OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INSTITUT FÜR INFORMATIONS- UND
KOMMUNIKATIONSTECHNIK (IIKT)

Driver Affect Recognition from Real-World Speech Data in In-Vehicle Driving Environments

DISSERTATION

zur Erlangung des akademischen Grades
Doktoringenieurin (Dr.-Ing.)

von

Alicia Flores Requardt geb. Lotz, M. Sc.

geb. am 29.05.1990 in Gießen

genehmigt durch die
Fakultät für Elektrotechnik und Informationstechnik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Herr Prof. Dr. rer. nat. Andreas Wendemuth
Herr Prof. Dr.-Ing. Ulrich Jumar
Frau Prof. Anna Anund

Promotionskolloquium am 16.12.2022

„Don't go wasting your emotion"
- ABBA

Zusammenfassung

In der heutigen Zeit der zunehmenden Autonomie im Straßenverkehr, gewinnen Systeme zur Erkennung des Fahrerzustandes immer mehr an Bedeutung. Über die Anaylse des Lenkverhaltens und des Blickverhaltens des Fahrers können einige Zustände bereits zum Stand der Technik in der Automobilindustrie gezählt werden. Der tatsächliche intrinsische Zustand des Fahrers, der z.B. durch Feedbacksignale aus Sprache, Mimik oder Gestik wiedergegeben werden kann, wird momentan noch vollständig außer Acht gelassen.

Ziel dieser Arbeit ist es, diese Forschungslücke weiter zu schließen, indem die Emotionalität des Fahrers anhand seiner Sprache erkannt und systemseitig berücksichtigt werden kann. Um dieses Ziel zu erreichen, muss die vollständige Prozesskette von der Datenerhebung, über die Datenvoranalyse und ggf. die Durchführung digitaler Signalverarbeitungs-Schritte, bis hin zur Datenklassifizierung und schlussendlich der Bewertung der erhaltenen Ergebnisse im Vergleich zu anderen Forschungsergebnissen aus diesem Bereich, berücksichtigt werden. Die Gesamtheit jedes einzelnen Prozessschrittes muss hierzu dem Leser nähergebracht werden. Dies begründet auch den Umfang der vorliegenden Arbeit.

Zu Beginn der Arbeit werden dem Leser folgende drei Forschungshypothesen vorgestellt, die im Laufe der Arbeit wiederholt aufgegriffen werden:

1. **Hypothese:** Es ist möglich dem Fahrer während der Fahrt naturalistische Emotionen zu induzieren.

2. **Hypothese:** Es ist möglich Störungen des Sprachsignales zu kompensieren.

3. **Hypothese:** Falls Hypothese 1 und 2 bestätigt werden, ist es möglich automatisch den emotionalen Zustand des Fahrers anhand prosodischer Sprachmerkmale zu erkennen.

Durch die sehr geringe Datenlage in diesem Forschungsgebiet wurden zwei Datenaufnahmen durchgeführt (simuliert und real). Anhand erster simulierter emotionaler Sprachdaten im Fahrzeug konnten erste Erkenntnisse über die Beschaffenheit der Daten und das Potential zur Erkennung des emotionalen Zustandes, erlangt werden. Anhand einer weiteren Datenaufnahme, induzierter Emotionen unter realen Fahrbedingungen, konnten Detailinformationen zur Erkennung von vier Fahrerzuständen ermittelt werden (*neutral*, *positiv*, *verärgert* und *ängstlich*). Die aufgenommenen Daten wurden, anhand der Selbsteinschätzung des Fahrers (unter Verwendung des *Geneva Emotional Wheel* und den *Self-Assessment Manikins*) und einer Auswertung ihrer bio-physiologischen Daten, auf ihre emotionalen Inhalte und ihre Verwendbarkeit validiert.

Da realitätsnahe verrauschte Sprachdaten sehr zeitaufwändig in ihrer Generierung sind und zu Beginn der Arbeit noch nicht vorlagen, wurden erste Untersuchungen anhand komprimierter Sprachdaten durchgeführt. Anhand dieser Daten wurde der Effekt digitaler Signalverarbeitungs-Algorithmen auf das Sprachsignal, die Sprach- und Signalqualität und die Erkennung der Emotionen analysiert. Es konnte festgestellt werden, dass die angewendeten Audio-Codecs je nach ihrem designierten Einsatzgebiet unterschiedliche Einflüsse auf die Sprach- und Signalqualität und die Erkennungsleistung der Emotionen haben. Vor allem Codecs, die für die Komprimierung von Musik entwickelt wurden, haben einen geringeren Einfluss auf die Emotionserkennung als Codecs, die für die Telekommunikation entwickelt wurden. Im Fall der Anwendung von Musik-Codecs konnte sogar eine Verbesserung der Erkennungsleistung im Vergleich zu unkomprimierten Sprachdaten erzielt werden. Ähnliche Untersuchungen wurden anhand der simulierten emotionalen Sprache im Fahrzeug durchgeführt, indem die im Original unverrauschten Sprachdaten mit ihren künstlich verrauschten Versionen verglichen wurden. Diese Untersuchung zeigte, dass die Natürlichkeit der Emotionen in der Sprache und die Natürlichkeit der Datenaufnahmen selbst, den Effekt der Fahrgeräusche auf die Signalqualität beeinflusst. Des Weiteren konnte ein eindeutiger Rückgang der Erkennungsleistung im Zusammenhang mit der Abnahme der Signalqualität erkannt werden.

Zur weiteren Nutzung der Datenaufnahmen im realen Fahrzeugumfeld wurde eine Annotation der Daten durchgeführt. Dies beinhaltete eine dimensionale und kategoriale Bewertung, die in sich eine hohe Übereinstimmung aufwiesen. Die Ergebnisse zeigen, dass die Emotionen des Fahrers auch in seiner Sprache widergespiegelt wurden und somit als Datenbasis für die automatische Erkennung natürlicher Emotionen im Fahrzeugumfeld genutzt werden können.

Aus der (automatischen) Sprachverarbeitung ist bekannt, dass die Anwendung von Sprachverbesserungs-Verfahren (engl. speech enhancement) zu einer bemerkenswerten Erhöhung der Erkennungsraten und des Sprachverständnisses führen kann. Um zu untersuchen, ob dieser Effekt auch einen Einfluss auf die Erkennungsrate der Emotionen hat, wurden Untersuchungen zur Anwendbarkeit dieser Algorithmen auf verrauschter emotionaler Sprache durchgeführt. Es konnte festgestellt werden, dass die Anwendung dieser Verfahren zu einer starken Manipulation des Merkmalsraums führt, die im Vergleich zu verrauschter Sprache jedoch keine Verbesserung der Erkennungsleistung mit sich bringt. Um die Manipulation des Merkmalsraum nicht als zusätzlichen Freiheitsgrad in die Prozesskette mit einfließen zu lassen, wird die Anwendung eines solchen Verfahrens nicht empfohlen.

