < 0.01$)

In both cases we observe small, but consistent improvements over the baseline when adding the represenations generated by the different AE models. The results for the control condition are similar to (IEMOCAP) or even below (MSP-IMPROV) the baseline. This indicates that it is in fact the *additional* speech data which helps improving the performance. It can also be seen that adding more data increases the performance further, as the best results are achieved with the full Tedlium corpus. Because the absolute improvements are relatively small, we conducted Kruskal-Wallis H-tests (Kruskal and Wallis, 1952) to see which configurations yield significantly better results than the baselines.

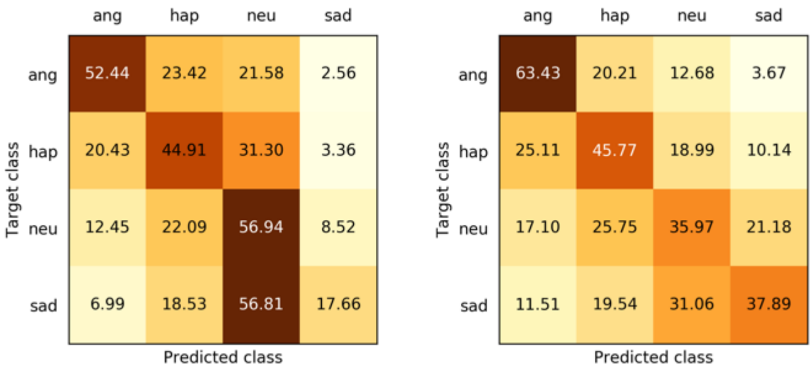


(a) ACNN baseline.

(b) ACNN+AE (full Tedlium).

Figure 4.2: Confusion matrices for mean results on IEMOCAP.

To gain more insights about the models' predictions, we analyzed error patterns, depicted in the confusion matrices in Figures 4.2 and 4.3. They represent the mean results across all ten runs of the particular setup. For within-corpus evaluation on



(a) ACNN baseline.

(b) ACNN+AE (full Tedlium).

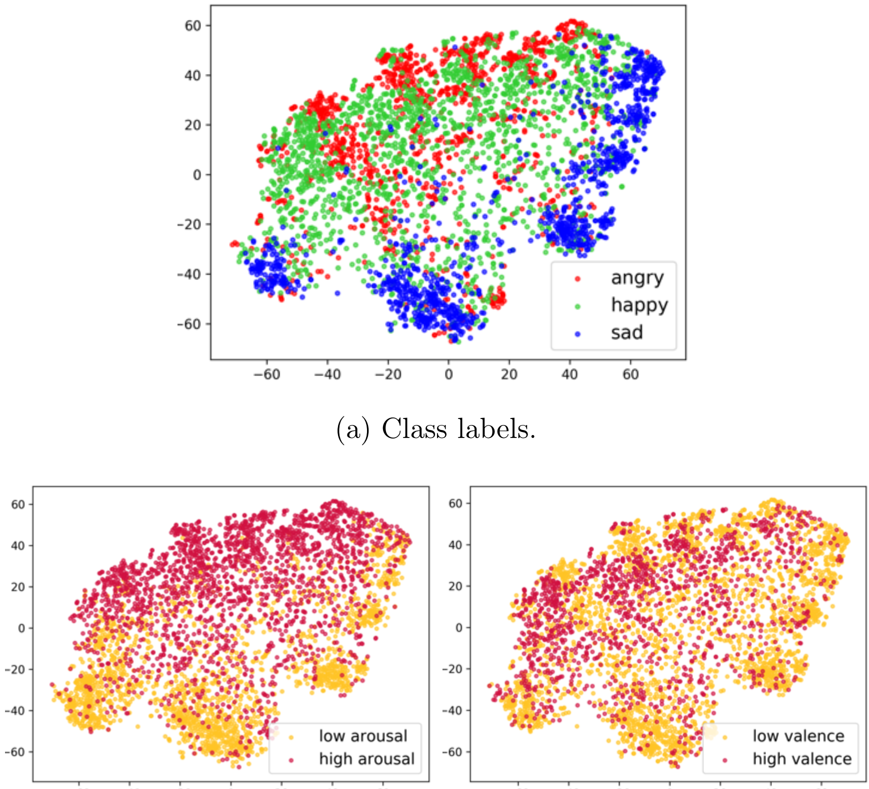
Figure 4.3: Confusion matrices for mean results of cross-corpus evaluation on MSP-IMPROV.

IEMOCAP (Figure 4.2) we see that the ACNN+AE model has a higher recall for *sadness* (marginal difference), *neutral state*, and *anger*. However, for *happiness* the recall drops below the baseline. The proportions of *happiness-anger* confusions are more balanced when adding the AE representations, indicating that the baseline model has a stronger bias towards *happiness*, which is counterbalanced to a certain extent in the ACNN+AE model. Overall, the confusion patterns between classes are similar between the two models.

For cross-corpus evaluation on MSP-IMPROV (Figure 4.3), there are significant differences for the classes *neutral* and *sadness*. Whereas the baseline ACNN trained on IEMOCAP predicts *happiness* and the *neutral state* for a large proportion of samples with high error rates across all target classes, the predictions of the ACNN+AE are more balanced. Besides a higher recall for *anger*, the main difference lies in the notably higher proportion of predictions for *sadness*, which in turn shift the *sadness-neutral* confusion and lead to a significantly lower recall for the *neutral state*. A possible reason for the different patterns compared to within-corpus evaluation on IEMOCAP is that samples from the class *sadness* exhibit different characteristics in the two corpora, and that the additional AE representations are well suited for distinguishing *sadness* from other classes.

Visualization of Speech Representations

To learn more about what kind of information might be contained in the automatically learned speech representations, we have a look at visualizations of both the learned ACNN representation (i.e. the last hidden layer before the final classification layer) and the representation from the AE trained on Tedlium. The objective here is to make visible what the two models learn – or do not learn – with regard to different aspects, such as emotion class, arousal/valence labels, speaker identity and gender. Figures 4.4 and 4.5 show two-dimensional projections of the data generated with t-distributed stochastic neighbor embeddings (t-SNE) (Maaten and Hinton, 2008). t-SNE is a nonlinear dimensionality reduction method for visualizing high-dimensional data



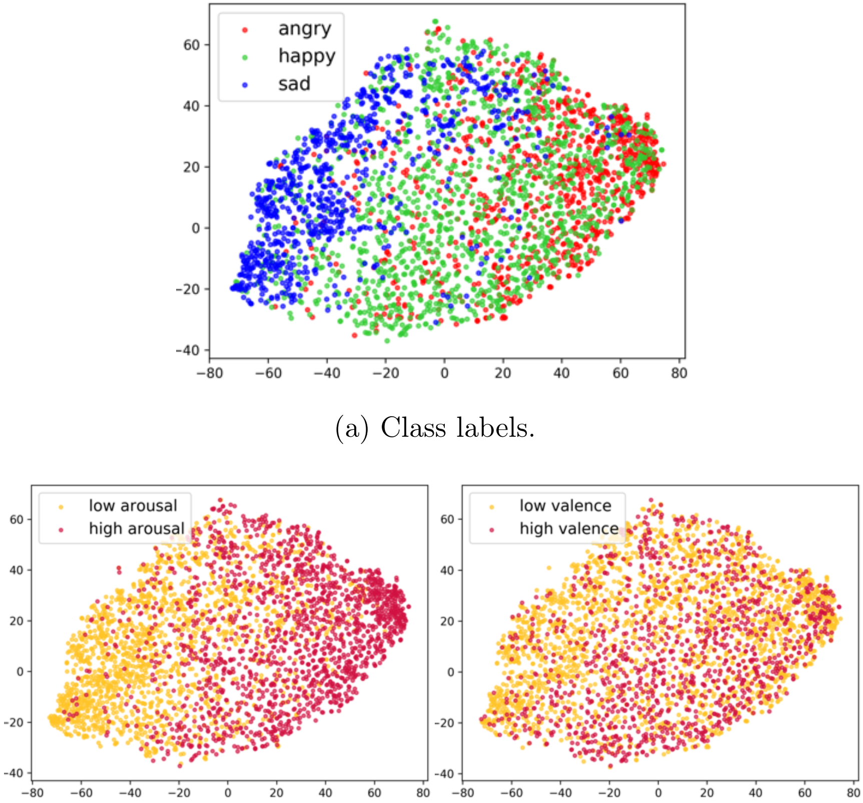
(a) Class labels.

(b) Arousal scores.

(c) Valence scores.

Figure 4.4: t-SNE visualizations of the last hidden layer of the ACNN for IEMOCAP.

in a two- or three-dimensional space. Data points are projected in a way that similar objects are close to each other and dissimilar ones are more distant from each other with high probability. The data points in the resulting scatter plots can then be col-



(a) Class labels.

(b) Arousal scores.

(c) Valence scores.

Figure 4.5: t-SNE visualizations of the AE representations for IEMOCAP (AE trained on full Tedlium).

ored according to certain information in order to reveal details of the data distribution. With this approach, we can find out whether clusters in the data (if there are any distinct clusters) belong for example to a certain emotion label or represent a certain group of speakers.

In Figures 4.4a, 4.5a and 4.7 the class *neutral state* was excluded for visual clarity; in all cases neutral samples are distributed across the whole plot and do not form a well-defined cluster. This finding has also been reported by Ghosh et al. (2016a). It can be seen that the ACNN is capable of separating *sadness* from *anger* to a certain extent. The class *happiness*, however, forms a high-variance cluster that largely overlaps with *anger*. This explains the high confusion rates seen in Figure 4.2 on page 109. The plots for arousal and valence annotations show that the ACNN model is much more discriminative for arousal than for valence – a well known characteristic of SER (cf. section 3.5 in the previous chapter).

Looking at the visualizations of the AE representation in Figure 4.5 on the preceding page, we observe similar patterns, despite no emotion labels are involved in training the model. This indicates that the learned speech representations implicitly contain information about low and high arousal, and therefore *anger* and *sadness* samples can be distinguished surprisingly well (which explains the boost for *anger* in the results in Figure 4.2 on page 109). Similar findings with respect to discriminative power between *anger* and *sadness* have also been reported by Ghosh et al. (2016a) and Lakomkin et al. (2017).

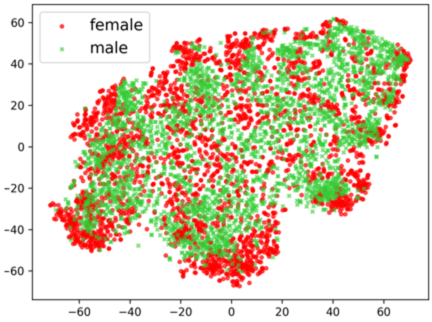
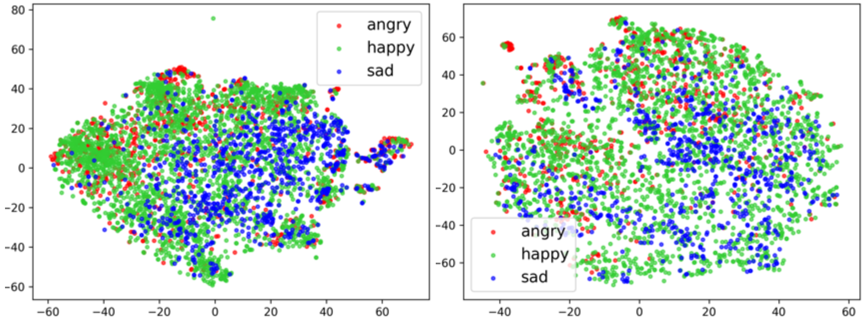


Figure 4.6: t-SNE visualizations of the ACNN representations colored by speaker gender.

Regarding speaker gender and identity we found that both representations are invariant to these factors, i.e. no separable clusters can be found in the t-SNE projections. An example of this is shown in Figure 4.6.

So far, we looked at visualizations of IEMOCAP data (i.e. within-corpus evaluation). To conclude this analysis, we will also have a look at the same kind of t-SNE projections for MSP-IMPROV data, which was used for cross-corpus evalution. Therefore, there is a difference in that the ACNN (visualizations in Figure 4.7a) was trained on IEMOCAP and then applied to MSP-IMPROV to harvest the last hidden layer's activations for visualization. For the AE representations depicted in Figure 4.7b the same autoencoder, trained on Tedlium data, was used to generate representations of the MSP-IMPROV samples. We see that in both t-SNE projections no single emotion class forms a well defined, separable cluster. This partially explains the low recall for *sadness* with the ACNN (cf. results in Figure 4.3) and it indicates that the two corpora in fact have different characteristics with respect to emotion classes. However, the visualization in Figure 4.7b does not reveal anything about the shift towards a higher proportion of *sadness* predictions in the ACNN+AE model. Assumingly, further anaylsis of the combined model would be necessary to get more insights.



(a) ACNN representations.

(b) AE representations.

Figure 4.7: t-SNE visualizations of the ACNN's last hidden layer (trained on IEMOCAP) and of the AE representations for MSP-IMPROV (AE trained on full Tedlium).

4.2 CycleGAN-based Emotion Style Transfer for Feature Generation

The problem of data scarcity may be approached from many different angles. In the previous section, we addressed the particular circumstance that *annotated* data is scarce, whereas plenty general, unlabeled speech data is available that can be used for unsupervised representation learning. Now, in this section we are going to approach the problem from a different perspective, that is generating more (labeled) training data artificially, a method known as data augmentation. While a large variety of data augmentation methods exist, for example noise injection (Ko et al., 2017) or speed variation (Ko et al., 2015), we focus here on a promising direction of utilizing generative adversarial networks to generate realistic data samples based on some target data distribution.

4.2.1 Related Work

As mentioned before, data scarcity is one of the major challenges in SER (Schuller et al., 2013), which is reflected not only in the lack of large, naturalistic labeled speech corpora, but also by the unbalanced distribution of emotions in the data (El Ayadi et al., 2011). To approach both problems, we proposed a method based on CycleGANs (Zhu et al., 2017) to generate feature vectors representing a certain target emotion. This way, the proportion of emotional categories can be controlled, thus building a large and balanced synthetic dataset.

A CycleGAN is a variation of a generative adversarial network (GAN) (Goodfellow et al., 2014). GANs have successfully been applied to a variety of computer vision tasks as well as to speech-related applications, such as speech enhancement (Pascual et al., 2017) and voice conversion (Kaneko and Kameoka, 2018). The basic principle of GANs is that they are composed of two neural networks that are trained in an adversarial manner: the generator and the discriminator. As the names suggest, the generator network is trained to generate data samples, while the discriminator's objective is to distinguish between real samples and those produced by the generator. The main idea is that during the joint training process both networks become better and better with respect to their goal, which results in generated data samples that come as close to real samples as possible.

Adversarial training schemes have also been used for SER. Sahu et al. (2017) deployed adversarial autoencoders (Makhzani et al., 2016) to represent emotional speech in compressed feature space while maintaining the discriminability between emotion categories. Chang and Scherer (2017) utilized a deep convolutional GAN to learn a discriminative representation of emotional speech in a semi-supervised way. Yet another adversarial training framework was proposed by Han et al. (2018): two separate networks are trained in an adversarial manner. One learns to predict dimensional representations of emotions, while the other aims at distinguishing between the first network's predictions and the actual target labels from the dataset. Furthermore, GANs can also be used for synthetic data generation to improve classification performance. Sahu et al. (2017) have shown this by using reconstructed samples from their adversarial autoencoder as synthetic training data. In a follow-up study, they investigated the use of a vanilla GAN and a conditional GAN for generating high-dimensional feature vectors from a low-dimensional (2-D) space (Sahu et al., 2018). It was shown that a vanilla GAN cannot achieve convergence and the conditional GAN only converges when it is initialized with pre-trained weights and the power of its discriminator is limited. The classification performance has been improved by augmenting the original training dataset with synthetic feature vectors. Inspired by this work, we proposed to generate synthetic feature vectors through emotion style transfer.

Previously, emotion style transfer had mainly been researched in the area of speech synthesis (Tao et al., 2006; Inanoglu and Young, 2009). Our approach was inspired by advances in unsupervised image-to-image translation (Zhu et al., 2017; Kim et al., 2017b; Shrivastava et al., 2017; Taigman et al., 2017; Liu et al., 2017; Yi et al., 2017). All these works have in common that a mapping between source and target domain can be learned without paired training data. For our work this means that there is no need for parallel speech corpora in which identical samples exist in a neutral and an emotional version. Instead, a CycleGAN has the ability to learn a mapping from the entirety of a provided source speech corpus to the entirety of a given emotional speech corpus as target.

4.2.2 Methods

CycleGAN as Main Building Block

Given a labeled dataset with N emotion classes, the proposed framework generates synthetic samples for each emotion class i using one individual CycleGAN. Figure 4.8 depicts the entire framework, in which the N CycleGANs build the foundation. As shown in the top half of the figure, one CycleGAN establishes a bijective mapping between a source domain S and a target domain T_i, where S can be any external (unlabeled) dataset and T_i represents the samples of emotion i in the labeled dataset. The two mapping functions G_i and F_i are used for translating from source to target and from target to source, respectively. The adversarial discriminator D_i^T encourages G_i to generate synthetic targets indistinguishable from real samples. The adversarial loss for G_i and D_i^T is defined as

$$\begin{aligned}\mathcal{L}_i^{\mathrm{GAN}}(G_i, D_i^{\mathrm{T}}, \mathrm{S}, \mathrm{T}_i) = &\underset{\mathrm{t}\sim p_{\mathrm{t}}}{\mathbb{E}}[\log D_i^{\mathrm{T}}(\mathrm{t})] \\ &+ \underset{\mathrm{s}\sim p_{\mathrm{s}}}{\mathbb{E}}[\log(1 - D_i^{\mathrm{T}}(G_i(\mathrm{s})))]\end{aligned} \tag{4.1}$$

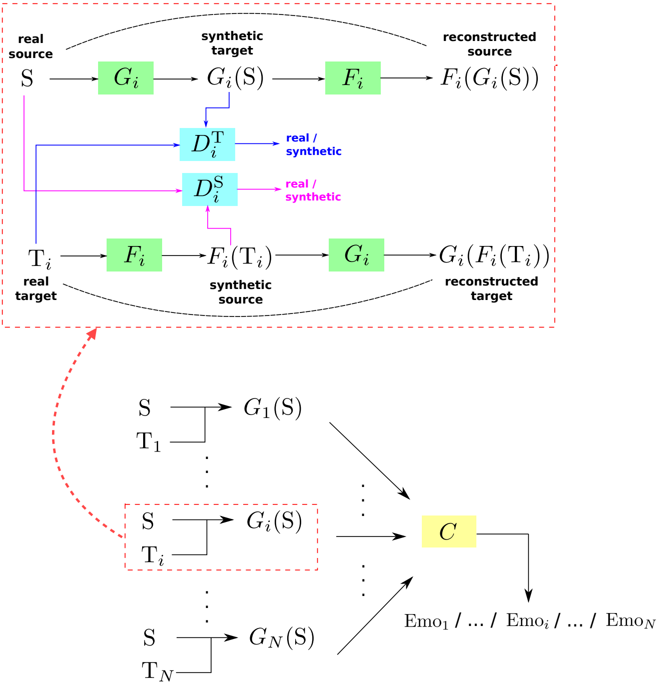


Figure 4.8: Illustration of the proposed framework. It consists of N CycleGANs, where N is the number of emotion classes. For each emotion i, there is one CycleGAN with two discriminators D_i^{T}, D_i^{S} and two mapping functions G_i, F_i as generators. The output of the mapping $G_i(\mathrm{S})$ are the desired synthetic samples. Cycle-consistency loss is built between real samples and their corresponding reconstructed samples. The domain classifier C is added to ensure the discriminability between the generated samples.

Analogous, there is an adversarial loss for the generator F_i and the discriminator D_i^{S}, $\mathcal{L}_i^{\mathrm{GAN}}(F_i, D_i^{\mathrm{S}}, \mathrm{S}, \mathrm{T}_i)$.

The total adversarial loss is defined as the sum of these two functions:

$$\begin{aligned}\mathcal{L}_i^{\mathrm{GAN}}(G_i, F_i, D_i^{\mathrm{T}}, D_i^{\mathrm{S}}, \mathrm{S}, \mathrm{T}_i) = \mathcal{L}_i^{\mathrm{GAN}}&(G_i, D_i^{\mathrm{T}}, \mathrm{S}, \mathrm{T}_i) \\ + \mathcal{L}_i^{\mathrm{GAN}}&(F_i, D_i^{\mathrm{S}}, \mathrm{S}, \mathrm{T}_i)\end{aligned} \tag{4.2}$$

The generators G_i and F_i try to minimize this loss, while the discriminators D_i^{T} and D_i^{S} try to maximize it. Intuitively, this means that the generators goal is to produce samples that the discriminators classify as *real* samples, so that the term $D_i^{\mathrm{T}}(G_i(\mathrm{s}))$ in equation 4.1 gets close to 1. The discriminators objective on the other hand is to assign high probability (close to 1) only to real samples t, and low probability (close to 0) to generated samples.

Additionally, a CycleGAN regularizes the adversarial training with a cycle consistency loss. The generated target samples $G_i(\mathrm{S})$ are translated back to the source domain and the mean squared error (MSE) between the real source S and reconstruction $F_i(G_i(\mathrm{S}))$ is computed. The same is done for T_i and the reconstructed target $G_i(F_i(\mathrm{T}_i))$. This cylce consistency loss, which can be compared to back translation in neural machine translation (He et al., 2016), prevents the problem of mode collapse, where all inputs are mapped to the exact same output.

The total cycle consistency loss is defined as follows:

$$\begin{aligned}\mathcal{L}_i^{\text{cyc}}(G_i, F_i, \mathrm{S}, \mathrm{T}_i) = &\mathop{\mathbb{E}}_{\mathrm{s}\sim p_{\mathrm{s}}}[\|(F_i(G_i(\mathrm{s})) - \mathrm{s})\|_2^2] \\ &+ \mathop{\mathbb{E}}_{\mathrm{t}\sim p_{\mathrm{t}}}[\|G_i(F_i(\mathrm{t})) - \mathrm{t}\|_2^2]\end{aligned} \tag{4.3}$$

Discriminability between Generated Samples

The bijective mapping of the CycleGAN ensures *similarity* between the distribution of real and synthetic data. However, to improve classification performance, we need to learn a generalized distribution from real data samples instead of merely reconstructing the exact same distribution. Therefore, we added a classifier C to the framework to discriminate between the generated data of each emotion class, which is illustrated in the bottom half of Figure 4.8. This additional loss function, the classification loss, can be defined as a softmax cross-entropy loss:

$$\mathcal{L}^{\text{cls}} = \sum_i y_i \text{log}(C(G_i(\mathrm{S}))) \tag{4.4}$$

where y_i is the label of the target emotion i. Ultimately, the total loss function for the proposed model is then defined as

$$\mathcal{L} = \sum_i \mathcal{L}_i^{\text{GAN}} + \lambda^{\text{cyc}} \sum_i \mathcal{L}_i^{\text{cyc}} + \lambda^{\text{cls}} \mathcal{L}^{\text{cls}} \tag{4.5}$$

The parameters λ^{cyc} and λ^{cls} are weights to control the impact of the cycle-consistency and classification loss, respectively. They affect the similarity of generated feature vectors to real data and the discriminability between emotions.

4.2.3 Data and Features

The following datasets were used for experiments with the CycleGAN framework: IEMOCAP as labeled target data, Tedlium 2 as unlabeled source data, and MSP-IMPROV for cross-corpus evaluation (cf. sections 2.2 and 4.1.3 for details about the datasets).

As mentioned before, to ease model training and optimization, we used feature vectors instead of raw audio signals throughout all experiments. The openSMILE toolkit (Eyben et al., 2013) was used to extract the 'emobase2010' reference feature set for each utterance. It is based on the Interspeech 2010 Paralinguistic Challenge feature set (Schuller et al., 2010a) and consists of 1,582 features (see section 2.3 for details). This feature set was selected because it is a widely used reference set for SER and because it was employed in the experiments in Sahu et al. (2018), with which we compare our results. In addition, we ran our experiments with the 88-dimensional eGeMAPS feature set (cf. section 2.3) in order to compare results using both a very large and a minimalistic feature set.

4.2.4 Experimental Results

Setup

Since there are four emotions to be classified, the model consists of four generators, four discriminators and one classifier. They are all implemented by feed-forward neural networks. Each generator has three hidden layers with 1000, 500, and 1000 neurons for emobase2010 features (and 64, 32, and 64 neurons for eGeMAPS). Each discriminator has two hidden layers with 1000 neurons each (2 x 64 for eGeMAPS). The classifier has two hidden layers with 100 neurons each (2 x 64 for eGeMAPS). For all hidden layers Leaky Rectified Linear Units (leaky ReLUs) were used as activation function.

Due to the difficulty for generators to learn a high-dimensional distribution, we pre-trained each pair of the generators G_i and F_i on their corresponding source and target data for 10,000 epochs with a learning rate of 0.0002 and a dropout of 0.2. The source data S consists of the full Tedlium corpus in each case, and the target data T_i consists of the particular portion of IEMOCAP annotated with emotion class i. As mentioned in section 4.2.1, the advantage of the CycleGAN framework is that no paired data is needed, which also means that source and target datasets can be of different sizes because the learned mapping is not a 1-to-1 mapping on sample level, but rather a domain-to-domain mapping.

Initialized with the pre-trained weights for generators, the complete model was then trained for 2,000 epochs with four parallel CycleGANs that transfer the unlabeled data to each of the target emotions individually. This procedure yields 92,973 synthetic feature vectors for each emotion class, i.e. a total of 371,892 samples. To reduce loss oscillation, the initial learning rate was set to 0.0002 and linearly decayed every 50 epochs by a factor of 0.8. To balance the generators and discriminators, the generators were updated twice and the discriminators once at each iteration. Besides that, we used one-sided label smoothing as introduced by Salimans et al. (2016). For both training and pre-training we used Adam optimization and a batch size of 64.

The model was implemented with TensorFlow (v 1.10.0). In terms of preprocessing, min-max normalization was used for synthetic features generation. For the emotion classification task we scaled the features on each dataset with z-normalization separately, because Zhang et al. (2011) have shown that z-normalization yields an improvement over min-max normalization for cross-corpus classification.

Experiment 1: Emotion Transfer

The aim of this first experiment is to test the feasibility of adapting CycleGANs to emotion style transfer in feature space. Our main objective was to generate feature vectors that preserve the distribution of the real target samples. The proposed cycleGAN framework was trained in two different configurations: (a) *without* classification loss (we set $\lambda^{\text{cls}} = 0$ in equation 4.5), and (b) *with* classification loss, setting $\lambda^{\text{cls}} = 2$. For both setups we set $\lambda^{\text{cyc}} = 5$, which controls the impact of the cycle-consistency loss.

We compared the distribution of the unlabeled source data, the emotional synthetic data and the target data in feature space (using the emobase2010 feature set). First, these distributions were visualized by plotting the mean and standard deviations for each feature dimension individually. A small subset of features is shown in Figure 4.9, which shows that the synthetic and target feature vectors are similar in both mean and standard deviation, which means the source data are transferred to the four target distributions successfully. Also, we observe that without classification loss ($\lambda^{\text{cls}} = 0$) the similarity is higher, which is expected because setting $\lambda^{\text{cls}} = 2$ produces more variability between the samples of different emotion classes, and in turn changes the feature distribution. To verify what is exemplified in this graphic, we manually inspected the plots for all 1,582 features and found that the demonstrated tendencies hold for the complete feature set. For this investigation all features had been normalized using min-max normalization.

In addition to this visual inspection of means and standard deviations, we explored the use of Fisher's discriminant ratio (Fisher, 1936) as a measure of overlap of feature values, as in

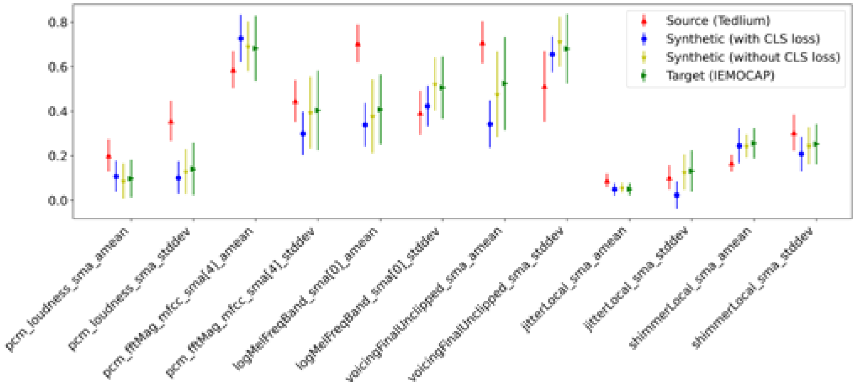


Figure 4.9: Normalized feature distributions of source, target, and synthetic feature vectors (exemplified with a subset of 12 features).