Unter Berücksichtigung der vorangegangenen Ergebnisse wurden schlussendlich zwei unterschiedliche Klassifikationsverfahren angewandt (Support Vector Machines und Random Forests), um die Emotion des Fahrers anhand realer Fahrzeugdaten zu erkennen. Die Klassifizierer wurden dazu in einem *leave one speaker out* Kreuzvalidierungsverfahren trainiert, optimiert und getestet. Die Optimierung erfolgte dabei durch die Anwendung eines *random search*-Verfahrens zur Hyper-Parameter-Optimierung, einer *wrapper* basierten Feature Auswahl und einer gezielten Reduzierung/ Auswahl der verwendeten Sprachdaten aus dem Datensatz. Unter Berücksichtigung all dieser Aspekte, konnte als bester Klassifizierer ein Random Forest entworfen werden, der dazu in der Lage ist, den emotionalen Zustand des Fahrers, im vorliegenden 4-Klassenproblem, mit einer *precision* von über 52% und einem *recall* von über 35% zu erkennen.

Abstract

With an increase of autonomy in vehicles, also the importance of driver state detection systems is becoming more relevant. By analyzing the driver's steering behaviour and her/ his gaze direction, the modern automotive industry is able to detect a limited number of driver states (e.g. tiredness or attention). The true intrinsic state of the driver, which is, for example, communicated through feedback signals in her/ his speech, facial expressions or gestures, is still being neglected.

The goal of this Thesis is to close this research gap by considering the driver's speech data to detect her/ his emotional state. This does not only include the design of a classifier, but the whole process chain of performing a suitable data collection, pre-processing of said data, implementation of relevant signal processing steps (e.g. speech enhancement) and finally also validating the designed classifier. This broad field of research also reasons the size of the Thesis.

At the beginning of the Thesis the following three research hypotheses are introduced to the reader and will accompany her/ him throughout the Thesis:

1. **Hypothesis:** It is possible to induce naturalistic emotions in the driver, while driving in a real vehicle.

2. **Hypothesis:** It is possible to compensate effects of speech distortion.

3. **Hypothesis:** Under the assumption that hypotheses 1 and 2 apply, it is possible to automatically detect the emotional state of the driver by only considering the speech signal of the driver and its prosodic features.

Because of the relatively low amount of freely available emotional speech data in in-vehicle environments, two data collections focusing on this noise environment (simulated and real-world) were performed. The simulated data was used to receive first insights on the noisy speech characteristics and its potential to be used to detect the driver's emotional state. A second real-world data collection was performed afterwards, and used to gain detailed information on the four most relevant emotional states occurring while driving (*neutral*, *positive*, *angry* and *anxious*). By using the drivers' self-reports (obtained by utilizing the *Geneva Emotional Wheel* and the *Self-Assessment Manikins*) and the recordings of their bio-physiological parameters, it was possible to validate the emotion inducement method and the usability of the collected real-world data.

The just mentioned data collections are highly time consuming to conduct and were not available at the start of the Thesis. Therefore, the first investigation presented in this Thesis, was performed on compressed speech data. This degraded data was used to analyze the effects signal-processing can have on the speech signal itself, the

signal quality and the ability to correctly classify the emotional state of a speaker. It was identified that, especially for speech emotion recognition, codecs developed for music compression are more suitable than codecs developed for speech compression. In some cases it was even possible to increase the recognition performance by applying music compression algorithms, compared to the recognition performance on uncompressed speech. Similar investigations on noisy speech were performed on the simulated in-vehicle speech data. By comparing the original emotional speech samples with their degraded noisy counter parts, it was possible to identify that the naturalness of the original speech samples plays a decisive role on the effect in-vehicle noises have on the signal quality. Furthermore, with decreasing signal quality also the recognition performance of the classifier decreased.

To verify the usability of the real-world driving data, a further annotation of the speech samples considering their emotional content was needed. This annotation was done utilizing a dimensional (valence vs. arousal) and a categorial (4 considered emotional states) labeling approach. In this process both approaches showed a high consistency in their results. These results show that the emotional state of the driver is also mirrored in the speech signal and that the recorded data is suitable for automatic speech emotion recognition in a real-world driving environment.

In case of noisy speech data and (automatic) speech recognition, it is known that by applying speech enhancement algorithms, significant increases in recognition rate and speech understanding can be achieved. To identify if these effects also occur in case of speech emotion recognition, suitable speech enhancement algorithms were applied to the simulated in-vehicle data. It was identified that by applying this method of signal processing steps to the noisy speech samples, the features used for the speech emotion recognition task were altered significantly but the recognition performance was not improved. To prevent this additional factor from influencing the emotion recognition task, it was decided to not apply speech enhancement in the further scope of the Thesis.

Finally, by considering and utilizing the above findings, two classification approaches (*Support Vector Machines* and *Random Forest*) were used to identify the driver's emotional state in a real-world driving scenario. By utilizing a *leave one speaker out* cross-validation scheme the classifiers were trained, optimized and tested. The optimization step included a hyper-parameter optimization using *random search*, a *wrapper* based feature selection and an adjusted of the data set, by reducing the data set to a tailored selection of speech samples. With regard to this approach, as best performing classifier for the present four class classification task, a random forest with a *precision* of over 52% and a *recall* of over 35% was designed.

References Related to the Author

Böck, R.; Egorow, O.; Höbel-Müller, J.; Requardt, A. F.; Siegert, I. & Andreas, W. (July 2019). 'Anticipating the User: Acoustic Disposition Recognition in Intelligent Interactions'. In: *Innovations in Big Data Mining and Embedded Knowledge.* Ed. by Esposito, A.; Esposito, A. M. & Jain, L. C. Vol. 159. Intelligent Systems Reference Library (ISRL). Springer, Cham, pp. 203–233.

Egorow, O.; Lotz, A.; Siegert, I.; Böck, R.; Krüger, J. & Wendemuth, A. (2017). 'Accelerating manual annotation of filled pauses by automatic pre-selection'. In: *2017 International Conference on Companion Technology (ICCT 2017).* Ulm, Germany: Institute of Electrical and Electronics Engineers (IEEE), s.p.

Höbel-Müller, J.; Siegert, I.; Heinemann, R.; Requardt, A. F.; Tornow, M. & Wendemuth, A. (2019). 'Analysis of the Influence of Different Room Acoustics on Acoustic Emotion Features'. In: *Elektronische Sprachsignalverarbeitung 2019: Tagungsband der 30. Konferenz.* Ed. by Birkholz, P. & Stone, S. Vol. 93. TUDpress, pp. 156–163.

Lotz, A. F.; Faller, F.; Siegert, I. & Wendemuth, A. (2018). 'Emotion Recognition from Disturbed Speech - Towards Affective Computing in Real-World In-Car Environments'. In: *Elektronische Sprachsignalverarbeitung 2018.* Ed. by Berton, A.; Haiber, U. & Minker, W. Vol. 90. Studientexte zur Sprachkommunikation. TUDpress, pp. 208–215.

Lotz, A. F.; Ihme, K.; Charnoz, A.; Maroudis, P.; Dmitriev, I. & Wendemuth, A. (2018). 'Recognizing Behavioral Factors while Driving: A Real-World Multimodal Corpus to Monitor the Driver's Affective State'. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Ed. by Calzolari, N. et al. Miyazaki, Japan: European Language Resource Association (ELRA), pp. 1589–1596.