Ho and Basu (2000). It is defined as:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{4.6}$$

where $\mu_1, \mu_2, \sigma_1, \sigma_2$ are the means and standard deviations of the two distributions for a specific feature dimension. For similar feature distributions with large overlap f becomes small, and the smaller the overlap is, the larger the value for f becomes. Since we want to look at multi-dimensional distributions, we have two options for aggregating these discriminant ratios, which depend on what we want to investigate: computing the average or the maximum over all feature dimensions.

For this analysis, the two main goals are: (1) high similarity between the target (emotional speech) and the generated feature distribution, and (2) high discriminability between individual emotion classes *within* the synthetic dataset. For (1) the average discriminant ratio over all features is most appropriate to measure the overlap because the distributions should be similar in all dimensions, while for (2) the maximum value over all features is preferred to measure discriminability because as long as one highly discriminating feature exists, emotion classes can be separated regardless of other features with possibly small values for f. Figures 4.10 and 4.11 show the results for these two measures.

For the overlap between datasets in Figure 4.10 the values in the left-most column are all identical because the source data (Tedlium) does not have emotion labels, so we computed the overlap between the complete source and target datasets. For the other comparison, we split up the calculation by emotion classes. We can see that the average overlap between synthetic and target data is notably higher (indicated by a small value of f) than between source and target data. Concerning the classification loss (i.e. when $\lambda^{\text{cls}} = 2$), the plot shows a mixed result: the feature overlap for *anger* and *happiness* samples decreases compared to the middle column, while it increases for *neutral state* and *sadness.*

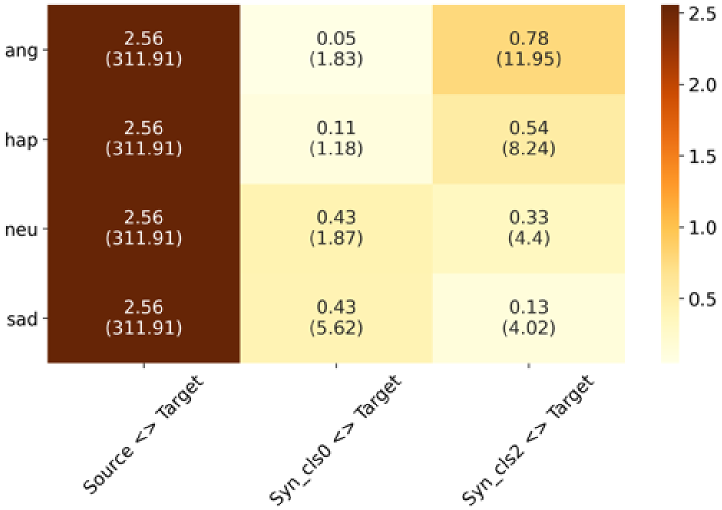


Figure 4.10: Overlap between datasets for each emotion class measured as average Fisher's discriminant ratio f over all features (maximum f is given in parentheses). Syn: Synthetic features; cls0: $\lambda^{\text{cls}} = 0$, cls2: $\lambda^{\text{cls}} = 2$.

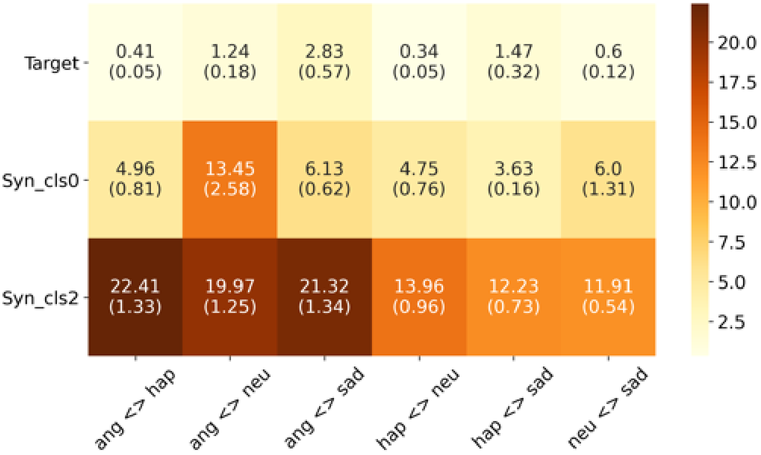


Figure 4.11: Overlap between emotion classes for each dataset measured as maximum Fisher's discriminant ratio f over all features (average f is given in parentheses). Syn: Synthetic features; cls0: $\lambda^{\text{cls}} = 0$, cls2: $\lambda^{\text{cls}} = 2$.

The comparison of emotion classes within one dataset in Figure 4.11 (shown as the maximum overlap across all features) demonstrates the consequences of the introduced classification loss more clearly. For the target dataset (IEMOCAP) the maximum values for f are all relatively low, i.e. the overlap of feature values between emotion classes is rather high. For both synthetic datasets these values are notably higher, reaching the highest ratios when classification loss is used (bottom row in the figure). Primarily, this means that the feature distributions between emotion classes are enforced to be more different in the latter case. However, from this analysis nothing can be inferred about how representative these synthetic feature vectors are for a given emotion class. This leads us to the next experiment, where we used the generated data to test SER performance on IEMOCAP.

Experiment 2: Within-corpus Evaluation

For evaluating the usefulness of the generated samples, we implemented three feed-forward neural network classifiers that were trained on: (i) only real samples taken from IEMOCAP, (ii) only synthetic features and (iii) the combination of both.[2] We per-

[2] These models were created and trained independently from the CycleGAN framework, i.e. whenever the term *classifier* is used in this and the following section, we do not mean the classifier component C in Figure 4.8 but a separate classifier that is trained and evaluated on the created data.

	Real	Syn.	Real + Syn.
Hidden layer sizes	100, 100	200, 200	1000, 1000
Dropout rate	0.2	0.5	0.5
Batch size	64	256	256
Init. learning rate	1e-5	1e-5	5e-6
Number of epochs	70	5	30
Optimizer	Adam		
Activation function	Leaky ReLU		

Table 4.2: Hyper-parameters of feed-forward neural networks for within-corpus evaluation.

formed leave-one-session-out cross validation on IEMOCAP to ensure that results are speaker-independent. The hyper-parameters for these models are listed in Table 4.2. When training only with synthetic data, we observed that the model overfits very fast, hence the small number of training epochs. We report unweighted average recall (UAR) as performance measure. All experiments were repeated five times and we report mean and standard deviation of the results. For the eGeMAPS feature set we set the layer size to 100 units per layer for all settings due to the smaller size of the feature set, the remaining hyper-parameters were kept unchanged.

Table 4.3 on the following page shows the results and, for comparison, the results reported in Sahu et al. (2018) for the three experimental settings. It can be seen that the baseline performance with the emobase2010 feature set is comparable to Sahu et al. (2018). Using the combined dataset (real + syn.), we achieved an improvement over the baseline when incorporating the classification loss into the cycleGAN ($\lambda^{\text{cls}} = 2$). Augmenting the dataset with synthetic features generated without this loss did not yield an improvement with emobase2010 features.

	Real	**Syn.**	**Real + Syn.**
Sahu Sahu et al. (2018)	59.42	34.09	60.29
emobase2010, $\lambda^{\text{cls}} = 0$	**59.48 ± 0.71**	51.57 ± 0.60	58.79 ± 0.77
$\lambda^{\text{cls}} = 2$		46.59 ± 0.75	**60.37 ± 0.70**
eGeMAPS, $\lambda^{\text{cls}} = 0$	54.28 ± 1.03	54.83 ± 0.80	54.89 ± 0.59
$\lambda^{\text{cls}} = 2$		**55.40 ± 0.83**	55.26 ± 0.72

Table 4.3: Results for cross validation evaluation on IEMOCAP.

Using only synthetically generated samples as training data, we observe a significantly higher performance on the test set (51.57%) than reported in Sahu et al. (2018), which implies that our cycleGAN approach generates feature vectors that are closer to the underlying distribution of real data. Interestingly, with the emobase2010 features the UAR for the setting with $\lambda^{\text{cls}} = 2$ is notably lower than for $\lambda^{\text{cls}} = 0$.

With eGeMAPS features, we observe overall only small differences between the different configurations. While the baseline on only real data and the results on the augmented dataset are notably lower than with emobase2010 features, the result when training only on synthetic features higher (55.40%). This suggests that with the much smaller eGeMAPS feature set, there is not as much variation between synthetic and real training data.

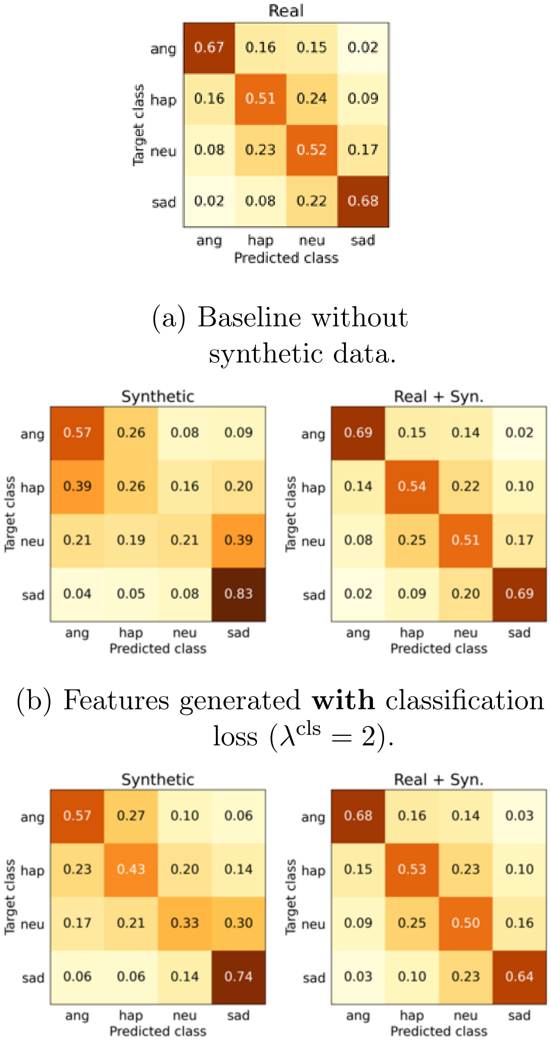
But the overall best result of 60.37% is achieved with the larger feature set.

To gain a deeper understanding of the performance differences, we analyzed the prediction errors (for models trained on emobase2010 features[3]), shown in Figure 4.12a- 4.12c.

It can be seen from the confusion matrices that the predictions and error patterns based on the augmented dataset (real + syn., right-hand sides of Figure 4.12b, 4.12c) are similar to the baseline (Figure 4.12a). For the setting *with* classification loss (Figure 4.12b), we observe improvements for the three classes *anger, happiness, sadness* – whereas the result for *sadness* drops below the baseline in the setting *without* classification loss (Figure 4.12c).

Substantial differences between the two configurations are found in the predictions when using only synthetic data as train set (left-hand sides of Figure 4.12b, 4.12c). For $\lambda^{\text{cls}} = 2$, the model appears to have a strong bias towards the classes *anger* and *sadness*, given the high proportions of incorrect predictions of those two classes. For $\lambda^{\text{cls}} = 0$, the proportions of samples wrongly predicted as *sadness* and *anger*, respectively, are also high, but Figure 4.12c presents a more balanced confusion matrix for synthetic samples overall. The total UAR of 51.57% is

[3] In addition, we also inspected the confusion matrices for models trained on eGeMAPS features. They exhibit similar patterns as the ones shown here.



(a) Baseline without synthetic data.

(b) Features generated **with** classification loss ($\lambda^{\text{cls}} = 2$).

(c) Features generated **without** classification loss ($\lambda^{\text{cls}} = 0$).

Figure 4.12: Averaged confusion matrices for IEMOCAP with emobase2010 feature set.

higher than for $\lambda^{\text{cls}} = 2$ and the bias towards *sadness* and *anger* not as severe.

These findings show that the proposed classification loss in our cycleGAN framework can in fact improve classification results (for real + syn.), but potentially introduces biases towards certain classes. In addition, we have recognized a strong overfitting problem when training *only* on synthetically generated feature vectors.

Experiment 3: Cross-corpus Evaluation

To investigate whether the generated synthetic samples are useful when applying a model to another, unseen dataset, we performed cross-corpus evaluation in the same three setups as described in the previous section, using the MSP-IMPROV dataset as test data. We took 30% of the samples as development set for hyper-parameter tuning and the remaining 70% as test set, keeping class proportions equal in both sets.[4] For (ii) *Syn.* and (iii) *Real + syn.*, the following hyper-parameters differ from the within-corpus setup: 200 neurons per layer (100 units with eGeMAPS features), dropout of 0.8, learning rate of 1e-5 and 20 training epochs for both setups. The high dropout rate ap-

[4] Note that hyper-parameter tuning was primarily done for the emobase2010 feature set. Due to resource limitations, we did not conduct the complete procedure again for eGeMAPS, which potentially affects the reults.

peared to be necessary because of the overfitting problem with synthetic samples.

	Real	**Syn.**	**Real + Syn.**
Sahu Sahu et al. (2018)	45.14	33.96	45.40
emobase2010, $\lambda^{\text{cls}} = 0$	**45.58 ± 0.40**	39.35 ± 0.33	42.61 ± 0.34
$\lambda^{\text{cls}} = 2$		**41.58 ± 1.29**	**46.52 ± 0.43**
eGeMAPS, $\lambda^{\text{cls}} = 0$	40.03 ± 0.67	24.39 ± 3.29	34.27 ± 1.79
$\lambda^{\text{cls}} = 2$		29.47 ± 2.17	33.03 ± 2.53

Table 4.4: Results for cross-corpus evaluation on MSP-IMRPOV.