Lotz, A. F.; Siegert, I.; Maruschke, M. & Wendemuth, A. (2017). 'Audio Compression and its Impact on Emotion Recognition in Affective Computing'. In: *Elektronische Sprachsignalverarbeitung 2017.* Ed. by Trouvain, J.; Steiner, I. & Möbius, B. Vol. 86. Studientexte zur Sprachkommunikation. TUDpress, pp. 1–8.

Requardt, A. F.; Egorow, O. & Wendemuth, A. (2020). 'Machine Learning-Assisted Affect Labelling of Speech Data'. In: *Elektronische Sprachsignalverarbeitung 2020.* Ed. by Wendemuth, A.; Böck, R. & Siegert, I. Vol. 95. Studientexte zur Sprachkommunikation. Best Student Paper Award. TUDpress, pp. 199–226.

Requardt, A. F.; Ihme, K.; Wilbrink, M. & Wendemuth, A. (2020). ‘Towards affect-aware vehicles for increasing safety and comfort: recognising driver emotions from audio recordings in a realistic driving study’. *IET Intelligent Transport Systems* 14.10, pp. 1265–1277.

Requardt, A. F.; Wilbrink, M.; Siegert, I.; Jipp, M.; Wendemuth, A. & Ihme, K. (2018). ‘An Experimental Paradigm for Inducing Emotions in a Real World Driving Scenario - Evidence from Self-Report, Annotation of Speech Data and Peripheral Physiology’. *Kognitive Systeme* 2018 (1), pp. 1–11.

Siegert, I.; Lotz, A. F.; Duong, L. L. & Wendemuth, A. (2016). ‘Measuring the Impact of Audio-Compression on the Spectral Quality of Speech Data’. In: *Elektronische Sprachsignalverarbeitung 2016.* Ed. by Jokisch, O. Vol. 81. Studientexte zur Sprachkommunikation. TUDpress, pp. 229–236.

Siegert, I.; Lotz, A. F.; Egorow, O.; Böck, R.; Schega, L.; Tornow, M.; Thiers, A. & Wendemuth, A. (2016). ‘Akustische Marker für eine verbesserte Situations- und Intentionserkennung von technischen Assistenzsystemen’. In: *Proc. of the 2nd Transdisziplinäre Konferenz "Technische Unterstützungssysteme, die die Menschen wirklich wollen".* SmartASSIST. Hamburg, Germany: Smart, AdjuStable, Soft and Intelligent Support Technologies (SmartASSIST), pp. 465–474.

Siegert, I.; Lotz, A. F.; Egorow, O. & Wendemuth, A. (2017). ‘Improving Speech-Based Emotion Recognition by Using Psychoacoustic Modeling and Analysis-by-Synthesis’. In: *Speech and Computer (SPECOM 2017).* Ed. by Karpov, A.; Potapova, R. & Mporas, I. Vol. 10458. Lecture Notes in Computer Science. Springer, Cham, pp. 445–455.

Siegert, I.; Lotz, A. F.; Egorow, O. & Wolff, S. (2018). ‘Utilizing Psychoacoustic Modeling to Improve Speech-Based Emotion Recognition’. In: *Speech and Computer (SPECOM 2018).* Ed. by Karpov, A.; Jokisch, O. & Potapova, R. Vol. 11096. Lecture Notes in Computer Science. Springer, Cham, pp. 625–635.

Siegert, I.; Lotz, A. F.; Maruschke, M.; Jokisch, O. & Wendemuth, A. (2016). ‘Emotion Intelligibility within Codec-Compressed and Reduced Bandwidth Speech’. In: *Proc. of the Speech Communication 12. ITG Symposium.* Paderborn, Germany: VDE Verlag GmbH, pp. 1–5.

Contents

List of Figures

List of Tables

Abbreviations

3GPP	3rd Generation Partnership Project
AAC	Advanced Audio Coding
AbS	Analysis-by-Synthesis
ACC	Accuracy
ACELP	Algebraic Codec-Excited Linear Prediction (CELP)
ADAS	Advanced Driver Assistant Systems
AL	Active Learning
AMR	Adaptive Multi-Rate
AMR-WB	Adaptive Multi-Rate (AMR) Wideband
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
Bagging	Bootstrap Aggregation
BSD	Bark Distortion Measure
CART	Classification and Regression Trees
CELP	Codec-Excited Linear Prediction
CELT	Constrained Energy Lapped Transform
CER	Compression Error Rate
CFS	Correlation-based Feature Selection
DES	Danish Emotional Speech Database
DLR	German Aerospace Center
DNN	Deep Neural Network
DSS	Decision Support System
DTFT	Discrete-Time Fourier Transform
DTW	Dynamic Time Warping
ECG	electrocardiogram
EEG	electroencephalogram
EmoDB	Berlin Emotional Speech Database
EmoDB-Car	re-recorded EmoDB under in-car recording conditions
ER	Error Rate
ETSI	European Telecommunication Standards Institute

Euro NCAP	European New Car Assessment Programme
FDR	Fisher's discriminant ratio
FLAC	Free Lossless Audio Codec
FN	False Negative
fNIRS	functional Infrared Spectroscopy
FP	False Positive
FT	Finger Temperature
FWER	Family-Wise Error Rate
GEW	Geneva Emotion Wheel
GMM	Gausian Mixture Model
GSR	General Safety Regulations
HCI	Human-Computer Interaction
HHI	Human-Human Interaction
HMI	Human-Machine Interface
HMM	Hidden Markov Model
HR	Heart Rate
ICA	Independent Component Analysis
ICC	Intra-Class-Correlation
IMCRA	Improved Minima Controlled Recursive Averaging
IRR	Inter-Rater-Reliability
LDA	Linear Discriminant Analysis
LLD	Low-Level Descriptor
LOSGO	Leave-One-Subject-Group-Out
LOSO	Leave-One-Subject-Out
LPC	Linear Predictive Coding
LSP	Line Spectral Pair
LSTM	Long Short-Term Memory
MDCT	Modified Discrete Cosine Transforms
MFCC	Mel-Frequency Cepstral Coefficient
MOS	Mean Opinion Score
MOS-LQO	MOS - Listening Quality Objective
MP3	MPEG-1/MPEG-2 Audio Layer-3
MPEG	Moving Picture Experts Group
MR-ACELP	Multi-Rate Algebraic CELP (ACELP)

NB	narrowband
NTSB	National Transportation Safety Board
OM-LSA	Optimally-Modified Log-Spectral Amplitude
OOB	Out-Of-Bag
PCA	Principle Component Analysis
PL	Passive Learning
PLP	Perceptual Linear Prediction
POLQA	Perceptual Objective Listening Quality Assessment
PSD	Power Spectral Density
RBF	Radial Basis Function
RF	Random Forest
RIFF	Research Interchanged File Format
RMS	Root Mean Square
SAE	Society of Automotive Engineers
SAM	Self Assessment Manikins
SCL	Skin Conductance Level
SEP	Smart Eye Pro
SII	Speech Intelligibility Index
SNR	Signal-to-Noise Ratio
SPX	Speex
SSL	Semi-Supervised Learning
STFT	Short-time Fourier Transform
SVM	Support Vector Machine
SVR	Support Vector Regression
SWB	super-wideband
TEO	Teager Energy Operator
TN	True Negative
TP	True Positive
UAP	Unweighted Average Precision
UAR	Unweighted Average Recall
VAM	Vera am Mittag
VAM-Car	re-recorded VAM under in-car recording conditions
VoIP	Voice over Internet Protocol

VoLTE	Voice over LTE
WAV	Waveform Audio File
WMA	Windows Media Audio
WoZ	Wizard-of-Oz
ZCR	Zero-Crossing-Rate

Chapter 1

Introduction

Contents

IN today's automotive industry a focus is drawn on increasing the safety of driving. The developed safety functions mainly cover the field of preventing critical/ dangerous driving situations by monitoring the vehicle itself and the surrounding environment (e.g. other road users or environmental conditions). The inter-individual differences of the drivers themselves, however, are still less considered, even though, it is widely known that the maturity of the driver plays a decisive role. When it comes to accessing the maturity of a driver a focus is often drawn on the driving practice, driving experience, involved accidents, number of traffic violations and the driver's self-evaluation over a given evaluation period. These measures however, do not contribute actively to a safer way of driving, as they give information in retrospect. To be able to also actively include the driver's inter-individual behavioural differences an online monitoring of the driver himself is needed. These so-called driver state monitoring systems include a recognition of the driver state and a constant tracking of changes of the driver's behavioural factors. By tracking these changes it is possible to identify differences in the driver's behavioural pattern and actively inform the driver about said changes.

One area of interest, which is recently receiving increased attention, is the monitoring of the driver's state with an increase in automated driving functions, as the driver is less involved in the driving task, but still holds the full responsibility. A recent statistics performed on data of the year 2019 shows that the number of Advanced Driver Assistant Systems (ADAS) has increased substantially in case of newly registered passenger cars compared to already registered cars. In case of drowsiness detection, one of the most prominent driver states, it is shown that the percentage of vehicles equipped with these systems in newly registered cars is twice as high (30%) as in registered used cars (15%) (cf. [Kords 2021]). This is also due to the General

Safety Regulations (GSR) approved by the EU-Parliament in 2019, which standardizes the presence of a drowsiness detection system in newly certified vehicle types from 2022 onwards and newly register vehicles from 2024 onwards. The drivers' affect and consequently their emotions, however, are still less considered. *Affect*, in the machine recognition literature (cf. [Picard 1997]), is considered in a broad sense to cover all conscious subjective aspects such as affection, passion, sensation, inclination, intention, inward disposition or feeling. *Emotion* is used in a narrower sense, as an invoked sensation reaction, primarily as basic emotions [Plutchik 1958] or such emotions which can be composed from basic emotions. Affects and emotions can play a decisive role when it comes to the comfort of the driver and, even more relevant, the safety of driving manually and automated. The methods applied to detect the driver's affective state are diverse and range from the recognition of facial expressions, over assessing bio-physiological signals to speech emotion recognition.

The mentioned methodologies hold different advantages and disadvantages, which need to be identified beforehand. A vision on how these kind of modalities can be used to identify critical driver states and mitigate the effect of incapacitated drivers was investigated in the research Project ADAS&ME [1], which was successfully reviewed and completed in 2020. The goal of the project was to develop a driver state detection system for multiple driver states (i.e. fatigue/ drowsiness, stress, inattention/ distraction and impairing emotions) in combination with multimodal, user oriented interaction strategies. Among others, this included the development of algorithms (e.g. for the detection of the individual driver states), sensing technologies (e.g. electrocardiogram (ECG)-steering wheel and ECG-seat), supportive technologies (e.g. vehicle automation and V2X-communication) and Human-Machine Interface (HMI)-components. As part of this research project, results were achieved in driver state emotion and in the development of the speech based detection systems as well as the late-fusion of the individual driver states. The research conducted for these topics, and the publishing and publicizing of the relevant results, were performed under the responsibility and lead, and with major contribution, of the author of this Thesis.

A great benefit of recognizing the emotional state from the drivers' speech, by using prosodic features only, is the potential to be used in a cross-cultural and cross-lingual setting. Especially when evaluating facial expression, this is not always the case, as emotions may be communicated differently throughout cultures (cf. [Jack et al. 2009]). Furthermore, the environmental conditions of in-vehicle emotion recognition are highly limited by the vehicle and the driving task itself. In many cases not the full frontal face of the driver may be in the field of view of the camera. This makes it challenging to analyze the facial expression of the driver. It further has been shown that facial expressions can be easily manipulated by the subject

[1] https://www.adasandme.com/, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 688900

himself and may not per se give insight into the actual intrinsic state of the driver. Considering bio-physiological signals, the main disadvantage lies in recent sensing technologies, as they are mainly based on body attached sensor systems. Especially with a focus on usability, these intrusive measurements are not seen as user-friendly and are rather inconvenient. With an increased connectivity between the vehicle and the driver, for example, by connected consumer electronics, an excess to the user's health tracker may be possible in the future. Until now, this connectivity is strongly limited to a number of manufacturers and cannot be seen as standard. Furthermore, the development of non-body attached sensor systems is highly cost consuming, which will increase the purchasing price of such vehicles and make them unaffordable for a majority of buyers (cf. [Calem 2019]). Therefore, as a preferred modality, this Thesis will focus on the detection of the driver's emotional state from speech.

Emotion recognition from speech has already received increased attention in the past years. While we have now reached a state where emotions in laboratory environments can be recognized with high recognition performances, the application of speech emotion recognition systems in everyday "in the wild" situations is still very challenging. Depending on the application domain the requirements on the speech emotion recognition system can vary strongly, as the environment in which the speech signal needs to be processed can be seen as non-static. This specific processing environment of in-vehicle speech will be evaluated in the scope of this Thesis, under the assumption that the driver is interacting with a co-driver or, for example, an integrated infotainment system, such that the speech signal is available and processable. With regard to this assumption, I will now present to the reader the three main motivational aspects the Thesis is based on (cf. Sections 1.1 - 1.3), as well as the three identified main research hypotheses (cf. Section 1.4).

1.1 Safety in Automated Driving

In these times and days the field of autonomous driving and ADAS has gained increased interest. A forecast of the global autonomous car market size from 2019 to 2023 indicates an increase from 24.10 billion US $ to 37.22 billion US $ (cf. [Statista 2021b]). With a focus on ADAS-functions, a growth from 17.6 billion US $ in 2020 to nearly 32 billion US $ by 2023 may be reached (cf. [Statista 2021a]). In the automotive industry a trend is observable from fully manual driving (standardized automation level by the Society of Automotive Engineers (SAE): no automation - SAE Level 0) over semi-automated driving (Partial automation - SAE Level 2) to fully automated driving (Full automation - SAE Level 5) (cf. [SAE 2018], for detailed information on the SAE Levels of driving automation). In Europe, we have now reached a state of conditional automation (SAE Level 3). At this state, the System is fully taking over the transverse and longitudinal guidance of the vehicle and the monitoring of the environment, which is, for SAE Level 2 and lower, in control of the human driver. However, in contrast to high or full automation, the human driver serves as the fallback solution in case of automation outage and needs to be physically able to take back the control from the automation function.