The results in Table 4.4 show similar characteristics for emobase2010 features as the results for within-corpus evaluation: adding synthetically generated training samples improves the classification performance (with classification loss, i.e. $\lambda^{\text{cls}} = 2$). Also when using only synthetic training samples, the UAR for $\lambda^{\text{cls}} = 2$ is higher than for $\lambda^{\text{cls}} = 0$, suggesting that the introduced classification loss is beneficial for cross-domain scenarios. However, with eGeMAPS features the baseline constitutes the best result, while adding synthetic data diminishes the performance. When training only on synthetic features, the UAR even drops below chance level ($< 25\%$) for $\lambda^{\text{cls}} = 0$. This shows that the findings need to be interpreted cautiously and that they do not necessarily generalize across different conditions. We assume that one reason for this discrepancy was the hyper-parameter tuning, which was primarily done with the emobase2010 feature set (we only reduced the size of the neural network when transferring to the smaller feature set).

4.3 Summary

In this chapter we presented experimental results on representation learning and feature generation (data augmentation) for SER. First, we have shown that incorporating representations generated by an autoencoder that was trained on a large dataset, leads to statistically significant and consistent improvements in recognition accuracy of the proposed ACNN model (for wihtin and cross-corpus evaluation). Visualizations of the learned representations (created with t-SNE dimensionality reduction) revealed the discriminative strength of those representations with regard to low and high arousal. With cross-corpus evaluation on MSP-IMPROV we further showed that speech samples of the two used datasets that are assigned to the same emotion class (e.g. *sadness*) seem to exhibit notably different acoustic properties. This became evident when comparing the recall for *sadness* as well as the visualizations of AE representations with regard to emotion classes between the two corpora

This investigation has shown that RL on additional speech data can be beneficial for SER performance. For future experimentation in this direction, *robustness* with respect to varying recording conditions and noise could be a possible subject of study. We hypothesize that speech representations that are automatically learned on large corpora with high variability in this respect, can improve robustness and increase generalizability of models.

The second part of the chapter addressed another topic related to data scarcity: data augmentation with generative adversarial networks. In contrast to previous methods, where synthetic feature vectors were generated from a low-dimensional space (Sahu et al., 2018), we proposed a CycleGAN-based method to transfer unlabeled data into different target emotions. The experiments have shown a considerable similarity between the distribution of synthetic and target feature vectors. Furthermore, we introduced a classification loss to the network architecture as an additional regularizer to enable the generated samples to be better distinguishable. Experimental results on IEMOCAP and MSP-IMPROV with the high-dimensional emobase-2010 feature set (1,582 features) have demonstrated improvements in classification performance over previous methods when training on synthetic features as well as on the combination of real and synthetic samples. The same models were additionally trained using the minimal eGeMAPS feature set (88 features; only layer sizes of the neural networks were adapted). These results showed a stable UAR across settings for within-corpus evaluation on IEMOCAP, with slight improvements for training on the synthetic and the combined data. However, for cross-corpus testing on MSP-IMPROV the performance with eGeMAPS features notably deteriorated when the generated feature vectors were involved. One possible reason for this is that the classifier's hyper-parameters were tuned with the emobase2010 features to work well in the cross-corpus setting, while this tuning step was not done for the additional experiments with eGeMAPS. In any case, this comparison shows that such experimental evidence needs to be interpreted cautiously and that it is generally necessary to test various conditions (such as different datasets, features, or model architectures) to derive generalizable conclusions. Furthermore, we observed a strong bias towards certain emotion classes in the synthetic data and strong overfitting when training only on these samples. These problems need to be solved in future work, possible directions are varying the weights of the differnt parts in the loss function (λ^{cls} and λ^{cyc}) to find the optimal balance between similarity and discriminability as well as utilizing additional speech emotion corpora as target data for cycleGAN training.

5 Audiovisual Emotion Recognition in Noisy Conditions

> Our intuition tells us that our senses are separate streams of information. We see with our eyes, hear with our ears, feel with our skin, smell with our nose, taste with our tongue. In actuality, though, the brain uses the imperfect information from each sense to generate a virtual reality that we call consciousness. It's our brain's best guess as to what's out there in the world. But that best guess isn't always right.
>
> David Ludden, *Hearing With Our Eyes, Seeing With Our Ears*, 2015

Emotion recognition has seen great advances both in speech and facial expression analysis, more and more blending into each other with the rise of multimodal machine learning. In this chapter we explore emotion recognition under noisy acoustic conditions and investigate in audiovisual feature fusion in order to improve the overall performance. The main research questions are:

(i) How does *speech* emotion recognition perform on noisy data? and

(ii) To what extend does a *multimodal* approach improve the accuracy and compensate for potential performance degradation at different noise levels?

We present an analytical investigation on two English audiovisual emotion datasets, MSP-IMPROV and CREMA-D. Acoustic noise was superimposed to the original audio files at different signal-to-noise ratios. For this, in order to simulate real-world scenarios with various kinds of background noise, randomly selected samples from three different domains of real noise recordings were used (instead of using static white noise). Focusing on noisy *acoustics*, we compared and analyzed the results with three different types of acoustic features. For multimodal fusion, visual features are extracted from the videos with a pre-trained CNN model for image recognition.

As one would expect, the experimental results show a strong performance degradation when audio-only models are applied to noisy audio – throughout different features, noise types and datasets. Both the addition of visual features and the addition of noisy samples (data augmentation) to the training data significantly improve the accuracy. Further, we found that multimodal fusion helps to distinguish the high-arousal emotion classes *anger* and *happiness* better from each other in clean audio conditions. Comparing three acoustic feature sets, one main finding is that a CNN with log Mel filter banks (lMFBs) as input performs most *stable* under noisy conditions, compared to feed-forward networks on the other feature sets, which exhibit strong biases towards single classes. However, the overall best performance was achieved with eGeMAPS features and a feed-forward network. These results have been published in Neumann and Vu (2021).

This chapter is structured as follows: We highlight the most relevant related work in section 5.1, before presenting the neural network architectures and the used feature sets in sections 5.2 and 5.3, respectively. Section 5.4 outlines the experimental design and the results are presented in section 5.5. The chapter is rounded off with a summary of the main findings in section 5.6.

5.1 Related Work

Research on automatic affect recognition has long focused on one modality individually (e.g., speech *or* facial expressions), but multimodal emotion recognition – and multimodal machine learning in general – has gained much attention in recent years. For a comprehensive survey on multimodal machine learning the reader can refer to Baltrušaitis et al. (2019). Regarding emotion recognition, Sebe et al. (2005) presented a survey on multimodal approaches and proposed probabilistic graphical models for fusing modalities. In another early work, Busso et al. (2004) compared early and late fusion and showed that acoustic and visual features contain complementary information about expressed emotions. More recently, deep learning and end-to-end learning gained traction in the field of (multimodal) emotion recognition, partially because of the availability of larger amounts of training data (Tzirakis et al., 2017; Han et al., 2019; Ghaleb et al., 2017; Wöllmer et al., 2013; Mallol-Ragolta et al., 2019).

Another aspect of increasing interest is the performance of systems outside of clean laboratory conditions, which is for example addressed by the 'Emotion Recognition in the Wild Challenge' (Dhall, 2019). The effect of noisy data has been investigated for *speech* in several studies (Schuller et al., 2006, 2007; You et al., 2006; Zhao et al., 2014) and speech enhancement methods are one promising direction for better SER quality (Chenchah and Lachiri, 2016; Avila et al., 2018; Zhang et al., 2016; Triantafyllopoulos et al., 2019). While we are aware of the different methods to attenuate noise effects in speech data, we focused specifically on a multimodal approach because only few studies have addressed the problem of noisy data in audiovisual experiments and we wanted to investigate the complementary effects of both modalities. Banda and Robinson (2011) focused on the effect of corrupted videos and showed that a multimodal system retains a reasonably high performance compared to video-only. Lin et al. (2013) added noise to both audio and video and presented a semi-coupled Hidden Markow Model to diminish the negative impact of noise. In contrast, we focused on noisy *acoustics*, leaving the videos untouched. An even more important aspect with respect to previous work is that the above-mentioned studies all have in common that the training and test data *match* (i.e. either clean or noisy data). A more realistic setting, however, is that training and test data vary from eath other with respect to noise type and noise levels. Therefore, we conducted experiments with different configurations, including models that are trained on clean audio data only and then applied to different noise levels.

5.2 Methods

For both unimodal and multimodal experiments we trained fully connected feed-forward (FF) neural networks (except for lMFBs, for which a CNN is applied). All models were implemented with PyTorch (Paszke et al., 2019). The focus of this investigation is on low complexity of the models to facilitate reproducibility

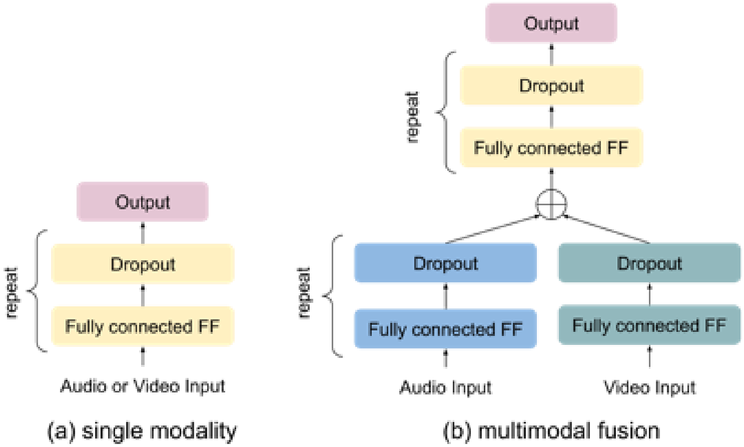


Figure 5.1: Scheme of neural network architectures that consists completely of fully connected feed-forward layers and are used for eGeMAPS and DeepSpectrum features.

and – since it is a comparative experiment – to reduce potential factors of variations. The study design is based on and inspired by the work in Wand et al. (2018). The FF networks are composed of a stack of fully connected layers with tanh non-linearity, each followed by dropout regularization, as shown in Figure 5.1. For lMFBs, which are a time-preserving 3-dimensional representation, we trained a strided CNN composed of two convolutional layers with ReLU activation, each followed by a dropout layer (depicted in Figure 5.2 on the next page). There is no pooling layer, but pooling is implicitly controlled by tuning the stride size of the convolution. The kernels of the first convo-

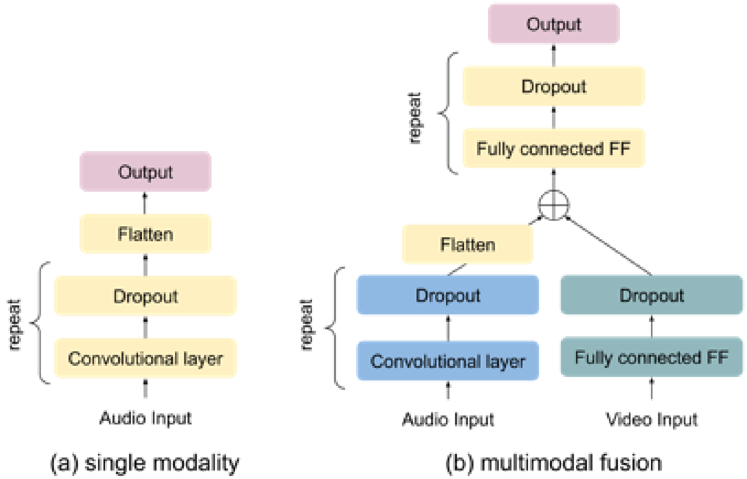


Figure 5.2: Scheme of neural network architectures that contain convolutional layers for acoustic log Mel filter bank features.

lutional layer span the entire input feature dimension (23 filter banks), i.e. the subsequent layer is a 1-D convolution over time. As explained in section 3.3, CNNs require a fixed input size. We set the sample length to 7.5s for MSP-IMPROV and 3s for CREMA-D (based on *mean duration + standard deviation* of each corpus) and applied zero-padding for shorter utterances. The two-dimensional output of the convolutional layers (a 'stack' of feature maps) is then flattened in order to feed into the fully connected output layer.

For multimodal fusion we used a hybrid approach, shown in Figures 5.1 (b) and 5.2 (b), respectively: audio and video input is fed into separate sub-networks, which consists of at least one layer. The sub-networks' outputs are then concatenated and fed into a joint network. For filter banks, the sub-network is a CNN (as described above) whose output is flattened and fed into the joint network.

5.3 Data and Features

5.3.1 Datasets

Two datasets were used for this experiment: MSP-IMPROV and CREMA-D (cf. section 2.2 for details). MSP-IMPROV audio files were downsampled to 16 kHz for all experiments (originally provided with 44.1 kHz). Identically to previously described experiments, 6-fold cross validation (leave-one-session-out) was applied as evaluation method to ensure speaker-independent evaluation. 10% of the training set were randomly selected as development set for hyper-parameter tuning.

As second dataset we used CREMA-D, a crowdsourced audiovisual dataset of emotional read speech. To facilitate comparisons between datasets, we used the same four classes as for MSP-IMPROV, resulting in 4,799 samples. As there are no default train and test splits for this dataset, we split the data by speaker IDs to avoid speaker overlap: speakers 1-63 as train, speakers 64-77 as development, and speakers 78-91 as test set. We have verified that these partitions are balanced regarding age and gender distributions. Video recordings show the speakers in front of a green screen in both datasets.

5.3.2 Acoustic Features

Three different types of acoustic features have been examined, including handcrafted (eGeMAPS) and model-based representations (DeepSpectrum). Noisy audio was generated by following the approach taken in Wand et al. (2018): Three different categories of noise samples from the Freesound database (Font et al., 2013) were superimposed to the clean audio at {-10, -5, 0, 5} dB signal-to-noise ratio (SNR) using the acoustic simulator presented in Ferras et al. (2016). The categories are ‘ambience-babble’ (151 samples), ‘ambience-music’ (96 samples), and ‘ambience-transportation’ (186 samples). For each category noise samples were randomly drawn to superimpose the noise to the emotional speech utterances one by one. Throughout all experiments, a sample rate of 16kHz was used (for clean and noisy audio).

eGeMAPS is a feature set recommendation for affective computing (see section 2.3 for details). We used openSMILE (Eyben et al., 2013) to extract the 88-dimensional eGeMAPS feature vectors.

As second feature set **lMFBs** were used as input to a CNN, as this approach has been shown to produce good results previously. Note that we use the terms lMFB and ‘filter banks’ interchangeably in the following sections. We used PyTorch’s torchaudio package (Paszke et al., 2019) to extract 23 lMFBs with the default settings.[1]

For the third feature set we used **DeepSpectrum** (Amiriparian et al., 2017b), a Python toolkit for acoustic feature extraction based on pre-trained image CNNs. Spectrograms are generated from the acoustic signal and fed into a pre-trained CNN. The activations of a specific layer form the feature vectors. We used the DeepSpectrum default settings, i.e. extract the activations of the ’fc2’ layer from AlexNet, resulting in a 4,096-dimensional feature vector for one utterance.[2] DeepSpectrum features have previously been shown to work well for emotion recognition (Cummins et al., 2017) and served as baseline features in the 2018 and 2019 Audio/Visual Emotion Challenge (AVEC) (Ringeval et al., 2018, 2019).

For all features z-score normalization was applied globally (i.e. across all speakers) on the training set and the computed feature means and standard deviations were then taken to normalize the test set. We compared global vs. speaker-wise normalization on

[1] 25ms window size, 10ms shift, frequency range for Mel bins: 20.0 - Nyquist frequency, Povey windowing function; `https://pytorch.org/audio/compliance.kaldi.html#fbank`

[2] `https://github.com/DeepSpectrum/DeepSpectrum`

MSP-IMPROV and found that there was no noticable difference in the normalized features.

5.3.3 Visual Features

Visual representations have been obtained as follows: Each video frame was converted to gray scale and Contrast-limited adaptive histogram equalization (Pizer et al., 1987) was applied to enhance contrast, followed by face recognition using dlib's (King, 2009) frontal face detector. The detected face region was cropped and resized to 100x100 pixels to feed it into the VGG 11-layer model (configuration *A* of the model described in Simonyan and Zisserman (2015)), which was trained on ImageNet (Russakovsky et al., 2015). The processing pipeline is depictd in Figure 5.3. We took the activations of the first fully connected layer as frame-level intermediate features and applied average pooling across all frames of the same utterance to obtain a 4,096-dimensional utterance-level representation.

5.4 Experimental Setup

To establish baselines, we assessed the accuracy on clean audio on the development set (average result across 6 folds for MSP-IMPROV) in a grid search for the following hyper-parameters: number of layers, number of neurons per layer, dropout rate (additionally for CNN: number and size of feature maps, stride

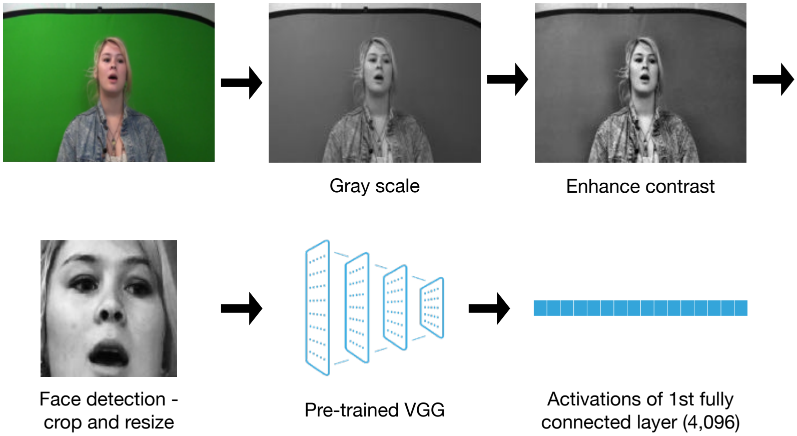

Gray scale
Enhance contrast
Face detection - crop and resize
Pre-trained VGG
Activations of 1st fully connected layer (4,096)

Figure 5.3: Illustration of the visual feature processing pipeline (which is applied to each frame in a video).[3]

size). This was done for each feature type (unimodal and multimodal) and dataset individually. Table 5.1 shows the number and size of layers for each input option. The CNN hyperparameters are: 128 feature maps of width 10 and stride 7 for MSP-IMPROV and CREMA-D unimodal, and 128 feature maps of width 15 and stride 3 for CREMA-D multimodal. For MSP-IMPROV, we applied dropout at a rate of 0.5 for all except the DeepSpectrum features, for which we selected 0.7 to prevent overfitting. For CREMA-D, overfitting appears more prevalent, presumably due to the artificial nature of the data. We obtained

[3] The following icon from the Noun Project (`https://thenounproject.com`) is used in Figure 5.3: *Convolutional neural network* by Oleksandr Panasovskyi

a dropout rate of 0.7 for all but the unimodal eGeMAPS features, for which 0.5 was applied. Model training was done with a batch size of 32 for 100 epochs for unimodal and 50 epochs for multimodal input. Early stopping was not employed because of loss oszillation. We inspected loss and accuracy on the training and development sets as a function over number of epochs and found that 100 and 50 epochs respectively were suitable for this data. We ran all experiments (except hyper-parameter tuning) three times and report mean and standard deviation in terms of unweighted average recall (UAR).

	MSP-IMPROV		CREMA-D	
	# layers	# neurons	# layers	# neurons
eGeMAPS	2	128	3	64
DeepS	2	128	2	128
Video (V)	3	256	3	128
eGeMAPS+V	1+1	256	1+1	128
Filterbanks+V	2+2	256	2+1	128
DeepS+V	1+1	128	1+1	128

Table 5.1: Hyper-parameters. 'x+y' means: x layers in each sub-network and y layers in the joint network.

5.5 Results and Analysis

5.5.1 Models trained on clean and applied to noisy audio

In this first experiment we applied trained models to data with different noise levels and compared the results with the clean reference data. The results are shown in Figure 5.5. Comparing different noise types, we observed the same tendencies across features and datasets. Overall, the recognition performance is slightly higher for transportation noise and lowest for babble noise. Figure 5.4 shows results for all three noise types on the MSP-IMPROV data with eGeMAPS features. Because of these similarities between noise types, we describe and analyze only the results for babble noise in greater detail in the following.

In general, we observed a large decline in performance from clean to noisy audio on both datasets, with decreasing UAR for higher noise levels (as one would intuitively expect). Adding visual features consistently improves the performance by a large margin. In the following, the detailed results for each dataset are presented.

For **MSP-IMPROV** (Figure 5.5a), the best audio-only result is 48.59% ±0.23% (clean audio, eGeMAPS), the best multimodal result is 53.50% ±0.14% (clean audio, DeepSpectrum). The video-only UAR is 44.24% ±0.59%.

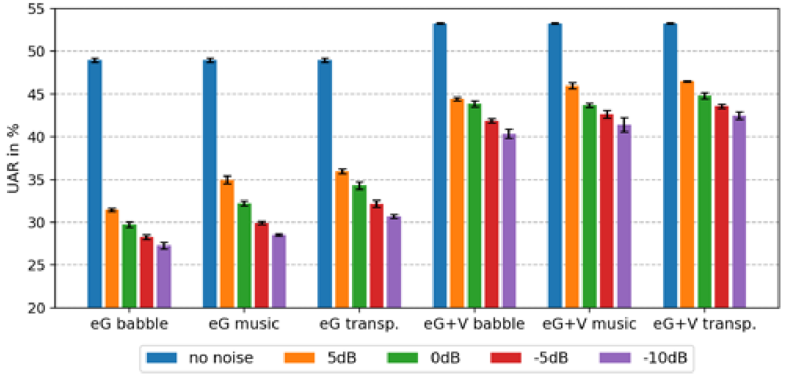
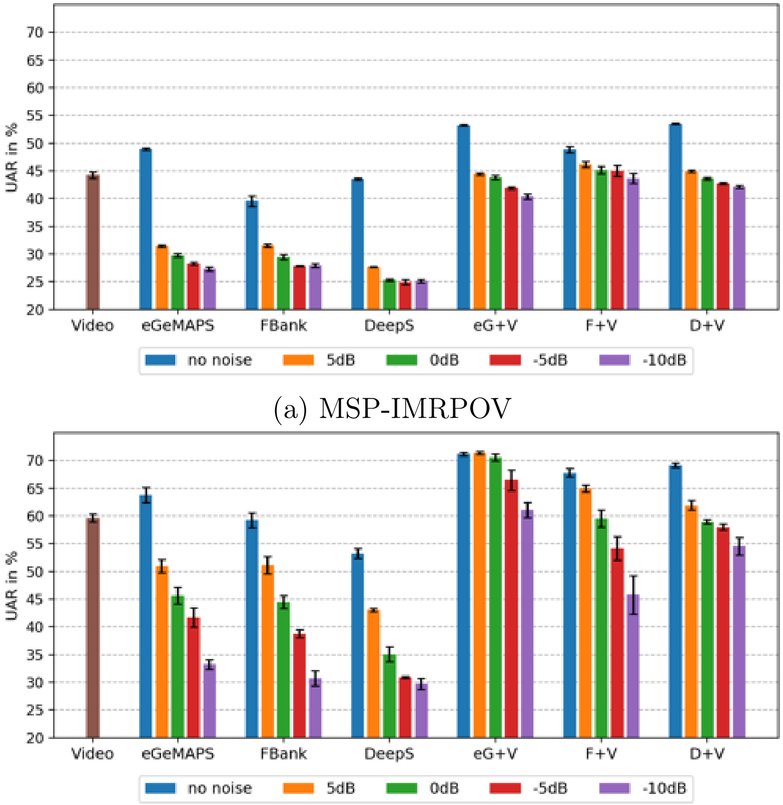


Figure 5.4: Results on MSP-IMPROV for training with clean audio only and applying the models to noisy audio data (eGeMAPS features, all noise types).

To gain more insights on the results, we analyzed the models' predictions. Exemplary confusion matrices are shown in Figure 5.6 on page 160 and Figure 5.7 on page 161. For clean audio, the individual class recalls with eGeMAPS and DeepSpectrum features are well balanced (cf. Figure 5.6a). For filter banks we observed a bias towards the class *neutral state* (Figure 5.7a). Adding visual features improves the recall for *anger* and *happiness* notably, which can partially be explained by the high recall for *happiness* in the video-only case (73.7%). When applied to noisy audio, we observed that *happiness* is predominantly predicted for all three feature types. With eGeMAPS (audio-only), this effect is most pronounced (cf. Figure 5.6b). Adding the visual modality improves recall for the other three classes significantly (cf. Figure 5.6d).



(a) MSP-IMRPOV

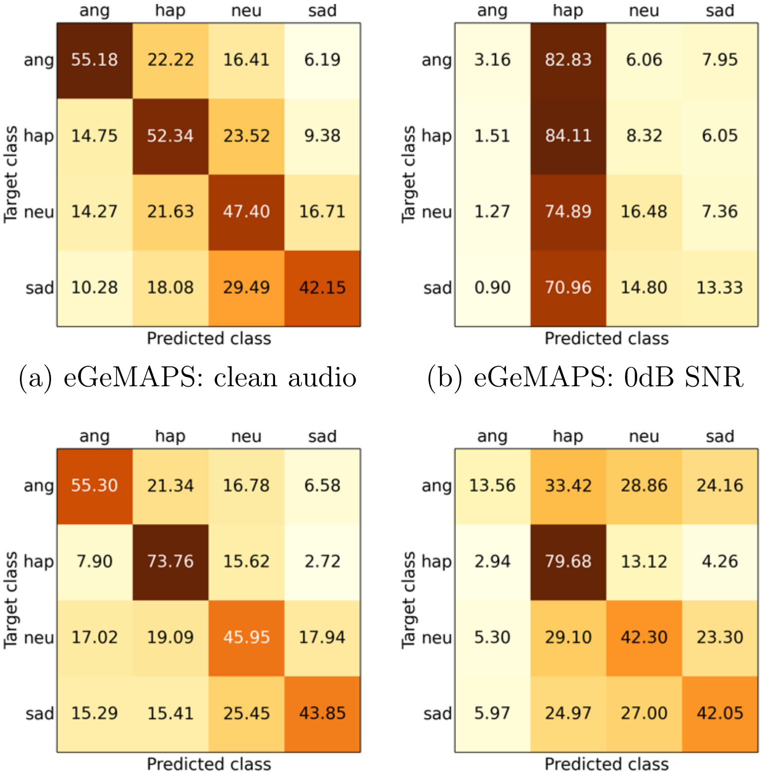
(b) CREMA-D

Figure 5.5: Results for training on clean audio only and applying the models to noisy audio data (babble noise).
FBank (F): filter banks, DeepS (D): DeepSpectrum, eG: eGeMAPS, V: Video.

With filter banks and DeepSpectrum features (audio-only), the majority of samples is predicted as either *happiness* or *neutral state* at low noise levels and the bias towards *happiness* increases with higher noise levels. Adding visual features improves the recall for *anger* and *sadness* considerably (cf. Figure 5.7a-5.7d), but a bias towards the high-arousal classes *happiness* and *anger* remains at higher noise levels.

One main finding of the analysis is that the performance *decline* on noisy data is smallest for the model with filter bank features. However, the reference performance on clean data is lowest in this case. The confusion matrices show that the model with filter bank features is more stable on noisy audio data with respect to the balance between classes and the bias towards one single class. Figure 5.6b and 5.7b illustrate this comparison: While with eGeMAPS features, the number of samples wrongly predicted as *happiness* is very high and almost all *anger* samples are predicted as *happiness*, these effects are less pronounced with filter bank features. This difference between the models and features becomes even larger at high noise levels.