Especially the “Tesla incident” in 2018, where a 38 year old Tesla driver died from major injuries after his car crashed into a non-operational crash attenuator without braking, while the car was driving in “autopilot”, led to huge debates on the safety of autonomous vehicles. This led the US - National Transportation Safety Board (NTSB) to introducing nine safety recommendations. The investigation of the incident further identified seven safety issues in 2020 that included, among others, the drivers’ distraction and the lack of a driver engagement monitoring, leading to the recommendation of a driver monitoring system for vehicles equipped with SAE Level 2 automation or higher by the American Society of Mechanical Engineering [ASME 2020]. Similar aspects were already included in the European New Car Assessment Programme (Euro NCAP)[2] roadmap 2025 of 2017 [NCAP 2017]. In this roadmap, the Euro NCAP recommends the integration of driver/ occupant monitoring systems into the vehicles to prevent the two major “human mistakes” of violations against given regulations and human errors occurring while driving in an vehicle. These include, among others, detecting and mitigating the effects of intoxicated drivers (e.g. alcohol or drug violation) and incapacitated drivers (e.g. fatigue or distraction).

We can now further deduce that not only the human driver is able to take over the vehicle in critical situations, but that the vehicle could also monitor the human’s driving behaviour while manually driving, to identify critical driving or driver states in which manual driving leads to an endangerment of the surrounding traffic, and take over the control from the driver. In case of sleepy drivers a warning of the driver is already state-of-the-art, and a majority of automotive companies has integrated driver drowsiness detection systems in their vehicle fleets. An intervention from the vehicle’s side, however, is not yet realized. The detection of other disruptions like distraction, stress and emotion, has, until now, rarely been addressed, although they influence the driving performance provably. A reliable monitoring of the driver’s state, however, serves as prerequisite to enable an intervention from the vehicle’s side.

1.2 Level of Adaptation Towards the Driver

To increase the safety and comfort of manual and automated driving, most vehicles are nowadays already equipped with a various number of ADAS systems. These systems support the driver in critical situations, which may be caused by the vehicle

[2] Providing recommendations, rating and standardizations regarding vehicle safety since 1996

environment or the driver himself (e.g. emergency brake assist, blind spot monitor, or lane departure warning), but can at the same time also increase the comfort of driving (e.g. intelligent speed assist, adaptive cruise control, or intelligent park assist). Considering the current development progress of such systems, it is possible to differentiate vehicles into three categories (cognitive cars, intelligent vehicles and empathic vehicles), which can be seen as different development stages regarding their level of adaptation towards the driver.

- Starting off with the *cognitive car*, this vehicle shows the least amount of adaptation towards the driver. It solely monitors the interaction between driver, vehicle and traffic and reacts in relevant situations [Heide & Henning 2006; Gadsden & Habibi 2009]. Considering the recent technology progress in automotive industry, this type of vehicle can be seen as current state-of-the-art.
- The next level of adaptation towards the driver is met by introducing the *intelligent vehicle*. This type of vehicles are cognitive cars that are additionally able to monitor critical states of the driver (e.g. drowsiness or distraction). Whenever a critical driver state is detected, the vehicle warns the driver and partly/ fully takes over control from the driver [Flemisch et al. 2013]. With regard to the above mentioned Euro NCAP recommendation, a progress towards intelligent vehicles can be seen. The considered driver states, however, mainly focus on the detection of "obvious hazards" caused by sleepy or distracted drivers. The impact of emotions on the driver is mostly neglected.
- While the aforementioned vehicle types focus on the pure prevention of dangerous incidents by warning the driver and taking over the control of the driver, *empathic vehicles* are additionally able to recognize, understand and give a tailored response to the driver's internal state of interest [Oehl et al. 2020]. This is, for example, realized by mirroring or balancing the emotions of the driver [Hernandez et al. 2014; Drewitz et al. 2017; Braun et al. 2019]. In contrast to cognitive cars and intelligent vehicles, empathic vehicles can be seen as a future technology trend.

With the increase of adaptation towards the driver also the complexity of the driver monitoring system increases. While cognitive cars and intelligent vehicles do not dialogically interact with the driver, the empathic vehicle needs to communicate to the driver using additional HMI components, which may lead to further distraction of the driver from the primary driving task. This also leads to additional requirements to develop such systems, for example, by including a Decision Support System (DSS) to prioritize and assess the driver state, and provide tailored mitigation strategies [Löcken et al. 2017; Braun et al. 2019].

1.3 Impact of Emotions while Driving

With regard to the aforementioned progress in automotive technology, emotions are still less considered. While drowsiness and distraction are already seen as highly relevant for vehicle safety (cf. [NCAP 2017; ASME 2020]), the impact of emotions on the primary driving task is mostly neglected, although its impact on the safety and comfort of driving has been proven for several years [Pêcher et al. 2010]. On the one hand, emotions may directly impact the driving behaviour as they affect driving-relevant cognitive capabilities, such as the build-up of a sufficient situation representation [Jeon 2015] or decision making [Freese & Jipp 2015], in both negative and positive ways. These impairments, however, are barely compensated by the driver, as the driver is often unaware of the effects they have on the driving behaviour, unlike impairments caused, for example, by distraction or drowsiness [Jeon 2015]. On the other hand, emotions, especially negative ones, may influence the user experience and, hence, the acceptance of technical systems and automated driving functions (cf. [Picard & Klein 2002], [Klein et al. 2002], [Koo et al. 2015] and [Drewitz et al. 2017]). Therefore, especially in the field of automated driving and with regard to the future technology of empathic vehicles, emotions should receive increased attention.

Most investigations based on emotions while driving concentrate on negative emotions with a focus on frustrated and anxious drivers. Frustration can not only lead to aggressive driving behaviours, but can culminate in so-called *road rage* [Shinar 1998]. According to German insurance companies, one third of deathly road incidents are caused by aggressive driving behaviour [Grasberger 2013]. In case of anxious drivers two behavioural patterns have been identified. On the one hand, anxiety may have a positive effect on manual driving, as it can lead to an increase of situation awareness [Lu et al. 2013], leading to less risk-taking and an adaptation of the driving behaviour towards the given environmental circumstances (e.g. heavy rain causing slowing down in speed). On the other hand, it may also have a negative effect, as it can cause a decrease of the driver's attention focus [Jeon et al. 2014]. Disregarding negative emotions, it has been shown that also positive events can affect the longitudinal and lateral driving parameters, for example, by listening to "happy music" [Pêcher et al. 2009; Steinhauser et al. 2011]. In [Taubman-Ben-Ari 2012] it is further shown that positive affects can lead to a greater willingness of reckless driving.

As a consequence, this Thesis is based on the following main motivational aspects: Increasing of safety in automated driving by monitoring the driver's ability to take over the vehicle in case of automation outage, and in manual driving by warning the driver in case of an identified critical driver state or even intervening by taking over the control from the driver, increasing the level of adaptation towards the driver to enable a more natural-like interaction between vehicle and driver and, consequently, decreasing the negative impact emotions have on the driving behaviour.