For **CREMA-D** (Figure 5.5b), the results show the same patterns as for MSP-IMPROV: a performance decline proportional to the noise level (even more pronounced at higher noise levels). An exception is the combination of visual and eGeMAPS features, for which the UAR at 5dB SNR is higher than on clean audio. We found that the benefit of adding visual features is



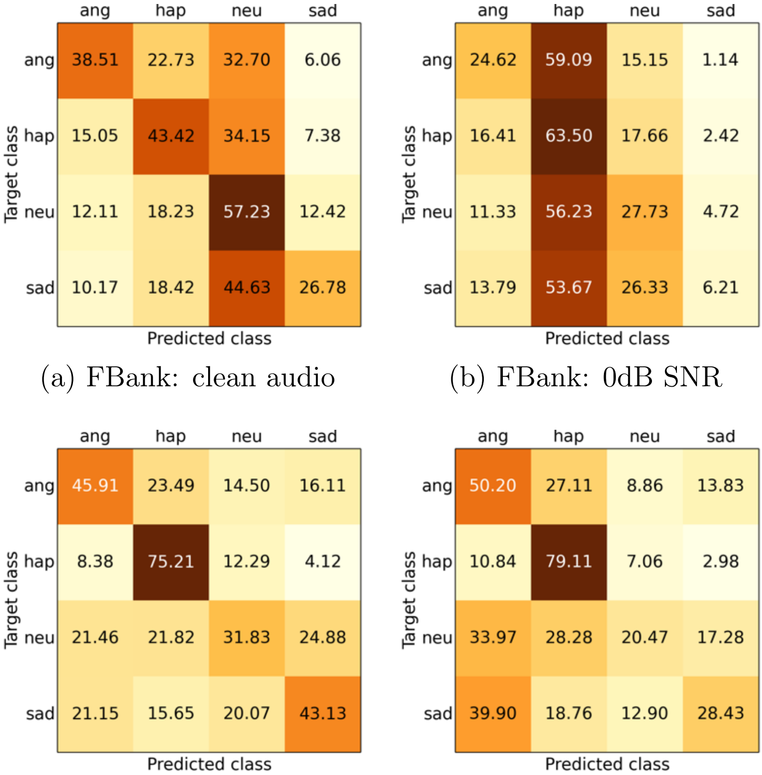
(a) eGeMAPS: clean audio

(b) eGeMAPS: 0dB SNR

(c) eG+V: clean audio

(d) eG+V: 0dB SNR

Figure 5.6: Results for MSP-IMPROV from uni- and multimodal models trained on clean audio only and tested on clean and noisy audio. eG: eGeMAPS, V: video.



(a) FBank: clean audio

(b) FBank: 0dB SNR

(c) F+V: clean audio

(d) F+V: 0dB SNR

Figure 5.7: Results for MSP-IMPROV from uni- and multimodal models trained on clean audio only and tested on clean and noisy audio. F: filter banks, V: video.

differently distributed across emotion classes. In general, the largest improvement is observed for the class *happiness*, which also has the highest recall for video-only (same as for MSP-IMPROV). As a result, it can happen that the total UAR is slightly higher on noisy data than on clean data because the imbalance between classes increases and certain biases are even more emphasized. This effect is not as strong for MSP-IMPROV because overall the differences in recall for individual classes are not as large.

The best audio-only result is 63.76% ±1.35% (clean audio, eGeMAPS), the best multimodal result is obtained with eGeMAPS features (71.38% ±0.28% at 5dB SNR and 71.17% ±0.29% on clean audio). The video-only UAR is 59.63% ±0.71%.

The inspection of confusion matrices showed high recall for the class *anger* on clean audio throughout all feature sets. With eGeMAPS the class recalls are most balanced, while with filter banks a high proportion of samples is wrongly predicted as *neutral state*. On noisy audio we observed a strong bias towards *happiness* with eGeMAPS, a strong bias towards *anger* with filter banks and high confusion between *sadness* and *neutral* with DeepSpectrum features. With higher noise levels the biases towards the high-arousal classes *anger* and *happiness* become stronger. The addition of visual features improves generally the recall for *happiness* and *neutral*.

5.5.2 Models trained on single noise levels

For the second analysis, we trained and evaluated the models at the same noise level (*matched* condition, as it was frequently done in related work). Figure 5.8 shows the results for MSP-IMPROV. The results for CREMA-D exhibit similar characteristics. In contrast to the first experiment the performance remains much more stable for noisy data when the model is trained on this kind of data. The results within one feature set are at a similar level with a tendency of slightly lower UAR for higher noise levels. This decrease is most pronounced for eGeMAPS features (audio-only). These observations paired with the findings of the first experiment emphasize the severe consequences that can be caused by a mismatch between train and test data.

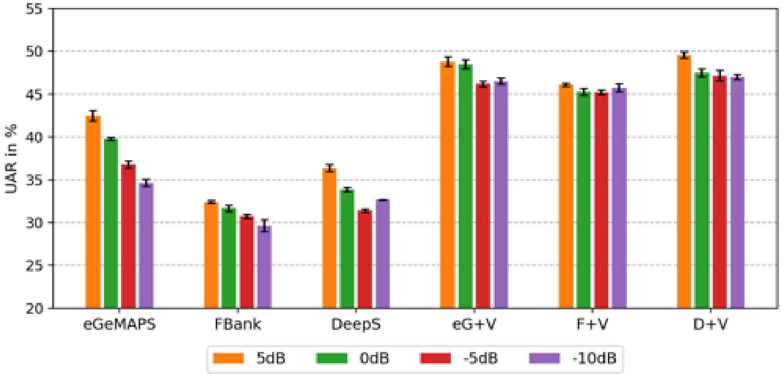


Figure 5.8: Results on MSP-IMPROV for training and evaluation on single noise levels (babble noise).

5.5.3 Data augmentation: Training on all noise levels

In the third experiment all models are trained on the union of data from all noise levels (including clean audio) and evaluated on the different noise levels separately (similar setup as in section 5.5.1). The results for MSP-IMPROV are shown in Figure 5.9. Again, CREMA-D results exhibit similar characteristics. Compared to Figure 5.5a, the performance on clean audio decreases throughout all features and multimodal combinations, while it improves remarkably on noisy data, especially for audio-only. The addition of visual features is especially useful with filter bank and DeepSpectrum features. These results show that data augmentation in the form of added noise is ben-

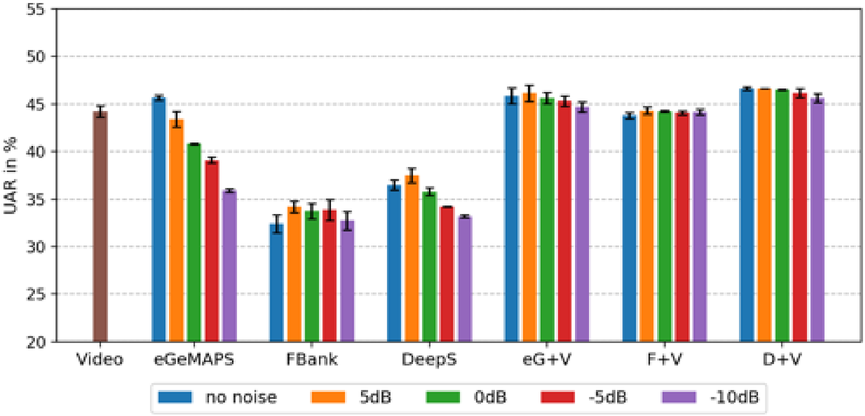


Figure 5.9: Results on MSP-IMPROV for training on all noise levels with babble noise (data augmentation).

eficial in noisy conditions. However, a trade-off between lower accuracy on clean and higher accuracy on noisy data needs to be accepted.

5.6 Summary

In this chapter on audiovisual emotion recognition in noisy acoustic conditions we have shown that the SER performance decreases significantly when training and test data do not match (clean vs. noisy), and that this effect can be dampened with audiovisual models. Intuitively, this does not seem to be surprising observation. However, it underlines the importance of evaluation in such mismatched conditions, which is neglected by a large proportion of related research work. As outlined in section 5.1, most studies have only investigated noisy data in the matched condition, where training and test data come from the same (or a similar) distribution with regard to background noise. Generally spoken, more cross validation needs to be done when the question at hand is whether developed models can be applied to real-world use cases. With cross validation, many varying parameters can be evaluated in order to verify the robustness of a model, including different noise conditions, recording conditions, languages, speaker characteristics such as sex and age, just to name a few. In this contribution we investigated noisy audio as one such varying parameter and demonstrated the potentially serious consequences of mis-matched conditions. Furthermore, we showed that data augmentation by adding noise to the training set increases the accuracy on noisy audio significantly, but can affect results on clean data negatively. Hence, a trade-off between the results obtained on clean and noisy data needs to be found (which may be dependent on the specific application or use case).

The comparison between feature sets showed that the eGeMAPS parameter set yields the best overall results, given the basic feed-forward networks that have been employed in this analysis. However, the inspection of error patterns revealed that the CNN with log Mel filter bank features yields more stable predictions under noisy conditions with respect to the magnitude of accuracy decline and the class balance in predictions.

One limitation of the present study is that the Lombard effect – the phenomenon that people speak differently than usual in noisy environments – is not taken into account. To consider this, future work needs to be based on real-world noisy speech data instead of superimposed noise.

6 Conclusion and Future Directions

This chapter summarizes the main findings with regard to the two overall research goals that we started off with: systematic investigations of certain aspects in SER, and extensions to basic modeling approaches. Furthermore, we present and discuss ethical considerations with respect to the work presented in this thesis and also to the broader perspective of affective computing. While this brief account on the ethical impact of our research is not intended to be exhaustive, we nevertheless want to highlight certain ethical aspects, since SER (and more generally, affective computing) is no longer confined to the work in research institutions, but it is employed in many real world applications and products that may have ethical consequences for individuals or society at large. Lastly, we conclude this work by outlining some directions for future work to address the limitations of this thesis.

6.1 Summary and Key Findings

We showed that convolutional neural networks are an efficient modeling approach to automatically learn high-level speech representations suitable for classifying basic emotions. With the proposed ACNN model we found that the particular choice of preprocessed acoustic features as input to the neural network does not substantially impact its performance, as long as enough information is contained in the input (only the very minimalistic prosodic feature set consisting of seven features yielded significantly lower accuracy). However, by comparing scripted conversations with improvised play, it was shown that the type of speech can have a significant impact on accuracy. Concerning the length of the input signal — basically asking the question, 'How long does the model have to listen before it makes a prediction?' — we showed that a short segment of the beginning of an utterance can be sufficient for prediction because the performance loss for shorter input down to about two seconds is marginal.

Apart from the input data in terms of features, speech type and signal length, we investigated the feasibility of multilingual and cross-lingual SER with experiments on English and French speech data. We showed that arousal level prediction is feasible across languages, especially when fine-tuning on the target data can be applied. However, predicting valence levels did not work well in these multi- and cross-lingual settings. These results should be interpreted cautiously with respect to the two languages used, because language alone was certainly not the only varying factors between the datasets. The last investigated aspect was noise robustness of SER. By evaluating feed-forward neural networks in mismatched noise conditions (i.e., the model was trained on clean and then applied to noisy audio data), we showed a severe performance degradation. Although this was an expected finding, we emphasized the importance in research to evaluate such mismatchted conditions, which has only rarely been done in the literature. As one solution to improve the accuracy, we showed that data augmentation by adding noisy data to the train set helps.

As for the second goal, possible extensions, several directions were explored. Using an attention mechanism to make the CNN model *attentive* to the most salient information, we showed that this yields slight improvements for certain constellations of input speech type and acoustic features. Beyond that, the analysis of the learned attention weights demonstrated that such an attention layer is a useful vehicle for analysis of the network. Looking at the problem of data scarcity, we presented two different promising appraoches: representation learning on unlabeled speech, and generating artificial samples that represent a certain target emotion. We showed that general-purpose speech representations learned by an autoencoder are useful as additional feature to improve SER performance and that a CycleGAN framework can be successfully employed for generating artifical training data by means of emotion style transfer. Finally, we explored multimodal processing of audio and visual information (facial expressions) to attenuate the performance degradation in noisy environments. With a model-level fusion approach of the two modalities, we showed that the addition of visual features does not only improve the general performance, but helps significantly to reduce the loss when the audio channel is corrupted by noise. Especially the high-arousal classes anger and happiness, which are difficult to distinguish based on acoustics alone, can be better told apart with a multimodal model.

6.2 Ethical Considerations

As emotion processing and related computational paralinguistics technologies are being used more and more in real-world applications for a variety of purposes, it becomes also increasingly important to put responsible use of such in the focus of the general discourse. Examples for such applications are found in the automotive industry (e.g. detecting stress or other states in drivers (Eyben et al., 2010c)), in call centers (e.g. reacting to angry customers (Burkhardt et al., 2009)), in the health sector (e.g. asssessing mood and emotional states to monitor neurological conditions (Cummins et al., 2018; Matton et al., 2019)), or in the entertainment sector (e.g. players' emotions influence the plot of video games (Lobel et al., 2016; Jones and Sutherland, 2008)).

Ethical guidance for artificial intelligence (AI) and its related areas and use cases exists, for example in the Institute of Electrical and Electronics Engineers' *Ethically Aligned Design* (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017) – however, *practical* guidelines and binding standards are still rather the exception. IEEE provides high-level recommendations in various areas of autonomous and intelligent systems, including affective computing. In a meta-analysis of existing ethical guidelines, Jobin et al. investigated whether there are globally agreed principles of ethical AI (Jobin et al., 2019). They identified five main principles that occurred most frequently: *transparency, justice and fairness, non-maleficence, responsibility*, and *privacy.* Focused on computational paralinguistics specifically, Batliner et al. (2020) give a comprehensive account on the specific ethical demands within the field and emphasize the importance of representative data and interpretability of outcomes.

Based on the literature, we identified the following pivotal ethical cornerstones with regard to SER and its potential applications:

- Performance – quality of models in terms of accuracy and robustness

- Representative data and models – different types of biases and how they can be addressed
- Transparency – disclosure of the inner workings and the underlying data, comprising interpretability/explainability
- Privacy – a large area that encompasses, among others, data pricacy and security, anonymizing algorithms, decentralized processing techniques, emotion-hiding methods
- Accountability – the question of who is responsible and accountable for decisions taken by autonomous systems is one of the 'General Principles' in *Ethically Aligned Design* (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017)

In the following, we discuss aspects that are directly related to the work presented in this thesis, which particularly pertain to the first two cornerstones, performance and representative data and biases.