1.4 Three Main Research Hypotheses

From the presented motivational aspects three main research hypotheses were identified. These hypotheses aim at covering the requirements to develop a speech based emotion recognition system in an in-vehicle environment.

1. **Hypothesis:** It is possible to induce naturalistic emotions in the driver, while driving in a real vehicle.

   The basis of a reliable speech based emotion recognition system is conditioned by the underlying/ utilized data base. A suitable database is, however, in most cases not available, especially when it comes to fundamental research in newly identified research areas. The data availability of in-vehicle emotional speech data is still very limited, especially with regard to highly natural and low-expressive emotions, as they occur in everyday driving situations. I hypothesize that it is possible to induce naturalistic emotions to the driver, while driving in a real vehicle. This Thesis, therefore, includes a data collection of real-world in-vehicle emotional data in three modalities. To the date of realization of the data collection in 2018, there did not exist a publicly available data set of this scope. The realized data collection is based on designated use cases and study designs, which were identified beforehand. The recorded data is afterwards validated to identify the reliability and usability for the present in-vehicle emotion recognition task, on the one hand by evaluating the corresponding bio-physiological signals and on the other by annotating the speech signal.

2. **Hypothesis:** It is possible to compensate effects of speech distortion.

   Considering “in the wild” speech emotion recognition environments, it can be assumed that the audio quality is strongly degraded compared to the audio quality of speech recordings obtained in a laboratory environment. With a focus on in-vehicle environments, there exist various factors that can affect the audio quality. The two main effects, causing these differences, are changes occurring in the acoustic characteristics, depending on the vehicle itself (e.g. type of vehicle, size and material of the cabin), and environmental noises (e.g. engine sounds, road surface and traffic noises). It needs to be identified how to cope with these effects and their influence on the speech signal. I hypothesize that it is possible to compensate for these effects, for example, by applying well-established speech enhancement or noise reduction algorithms. Therefore, I will first evaluate the effect of audio quality on the speech emotion recognition system and further evaluate the influence of digital signal processing steps on the features used to recognize emotions from speech and consequently the effect on the recognition performance.

3. **Hypothesis:** Under the assumption that hypotheses one and two apply, it is possible to automatically detect the emotional state of the driver by only considering the speech signal of the driver and its prosodic features.

   By utilizing machine learning algorithms it is possible to design a classification model to identify the current emotional state of a speaker. Depending on the classification task, identifying the optimal model can be quite challenging. Especially in case of low-expressive and highly natural emotions, the clusters of the individual emotional classes can strongly overlap. This makes it difficult to distinguish the emotional classes. There exist multiple ways to optimize the classification process, for example, by performing a feature reduction or parameter optimization. Therefore, I further hypothesize that, under the assumption that hypotheses one and two apply, it is possible to automatically recognize the natural emotional state of the driver by only considering the speech signal of the driver, without further evaluating the spoken content of the driver's speech. This is evaluated by designing a classifier that is able to detect the emotional state of the driver well above chance level. The performance of this classifier is further increased by applying customized feature reduction and parameter optimization methodologies.

1.5 Structure of the Thesis

The remainder of this Thesis is structured as follows.

In Chapter 2, I will first give insights to the reader on relevant state of the art research and methodological background, which serves as a prerequisite of the Thesis. The Chapter includes all relevant information on generating emotional speech, speech emotion recognition in general (e.g. how to model and evaluate a speech emotion recognition system and recent findings in speech emotion recognition), the effect of speech quality and speech disturbances on speech emotion recognition and finally an in-depth literature review on speech emotion recognition in in-vehicle surroundings. At this point I want to mention that Chapter 2 is the only chapter not based on own contributions.

Chapter 3 serves as basis of most of the investigations presented afterwards. In this Chapter, I present to the reader the realized data collections. This includes a collection of simulated and a collection of real-world emotional in-vehicle speech data. In case of the real-world data collection an additional validation of the data is presented. Furthermore, the collected data does not only include audio recordings but additional video recordings of the driver's face and recordings of the driver's biophysiological measures, which were not evaluated to their full extend in the scope of this Thesis.

Chapter 4 focuses on the evaluation of speech quality of compressed and disturbed speech samples and the ability to detect emotions from degraded speech. For the evaluation of the speech quality, state of the art quality measures as well as a newly developed quality measure that can be applied to both compressed and disturbed speech are presented.

In Chapter 5, I will examine the necessity of applying processing steps to the unprocessed emotional speech samples. On the one hand, I will describe the annotation process and present the annotation results of the conducted real-world driving study of Chapter 3. On the other hand, I will investigate the effect of speech enhancement on speech emotion recognition in terms of altered features, speech quality and the recognition performance.

The ultimate research hypothesis of this Thesis (hypothesis #3) will be evaluated in Chapter 6. I will present to the reader the design of a feature and parameter optimized speech emotion recognition system. Two classification approaches will be presented, evaluated and put in comparison. This process further includes a feature reduction and hyper-parameter optimization of the individual classification models.

Chapter 7 concludes this Thesis. In this Chapter, I will recapitulate the findings of the previous four Chapters with respect to the presented three main research hypotheses, and compare the main research results with recent state of the art investigations. I will further give insight on remaining open research questions and possible future development approaches, which have not been covered in the scope of this Thesis.

CHAPTER 2

State of the Art

Contents

IN every good book the reader should be provided with the most relevant information on the main characters, to build up a relation with the characters and empathize with the circumstances of the evolving story line. This is also the case for a good Thesis, where the reader should be introduced to related publications and relevant methodical approaches. The aim of this Chapter is to provide this information.

As already mentioned, this Thesis covers multiple research areas, therefore, it is also necessary to include an extensive state-of-the-art literature review covering each of the listed research areas. To limit the extent of this Chapter, further textbook knowledge is either referred to as citation or presented in additional appendices.

2.1 Generating Emotional Speech Data

I will start off with an introduction on how to generate emotional speech data. This will serve as basis for the realized data collection presented in Chapter 3 and the performed annotation in Section 5.1. First of all, I will emphasize the importance of defining the scope of the data collections correctly, as the obtained data should meet the requirements of the later application domain. As I focus on the collection of emotional speech data, I will introduce the reader to the concept of emotion, how emotions are expressed by humans and how they can be represented. Afterwards, I will present an overview of methodical approaches on how emotions can be elicited in a test subject. Especially for real-life data collections the probability of inducing a said emotion identically to different test subjects is rather low, due to changing environmental conditions and differences in the subjects themselves. It is assumed that the experimenter is a-priori unaware of the emotions felt by the subjects (cf. [Larradet et al. 2020]). Therefore, a posteriori annotation of the collected data needs to be conducted to generate a ground truth. The Section will be concluded with an overview on relevant benchmark data sets used in the scope of the Thesis.