In the experiments presented in the previous chapters, it became evident that the accuracy of SER is still at a relatively low level.[1] Performance, being expressed in terms of some evaluation metric, is usually not considered as an ethical aspect (but a

[1] Note, at the time of writing, state-of-the-art results in the literature for the used benchmark datasets are undoubtedly higher than the ones reported in this work, since the 'race for the highest accuracy' is moving at fast pace in the research community. However, this certainly comes at the cost of even lower generalizability. Usually, state-of-the-art results are presented on a specific dataset and assessing a model's performance

rather technical one) and is just assumed to be given; it appears that *optimal* performance is assumed in ethical discourses and guidance. However, it should be critically asked, whether an algorithm is 'good enough' for a particular use case. If this is not thoroughly assessed, potential unethical consequences can arise (e.g. if SER is used in personality tests for job recruitment, or for employee monitoring). As we have shown in chapter 5, when applied in conditions different from the training data (in this case acoustic noise), prediction accuracy can deteriote significantly. This is also generally the case for these small training data sets in other cross validation settings. For these reasons, we consider model performance and generalizability as critical ethical aspects, because the usage of inaccurate models can have severe consequences. A related problem is the 'veneer of accuracy', that is when applications or products are claimed to work optimally and accurately, given certain benchmark numbers, but they in fact fail to deliver this accuracy once released to the open, i.e. applied to unseen data.

Closely related to performance is the question how representative data and models are (including biases that emerge if data and/or algorithms are not representative). The prevailing discussions on bias are typically about training data bias (with the main theme often being 'AI itself is not biased, but the under-

with cross validation on different data is still not widely done. Generalizability remains a crucial challenge, also due to small and potentially biased datasets.

lying data are'). However, various distinct sources and types of biases exist and it is important to distinguish them. Already in early works on bias in computer systems, Friedman et al. identified three overarching types: *pre-existing, technial*, and *emergent* bias (Friedman and Nissenbaum, 1996). More recent work on algorithmic bias suggests a more fine-grained taxonomy, distinguishing biases with respect to *training data, algorithmic focus, algorithmic processing, transferring context,* and *interpretation of outcomes* (Danks and London, 2017). With regard to SER and the presented work in this thesis, we view training data bias (pre-existing) and several technical biases as the primary issues that need to be addressed in the research community. Considering the small datasets used for training and evaluating SER systems, the obtained results usually cannot be transferred to real world use cases yet and trained models are not representative for a broader population of speakers. Concerning technical bias, which comprises all tools and processes used to transform data into models that make predictions, Friedman (1996) described one source of bias as "the attempt to make human constructs amenable to computers – when, for example, we quantify the qualitative, make discrete the continuous, or formalize the nonformal." Dobbe et al. (2018) pointed out that technical bias is particularly domain-specific, but they identified four common sources of bias in the typical machine learning pipeline: *measurement, modelling, label*, and *optimization bias.*

Measurement bias describes how the procedure of transforming data and labels into machine-readable representations can induce unwanted effects (for example using a nominal instead of an ordinal scale to measure some variable). In SER, various types of measurements are used, both for the speech data itself (e.g. the unit of analysis can be a long utterance or a short audio frame; it can be modeled statically or time-continuously) and for labels (e.g. basic emotions on a nominal scale or dimensional representations on interval scales). Consequently, when putting SER into practice, these possible sources of bias need to be considered and the right measures for the particular application have to be chosen. Another critical issue in emotion processing technologies, which seems to be largely neglected in the technology-oriented AI community is *label bias*, which is about the question how representative the chosen output labels are for the actual phenomenon. The target labels we aim to predict (such as basic emotions) are unavoidably some discretized proxies for the actual emotional states because emotions cannot be objectively measured (no ground-truth exists). Some researchers in psychology are even calling for an overhaul on how emotions are understood in computational analysis by challenging the common view that certain physical movements (e.g. facial expressions) are intrinsic displays of specific emotional states (Barrett et al., 2019). While we acknowledge that it will remain difficult, if not impossible to truly 'read some-ones emotions', as emotion itself is a difficult concept to grasp even for humans, we argue that conducting research on acoustic parameters that correlate with certain states and traits of the speakers is important and useful for a large variety of applications. Consequently, depending on the application domain, it can be fully acceptable to utilize abstract proxies, like a closed set of emotion words to describe user states, whereas it might not be applicable to other contexts. To conclude, it is important to raise awareness in the community for these different types of biases and how they can affect the outcomes, which always need to be carefully interpreted.

6.3 Outlook

In this work, we focused on the *basic emotions* approach and therefore used mainly nominal emotion labels and treated the task of SER as classification problem. As already detailed in chapter 2, the second widely utilized type of emotion representation is the dimensional model with the two predominant dimensions of arousal and valence. Both approaches are broadly in use and have their own advantages and disadvantages. On the one hand, the basic emotions approach facilitates interpretation of the labels by users of a system because people usually have a common notion of how an *angry* or *sad* voice sounds like (at least in the prototypical sense). On the other hand, value ‘continuous’ dimensions eliminate the problem of artificially introduced clear-cut boundaries between classes and allow for more fine-grained nuances. A mapping from dimensional labels to basic emotions and vice versa is sometimes done (e.g. in cross-corpus evaluation), however it is not straightforward. It is conceivable that future work will examine such mappings further and how the different approaches can be made compatible.

In section 3.2 we described our proposed ACNN model with its necessity to provide fixed-length feature representations as input data. Although handling variable length input by means of zero-padding has been a common approach when working with CNNs, the approach has its limitations: the resulting data is either very sparse when we take the longest training sample as the threshold and pad everything else, or we lose information when we cut long utterances to a certain shorter threshold (and still have to pad many samples with zeros). While this has not been critical for the presented work because the used datasets consist of short dialog turns, it can pose problems to other types of data and use cases. One possible alternative could be a sliding analysis window, i.e. multiple shorter chunks of equal size are fed into a model and the predictions are fused in some way at the end for the complete utterance. Another broadly researched alternative, which was not in the scope of this work, is time-continous modeling where the task is not treated as classification, but as regression. This usually goes together with using continuous emotion dimensions, i.e. a model is trained to predict an arousal or valence value for each point in time at a certain rate. For this approach, recurrent neural networks and their variants have been established. With their ability to 'remember' information and therefore take preceding context into account, these types of neural networks are a promising avenue for SER applications.

In section 3.5.3 an analysis of the learned weights from the attention mechanism was presented. Another attempt to gain more insights about the inner workings of the model was presented in section 4.1.5, where the activations of the CNN's last hidden layer were visualized using t-SNE dimensionality reduction. These analyses provided useful insights, but were confined to looking at weights (without knowing what exactly is weighted) and outputs respectively. In recent years, explainable AI (XAI) has become a research field in its own right, which is concerned with uncovering what is actually represented and learned inside neural networks (Adadi and Berrada, 2018; Arrieta et al., 2020). Consequently, future work will provide more insight into the inner representations and how they can be interpreted in order to explain and justify predictions to the users.

Lastly, another area that is already gaining a lot of attention and which we have touched rather superficially is multimodal machine learning. Because of the complementary information that different modalities can provide – in the case of emotion recognition particularly speech and facial expressions – a large performance gain can be expected when the task is modeled holistically with as much available information as possible (which information channels are readily available strongly depends on the domain and use case). Outstanding challenges in this interdisciplinary area include the development of robust and efficient fusion methods (for potentially heterogeneous data), and handling noisy or missing data from one or the other modality. Coming back to the introductory quote by Peter Drucker, many challanges are still waiting to be solved on the way to let machines *"understand what isn't being said"*.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16.

Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., and Penn, G. (2012). Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE.

Abdelwahab, M. and Busso, C. (2018). Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Adel, H. and Schütze, H. (2017). Exploring different dimensions of attention for uncertainty detection. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics.*

Amiriparian, S., Freitag, M., Cummins, N., and Schuller, B. (2017a). Sequence to sequence autoencoders for unsupervised representation learning from audio. In *Proceedings of the DCASE 2017 Workshop*.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., and Schuller, B. W. (2017b). Snore sound classification using image-based deep spectrum features. In *Proceedings of Interspeech*.

Anand, N. and Verma, P. (2015). Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data. In *Technical Report*. Stanford University.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Avila, A. R., Alam, M. J., O'Shaughnessy, D. D., and Falk, T. H. (2018). Investigating speech enhancement and perceptual quality for speech emotion recognition. In *Proceedings of Interspeech*, pages 3663–3667.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 41(2):423–443.

Banda, N. and Robinson, P. (2011). Noise analysis in audio-visual emotion recognition. In *Proceedings of the 11th International Conference on Multimodal Interaction (ICMI)*. Citeseer.

Bao, F., Neumann, M., and Vu, N. T. (2019). Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition. In *Proceedings of Interspeech*, pages 2828–2832.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: new features and speed improvements. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning.*

Batliner, A., Hantke, S., and Schuller, B. W. (2020). Ethics and good practice in computational paralinguistics. *IEEE Transactions on Affective Computing.*

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. *Proceedings of a*

Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: A cpu and gpu math compiler in python. In *Proceedings 9th Python in Science Conf.*

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1042–1047.

Bradley, M. M. and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

Burkhardt, F., Van Ballegooy, M., Engelbrecht, K.-P., Polzehl, T., and Stegmann, J. (2009). Emotion detection in dialog systems: Applications, strategies and challenges. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4).

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S.

(2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211. ACM.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. (2017). Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1).

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

Chang, J. and Scherer, S. (2017). Learning representations of emotional speech with deep convolutional generative adversarial networks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Chenchah, F. and Lachiri, Z. (2016). Speech emotion recognition in noisy environment. In *2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 788–792. IEEE.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, pages 577–585.

Cibau, N. E., Albornoz, E. M., and Rufiner, H. L. (2013). Speech emotion recognition using a deep autoencoder. *Proceedings of the XV Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2013), San Carlos de Bariloche.*

Collobert, R. (2011). Deep learning for efficient discriminative parsing (video lecture). `http://videolectures.net/aistats2011_collobert_deep/` [Accessed March 18, 2021].

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug).

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., and Schröder, M. (2000). 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion.*

Creed, C. and Beale, R. (2008). Emotional intelligence: Giving computers effective emotional skills to aid interaction. In *Computational Intelligence: A Compendium*, pages 185–230. Springer.

Cummins, N., Amiriparian, S., Hagerer, G., Batliner, A., Steidl, S., and Schuller, B. W. (2017). An image-based deep spectrum feature representation for the recognition of emotional speech. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 478–484.

Cummins, N., Baird, A., and Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41 – 54. Health Informatics and Translational Data Analytics.

Danks, D. and London, A. J. (2017). Algorithmic bias in autonomous systems. In *IJCAI*, volume 17, pages 4691–4697.

Darwin, C. and Prodger, P. (1872). *The expression of the emotions in man and animals*. John Murray, London.

Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Fourth International Conference on Spoken Language Processing*.

Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4).

Dhall, A. (2019). Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*.

Dhall, A., Goecke, R., Joshi, J., Wagner, M., and Gedeon, T. (2013). Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516.

Dobbe, R., Dean, S., Gilbert, T., and Kohli, N. (2018). A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *Workshop on Fair-*

ness, Accountability and Transparency in Machine Learning (FAT/ML).

Ekman, P. (1970). Universal facial expressions in emotion. *California Mental Health Research Digest*, 8(4).

Ekman, P. (2000). *Handbook of cognition and emotion*, chapter 3. John Wiley &amp; Sons.

El Ayadi, M., Kamel, M. S., and Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3).

Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2).

Eskimez, S. E., Duan, Z., and Heinzelman, W. (2018). Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Eyben, F., Batliner, A., and Schuller, B. (2010a). Towards a standard set of acoustic features for the processing of emotion in speech. In *Proceedings of Meetings on Acoustics 159ASA*, volume 9, page 060006. Acoustical Society of America.

Eyben, F., Batliner, A., Schuller, B., Seppi, D., and Steidl, S. (2010b). Cross-corpus classification of realistic emotions–some pilot experiments. In *Proceedings LREC workshop on Emotion Corpora, Valettea, Malta.*

Eyben, F., Scherer, K. R., Schuller, B., et al. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2).

Eyben, F., Weninger, F., Gross, F., and Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM.

Eyben, F., Wöllmer, M., Poitschke, T., Schuller, B., Blaschke, C., Färber, B., and Nguyen-Thien, N. (2010c). Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. *Advances in human-computer interaction*, 2010.

Eyben, F., Wöllmer, M., and Schuller, B. (2010d). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Fayek, H. M., Lech, M., and Cavedon, L. (2016). On the correlation and transferability of features between automatic speech recognition and speech emotion recognition. In *Proceedings of Interspeech*.

Feraru, S. M., Schuller, D., et al. (2015). Cross-language acoustic emotion recognition: An overview and some tendencies. In *Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Ferras, M., Madikeri, S., Motlicek, P., Dey, S., and Bourlard, H. (2016). A large-scale open-source acoustic simulator for speaker recognition. *IEEE Signal Processing Letters*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. In *ACM international conference on Multimedia*.

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., and Schuller, B. (2017). audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *The Journal of Machine Learning Research*, 18(1).

Friedman, B. (1996). Value-sensitive design. *interactions*, 3(6):16–23.

Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347.

Fürnkranz, J. and Hüllermeier, E. (2010). *Preference learning and ranking by pairwise comparison*, pages 65–82. Springer.

Galindo-Aldana, G. M., Fraga-Vallejo, M., Menchaca-Díaz, R., Alvelais-Alarcón, M., and Machinskaya, R. (2017). Association between risky behaviors in adolescents and altered psychophysiological emotional responses. *Revista de la Facultad de Medicina*, 65(2):183–188.

Gendron, M. and Feldman Barrett, L. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review*, 1(4):316–339.

Ghaleb, E., Popa, M., Hortal, E., and Asteriadis, S. (2017). Multimodal fusion based on information gain for emotion recognition in the wild. In *2017 Intelligent Systems Conference (IntelliSys)*, pages 814–823.

Ghosh, S., Laksana, E., Morency, L.-P., and Scherer, S. (2016a). Learning representations of affect from speech. *International Conference on Learning Representations (ICLR).*

Ghosh, S., Laksana, E., Morency, L.-P., and Scherer, S. (2016b). Representation learning for speech emotion recognition. *Proceedings of Interspeech.*

Gideon, J., Khorram, S., Aldeneh, Z., Dimitriadis, D., and Provost, E. M. (2017). Progressive neural networks for transfer learning in emotion recognition. In *Proceedings of Interspeech.*

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics.*

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. Book in preparation for MIT Press.

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

Graves, A. (2013). Generating sequences with recurrent neural networks. technical report. *arXiv preprint arXiv:1308.0850*.

Green, J. (2012). *The Fault in Our Stars*. Penguin Young Readers Group.

Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136.

Han, J., Zhang, Z., Ren, Z., Ringeval, F., and Schuller, B. (2018). Towards conditional adversarial training for predicting emotions from speech. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada*.

Han, J., Zhang, Z., Ren, Z., and Schuller, B. (2019). Implicit fusion by joint audiovisual training for emotion recognition in mono modality. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5861–5865. IEEE.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of Interspeech*.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 820–828, Red Hook, NY, USA. Curran Associates Inc.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Ho, T. K. and Basu, M. (2000). Measuring the complexity of classification problems. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 43–47. IEEE.

Hozjan, V. and Kačič, Z. (2003). Context-independent multilingual emotion recognition from speech signals. *International journal of speech technology*, 6(3):311–320.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., and Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. In *LREC*.

Huang, C., Gong, W., Fu, W., and Feng, D. (2014). A research of speech emotion recognition based on deep belief network and svm. *Mathematical Problems in Engineering*, 2014.

Inanoglu, Z. and Young, S. (2009). Data-driven emotion conversion in spoken english. *Speech Communication*, 51(3):268–283.

James, W. (1884). What is an emotion? *Mind*, 9(34):188–205.

Jeon, J. H., Le, D., Xia, R., and Liu, Y. (2013). A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception. In *Proceedings of Interspeech*.

Jin, Q., Li, C., Chen, S., and Wu, H. (2015). Speech emotion recognition with acoustic and lexical features. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Jones, C. and Sutherland, J. (2008). Acoustic emotion recognition for affective computer gaming. In *Affect and emotion in human-computer interaction*, pages 209–219. Springer.

Jurafsky, D. and Martin, J. H. (2020). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition (third edition draft). Draft chapters available at `https://web.stanford.edu/~jurafsky/slp3/` [Accessed April 8, 2021].

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. *Proceedings of the 52nd ACL.*

Kaneko, T. and Kameoka, H. (2018). Parallel-data-free voice conversion using cycle-consistent adversarial networks. In *26th European Signal Processing Conference*, Rome, Italy.

Keren, G. and Schuller, B. (2016). Convolutional rnn: an enhanced model for extracting features from sequential data. *2016 International Joint Conference on Neural Networks (IJCNN).*

Kessler, V., Schels, M., Kächele, M., Palm, G., and Schwenker, F. (2015). On the effects of continuous annotation tools and the human factor on the annotation outcome. In *ISCT*, pages 174–180.

Kim, J., Truong, K. P., Englebienne, G., and Evers, V. (2017a). Learning spectro-temporal features with 3d cnns for speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017b). Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference for Learning Representations (ICLR)*.

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Kleinginna, P. R. and Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and emotion*, 5(4):345–379.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

Koolagudi, S. G. and Rao, K. S. (2012). Emotion recognition from speech: a review. *International journal of speech technology*, 15(2).

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621.

Lakomkin, E., Weber, C., Magg, S., and Wermter, S. (2017). Reusing neural speech representations for auditory emotion recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*.

Latif, S., Rana, R., Qadir, J., and Epps, J. (2018). Variational autoencoders for learning latent representations of speech emotion: A preliminary study. In *Proceedings of Interspeech*.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*. Citeseer.

Lefter, I., Rothkrantz, L. J., Wiggers, P., and Van Leeuwen, D. A. (2010). Emotion recognition from speech by combining databases and fusion of classifiers. In *International Conference on Text, Speech and Dialogue*. Springer.

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., and Sahli, H. (2013). Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Lim, W., Jang, D., and Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)*, pages 1–4. IEEE.

Lin, J.-C., Wu, C.-H., and Wei, W.-L. (2013). A probabilistic fusion strategy for audiovisual emotion recognition of sparse and noisy data. In *2013 1st International Conference on Orange Technologies (ICOT)*, pages 278–281. IEEE.

Liu, M.-Y., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 700–708. Curran Associates, Inc.

Lobel, A., Gotsis, M., Reynolds, E., Annetta, M., Engels, R. C., and Granic, I. (2016). Designing and utilizing biofeedback games for emotion regulation: The case of nevermind. In

Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, pages 1945–1951.

Ludden, D. (2015). Hearing with our eyes, seeing with our ears. `https://www.psychologytoday.com/intl/blog/talking-apes/201511/hearing-our-eyes-seeing-our-ears` [Accessed March 14, 2021].

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*.

Machine Elf 1735 (2011). Robert plutchik's wheel of emotions. `https://commons.wikimedia.org/wiki/File:Plutchik-wheel.svg`, Public domain, via Wikimedia commons [Accessed January 5, 2021].

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2016). Adversarial autoencoders. In *Proceedings of International Conference on Learning Representations (ICLR)*.

Mallol-Ragolta, A., Schmitt, M., Baird, A., Cummins, N., and Schuller, B. (2019). Performance analysis of unimodal and multimodal models in valence-based empathy recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–5. IEEE.

Mao, Q., Dong, M., Huang, Z., and Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8).

Marr, B. (2019). Artificial intelligence explained: What are generative adversarial networks (gans)? Forbes Magazine, `https://www.forbes.com/sites/bernardmarr/2019/06/12/artificial-intelligence-explained-what-are-generative-adversarial-networks-gans` [Accessed April 13, 2021].

Matton, K., McInnis, M. G., and Provost, E. M. (2019). Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. In *Proceedings of Interspeech*.

Meng, F., Lu, Z., Wang, M., Li, H., Jiang, W., and Liu, Q. (2015). Encoding source language with convolutional neural network for machine translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2).

Mohamed, A.-r. (2014). *Deep Neural Network Acoustic Models for ASR*. PhD thesis, University of Toronto.

Morgan, N. and Bourlard, H. (1989). Generalization and parameter estimation in feedforward nets: Some experiments.

In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, pages 630–637.

Moyers, B. and Flowers, B. (1989). *A World of Ideas: Conversations with Thoughtful Men and Women about American Life Today and the Ideas Shaping Our Future.* Doubleday.

Neumann, M. and Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In *Proceedings of Interspeech.*

Neumann, M. and Vu, N. T. (2018). Cross-lingual and multilingual speech emotion recognition on english and french. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP).*

Neumann, M. and Vu, N. T. (2019). Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton.

Neumann, M. and Vu, N. T. (2021). Investigations on audiovisual emotion recognition in noisy conditions. In *IEEE Spoken Language Technology Workshop (SLT)*, virtual.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848.

Palaz, D., Collobert, R., et al. (2015). Analysis of cnn-based speech recognition system using raw speech as input. In *Proceedings of Interspeech*.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Pandit, V., Schmitt, M., Cummins, N., Graf, F., Paletta, L., and Schuller, B. (2018). How good is your model 'really'? on 'wildness' of the in-the-wild speech-based affect recognisers. In *International Conference on Speech and Computer*. Springer.

Pantic, M. and Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390.

Paralinguistics (2020). Definition of paralinguistics. Lexico - Oxford UK Dictionary, `https://www.lexico.com/definition/paralinguistics` [Accessed December 13, 2020].

Parmar, M. (2019). Microsoft's cortana might respond better to your feelings in future. `https://www.windowslatest.com/2019/07/30/microsofts-cortana-might-respond-better-to-your-feelings-in-future/` [Accessed April 23, 2021].

Parthasarathy, S. and Busso, C. (2018). Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes. In *Proceedings of Interspeech*.

Pascual, S., Bonafonte, A., and Serra, J. (2017). Segan: Speech enhancement generative adversarial network. *Proceedings of Interspeech*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32*.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering*, volume 710, page 22.

Picard, R. W. (1995). Affective computing. Technical Report 321, MIT Media Lab, Perceptual Computing Group.

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., and Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Polzehl, T., Schmitt, A., and Metze, F. (2010). Approaching multi-lingual emotion recognition from speech-on language dependency of acoustic/prosodic features for anger detection. *Speech Prosody 2010 - Fifth International Conference.*

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 469–481.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., 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.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., et al. (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*, pages 3–12.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture*

Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE.

Rousseau, A., Deléglise, P., and Esteve, Y. (2014). Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Rozgic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., and Prasad, R. (2012). Ensemble of svm trees for multimodal emotion recognition. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *IJCV*.

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

Sagha, H., Matejka, P., Gavryukova, M., Povolny, F., Marchi, E., and Schuller, B. (2016). Enhancing multilingual recognition of emotion in speech by language identification. *Proceedings of Interspeech*.

Sahu, S., Gupta, R., and Espy-Wilson, C. (2018). On enhancing speech emotion recognition using generative adversarial networks. In *Proceedings of Interspeech*.

Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., and Espy-Wilson, C. (2017). Adversarial auto-encoders for speech based emotion recognition. In *Proceedings of Interspeech*.

Sainath, T. and Parada, C. (2015). Convolutional neural networks for small-footprint keyword spotting. In *Proceedings of Interspeech*.

Sainath, T. N., Kingsbury, B., Saon, G., Soltau, H., Mohamed, A.-r., Dahl, G., and Ramabhadran, B. (2015). Deep convolutional neural networks for large-scale speech tasks. *Neural Networks*, 64.

Sainath, T. N., Mohamed, A.-r., Kingsbury, B., and Ramabhadran, B. (2013). Deep convolutional neural networks for lvcsr. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*.

Schmidhuber, J. (2020). Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). *Neural Networks*, 127:58–66.

Schuller, B., Arsic, D., Wallhoff, F., and Rigoll, G. (2006). Emotion recognition in the noise applying large acoustic feature sets. In *Proceedings Speech Prosody 2006, Dresden.*

Schuller, B. and Batliner, A. (2013). *Computational paralinguistics: emotion, affect and personality in speech and language processing.* John Wiley & Sons.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011a). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9).

Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. IEEE.

Schuller, B., Seppi, D., Batliner, A., Maier, A., and Steidl, S. (2007). Towards more reality in the recognition of emotional speech. In *2007 IEEE international conference on acoustics, speech and signal processing-ICASSP'07*, volume 4, pages IV–941. IEEE.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C., and Narayanan, S. (2013). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C. A., Narayanan, S. S., et al. (2010a). The interspeech 2010 paralinguistic challenge. In *Proceedings of Interspeech*.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011b). Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009a). Acoustic emotion recognition: A benchmark comparison of performances. In *Automatic Speech Recognition and Understanding (ASRU), Workshop on*. IEEE.

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

Schuller, B., Wimmer, M., Mosenlechner, L., Kern, C., Arsic, D., and Rigoll, G. (2008). Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4501–4504. IEEE.

Schuller, B., Zhang, Z., Weninger, F., and Rigoll, G. (2011c). Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization. In *Proceedings 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel.*

Schuller, B. W., Steidl, S., Batliner, A., et al. (2009b). The interspeech 2009 emotion challenge. In *Proceedings of Interspeech.*

Schuller, B. W., Zhang, Z., Weninger, F., and Rigoll, G. (2011d). Using multiple databases for training in emotion recognition: To unite or to vote? In *Proceedings of Interspeech.*

Sebe, N., Cohen, I., Gevers, T., and Huang, T. S. (2005). Multimodal approaches for emotion recognition: a survey. In Santini, S., Schettini, R., and Gevers, T., editors, *Internet Imaging VI*, volume 5670, pages 56 – 67. International Society for Optics and Photonics, SPIE.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR.*

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1).

Taigman, Y., Polyak, A., and Wolf, L. (2017). Unsupervised cross-domain image generation. In *International Conference for Learning Representations (ICLR)*.

Tao, J., Kang, Y., and Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1145–1154.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2017). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. Technical report, IEEE.

Tian, L., Moore, J. D., and Lai, C. (2015). Emotion recognition in spontaneous and acted dialogues. In *Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Triantafyllopoulos, A., Keren, G., Wagner, J., Steiner, I., and Schuller, B. W. (2019). Towards robust speech emotion recognition using deep residual networks for speech enhancement. In *Proceedings of Interspeech*, pages 1691–1695.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8).

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE transactions on acoustics, speech, and signal processing*, 37(3).

Wand, M., Schmidhuber, J., and Vu, N. T. (2018). Investigations on end-to-end audiovisual fusion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3041–3045. IEEE.

Wang, Y. and Guan, L. (2004). An investigation of speech-based human emotion recognition. In *IEEE 6th Workshop on Multimedia Signal Processing, 2004.*, pages 15–18. IEEE.

Williams, F. and Sundene, B. (1965). Dimensions of recognition: Visual vs. vocal expression of emotion. *AV communication review*, 13(1):44–52.

Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of Interspeech incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*, pages 597–600.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., and Rigoll, G. (2013). Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163.

Wundt, W. (1897). Outlines of psychology (c.h. judd, trans.). *Leipzig: Wilhelm Engelmann*, 1:14.

Xia, R. and Liu, Y. (2015). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on Affective Computing*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

Xue, W., Huang, Z., Luo, X., and Mao, Q. (2015). Learning speech emotion features by joint disentangling-discrimination. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 374–379. IEEE.

Yannakakis, G. N. (2009). Preference learning for affective modeling. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE.

Yannakakis, G. N., Cowie, R., and Busso, C. (2018). The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*.

Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation.

In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876. IEEE.

Yin, W., Schütze, H., Xiang, B., and Zhou, B. (2016). Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*.

You, M., Chen, C., Bu, J., Liu, J., and Tao, J. (2006). Emotion recognition from noisy speech. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1653–1656. IEEE.

Young, P. T. (1973). Feeling and emotion. In *Handbook of General Psychology*. Prentice Hall: Englewood Cliffs, New Jersey.

Yu, D. and Deng, L. (2016). *Automatic Speech Recognition. A Deep Learning Approach*. Springer.

Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., and Schuller, B. (2016). Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks. In *Proceedings of Interspeech*.

Zhang, Z., Weninger, F., Wöllmer, M., and Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE.

Zhao, X., Zhang, S., and Lei, B. (2014). Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications*, 24(7-8):1539–1553.

Zhao, Y., Jin, X., and Hu, X. (2017). Recurrent convolutional neural network for speech processing. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5300–5304. IEEE.

Zheng, W., Yu, J., and Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Affective Computing and Intelligent Interaction (ACII)*. IEEE.

Zhou, F. and De la Torre, F. (2015). Generalized canonical time warping. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):279–294.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232.