2.1.1 Scope of the Data Collection

The speech signal is a highly variable signal, which is strongly dependent on the speaker characteristics (e.g. age, gender, health, cultural differences), speech style, speaking rate, dialect differences, non-native accents (cf. [Babel & Munson 2014; Docherty & Mendoza-Denton 2011]), present background noises (also see Section 2.5.1), side talk, as well as the recording environment (e.g. microphone setup in the wild, anechoic chamber, tv-studio) [Yu & Deng 2015]. A data collection which covers all these variations could be seen as an *universal* data set, used to describe the population as a whole. This, however, is highly challenging when it comes to enriched data, as it is the case for emotional speech [Böck et al. 2019]. While there exists a huge amount of transcribed speech data in various languages for speech recognition, the amount of annotated/ labeled emotional speech data is still limited. With regard to the aforementioned variations of the speech signal, it is rather unlikely, that there exists a data set, which contains enough data to represent the population as a whole. This issue will also be addressed later, when it comes to the recent findings of speech emotion recognition in the wild (see Section 2.3), as, especially in data driven machine learning, the utilized data set needs to represent the population in a sufficient way. When it comes to the collection of suitable, reliable, trustworthy and reproducible speech data, it is, therefore, of high relevance to design the data collection with regard to the above mentioned factors and specify each factor in a distinct way, such that it meets the requirements of the desired application domain without the pretension to cover the whole population. Hence, the desired application domain determines the scope of the data collection.

Additional factors, which need to be considered when it comes to emotionally enriched speech data, are the naturalness of the expressed emotion and the coverage of the search space. Typically, it is distinguished between acted, scripted and naturalistic emotions (in increasing level of naturalness) (cf. [Siegert 2015]). While data sets of acted emotions mainly contain sentences of emotional content uttered by actors in a highly expressive way (e.g. Berlin Emotional Speech Database (EmoDB) [Burkhardt et al. 2005], Danish Emotional Speech Database (DES) [Engberg et al. 1997] or Polish-EMO [Staroniewicz & Majewski 2009]), the expressiveness further decreases with an increase of naturalness. In case of scripted emotions, the speaker is still aware of the purpose of the data collection, but, in contrast to acted emotions, is prompted to express said emotion in a more natural way based on scripted plays or spontaneous hypothetical scenarios (e.g. eNTERFACE'05 [Martin et al. 2006] and IEMOCAP [Busso et al. 2008]). The highest level of naturalness is obtained in case of naturalistic emotions. Here, the emotions are naturally induced to the speaker while she/ he is unaware of the actual purpose of the data collection (e.g. FAU Aibo Emotion Speech corpus [Batliner et al. 2004] or RECOLA [Ringeval et al. 2013]). The advantages of natural emotions will be addressed in the further course of this Section, with regard of how emotions affect the human bodily function. As second factor, the collected emotional data needs to cover the search space sufficiently. Ideally, this would imply that all considered emotions are equally represented in the data set, by an equal number of male and female speakers for each age group (just to name some features considered in the search space). For acted and scripted emotions, the equal distribution of all considered emotions is controlled easily by giving correct instructions to the speaker. In case of natural emotions, this is more challenging and controlled by the used experimental design and employed inducement method. The true coverage of the search space and true emotional state of the speaker can, in this case, only be assessed by annotating the recorded speech samples (see Section 2.1.4).

2.1.2 Concept of Emotion

Until now I have shortly presented to the reader which factors impact the scope of emotionally enriched data collections. One aspect which has not been addressed so far, is the definition of emotions, how they are expressed by humans and how it is possible to represent the emotional state.

The nature of emotions has already been addressed by ancient Greek philosophers like Socrates or his student Aristotle. Even though they did not literally used the word emotion, their work on ‘passion’ or ‘mood’ show a strong consistency in their definition compared to today’s appraisal theory (cf. [Solomon 2000]). In [Aristotle & McKeon 1941][1] Aristotle, for example, defined emotions as a state affecting one’s judgment and being accompanied by pleasure and pain. He further names concrete emotional states such as anger, fear, and pity, and even gives clear definitions on single states, such as anger. This definition does not only include a distinct cognitive component, but also a specific social context, a behavioural tendency, and a recognition of physical arousal. Comparing the following definition with most recent definitions in appraisal theory, there exist astonishing similarities.

The recent definitions of emotion in appraisal theory have started to evolve since the 1960th (cf. [Stearns 2000]) leading to the introduction of the *component process model* of emotions by Scherer [Scherer 1987; Scherer 2009]. Based on this assumption Scherer defines emotions as...

> “... an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism.” [Scherer 1987]

With regard to this definition, an emotion is not a state but a process consisting of changes in the five organismic subsystem. Scherer further defines five components of an emotion episode which are used to describe this process. These components are:

1. Cognitive component: Evaluation of objects and events as information processing unit, leading to continuous changes in the appraisal processes of the central nervous system.
2. Neurophysiological component: System regulation as support unit, leading to changes in the response pattern in the neuroendocrine, autonomic, and somatic nervous system.

[1] As translated in [Solomon 2000]

3. Motivational component: Preparation and direction of actions as executive unit, leading to changes in behavioral tendencies.
4. Motor expression component: Communication of reaction and behavioral intention as action unit, leading to changes in the facial expression, body movement and vocal expression.
5. Subjective feeling: Monitoring of internal state and organism-environment interaction as monitoring unit, leading to subjective changes in the own condition of interaction with the environment.

There are several ways to measure the changes occurring in the different components of an emotion episode (cf. [Scherer 2005b]). The cognitive component (1.) and subjective feeling (5.) can be assessed by employing self-reports or measuring neural activity through electroencephalogram (EEG) or functional Infrared Spectroscopy (fNIRS). Changes in the physiological component (2.) lead to changes in the arousal and can be assessed through physiological (e.g. heart rate, skin conductance or blood pressure) and hormonal parameters. Behavioral changes (3.) again are hard to assess and are highly domain dependent. While it is possible to measure especially expressive behavior, behavioral changes of lower expressiveness are still challenging to assess [Harrigan et al. 2005]. Changes in the facial expression, body movement and vocal expression (4.), however, can be tracked by analyzing video and audio signals of the subject [Schirmer & Adolphs 2017]. With regard to these measurement options, this Thesis focuses on the detection of emotions by the changes occurring in the *motor expression component*, represented by changes in the speech of the speaker. In [Ekman 1999], the author gives further evidence on how certain emotions correlate with the five components.

We now know how emotions are naturally produced and expressed by the human organism, however, it now needs to be further defined how emotions can be represented in a distinct way. In general it is distinguished between two types of emotion representation methods, categorial and dimensional. A good overview on the representation and elicitation of emotions is presented in [Becker-Asano 2008] and [Siegert 2015].

Categorial Emotions

The categorial representation of emotions relies on the assumption that emotions can be described using so-called psychological primitive building blocks (cf. [Ortony & Turner 1990]). These emotions are also referred to as primary or basic emotions (cf. [McDougall 1908]). These basic emotions differ one from another in important ways, which lead to an unambiguous description of the subject's emotional state. There have been different assumptions on how many basic emotions exist. The most commonly used definition of basic emotions is introduced in [Ekman 1972], where the authors defined six basic emotions based on universal facial expressions which occur cross-culturally when humans express certain emotions, namely, *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*. Ekman later extended this list of basic emotions to also include *contempt*, considering 12 characteristics which are found in nearly all basic emotions and changed his view on the definition of basic emotions (cf. [Ekman 1999; Ekman & Cordaro 2011]).

Another approach suggesting the existence of basic bipolar emotions is presented in [Plutchik 1958]. Plutchik agrees on Ekmans general concept of basic emotions, but introduces a new concept, the *wheel of emotions* (cf. [Plutchik 2001]). His concept is based on the assumption that there exist primary emotions (i.e. basic emotions), which can be conceptualized analogously to a color wheel, as already supposed in [McDougall 1908]. The *wheel of emotions* includes eight primary emotions (*joy, trust, fear, surprise, sadness, disgust, anger* and *anticipation*), which are ordered in a circumplex fashion with similar emotions close together and their complementary emotions lying on the opposite side (cf. [Plutchik 1980]), just like complementary colors in a color wheel. The eight primary emotions can be mixed together to form other emotions (e.g. *joy* and *trust* mix together to form *love*). Plutchik further adds a third dimension to the wheel, which indicates the intensity of emotion (e.g. *annoyance* < *anger* < *rage*). The total concept of his emotional representation is shaped like a cone (depicted in [Plutchik 2001], Figure 6 on Page 349), but can also be represented in a flattened 2-dimensional way (see Figure 2.1).

Dimensional Emotions

Contrary to the assumption of Ekman on basic emotions, Mehrabian and Russell argue that emotions are not discrete and separate, but need to be described using the three dimensions of *valence* (negative vs. positive), *arousal* (high vs. low) and *dominance* (high vs. low) [Mehrabian 1996; Mehrabian & Russell 1974]. Synonyms also used in the field of emotion science are *pleasure* instead of valence, *activation* instead of arousal and *control* instead of dominance. Russell further presents the circumplex model of affect, where the dimension of valence and arousal span a two-dimensional space of *core affect* [Russell 1980]. He assumes that discrete emotion categories and the core affect are related to each other systematically and presents a schematic map of the core affect, as well as a mapping of the six basic emotions onto the core affect (cf. [Russell & Barrett 1999] and Figure 3.5 on page 93).

In the course of this Thesis, the reader will be confronted with both representation methods of categorial and dimensional emotions.

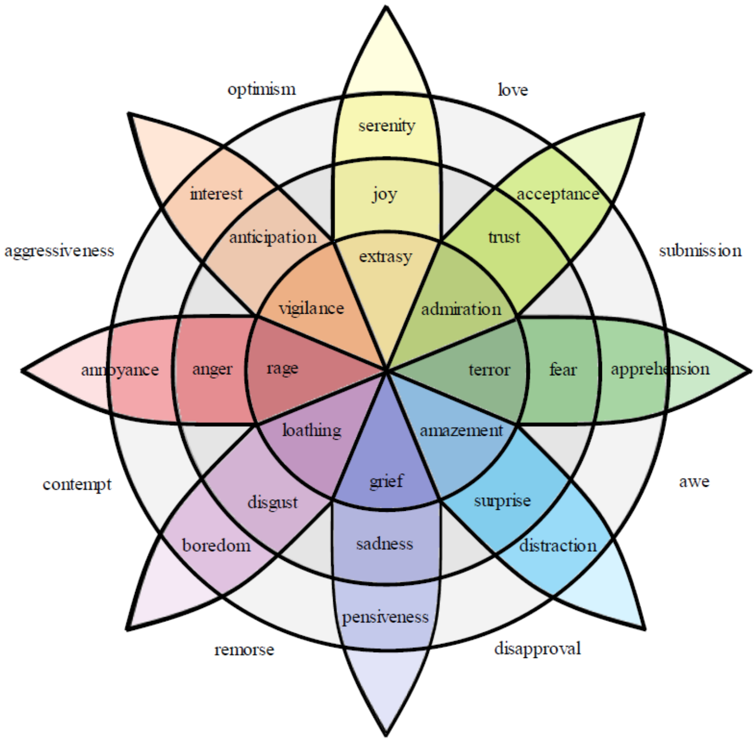


Figure 2.1: Wheel of emotions in 2-dimensional representation (adapted from [Plutchik 2001]).

2.1.3 Inducing Emotions

With the knowledge of how emotions are expressed by humans and how they can be represented in a distinct way, we will now focus on how to reliably induce the desired emotional state to a subject.

When collecting emotional data, it is of high importance to ensure the correct inducement of the desired emotional state to the subject. This is especially relevant when used in natural emotion recognition and applied for data-driven classification tasks, as used in the scope of this Thesis. Depending on the research question, individual inducement methods can be applied and will lead to an individual satisfactory level regarding the correctness of inducement. In the following, multiple inducement methods, their impact on the naturalness of the uttered emotion and their suitability for different recognition tasks will be presented.

There exist *simple* ways to collect emotional data for speech, facial expressions or body postures by using no emotion elicitation method at all, but pretending to experience a certain emotion (e.g. acted and scripted emotions). This approach is rather challenging when it comes to the elicitation of emotions in the subject's physiological reaction, as the subject may not be able to simulate this reaction.

Furthermore, early research has already shown that acted emotional data can differ significantly from emotions experienced in real-life (e.g. [Hoque et al. 2012]) and, hence, will not lead to reliable results when used to train a data-driven classifier for a later real-world application (cf. Section 2.3 and, for example, [Devillers et al. 2005; Healey et al. 2010; Wilhelm & Grossman 2010] and [Xu et al. 2017]). In the literature several methodological approaches can be found, used to induce/ elicit emotions in a more natural fashion, leading to certain response patterns in the subject's physiological data. A more recent publication on how emotions can be induced is [Larradet et al. 2020]. Here, the authors distinguish between seven different inducement methods ranging from self-inducement and retrospection (i.e. thinking about and narrating specific situations in which the subject experience a certain emotion) (cf. [Vrana 1993; Pasupathi 2003]), over validated experimental protocols (i.e. exposing subjects to pre-defined and pre-validated emotion stimuli like pictures (e.g. IAPS [Lang et al. 2008], GAPED [Dan-Glauser & Scherer 2011],...), movie extracts (e.g. LIRIs-ACCEDE [Baveye et al. 2015], FilmStim [Schaefer et al. 2010], E-MOVIE [Maffei & Angrilli 2019],...) or music (e.g. film music data set [Eerola & Vuoskoski 2011], DEAM [Aljanaki et al. 2017],...) and active participation of the subject using video games or virtual reality) (cf. [Dikecligil & Mujica-Parodi 2010; Fox et al. 2010; Schmidt et al. 2011; Walter et al. 2011; Rooney et al. 2012; Konečni 2008; Kreutz et al. 2008; Vuoskoski & Eerola 2011; Tognetti et al. 2010; Ververidis et al. 2008; Bassano et al. 2019] and [Kim & André 2008]), self-reporting using annotation interfaces (i.e. self-reported affect ratings of subjects while experiencing certain media) (cf. [Melhart et al. 2019; Girard 2014]), methods based on facial feedback theory [Tomkins 1962; Izard 1977] (i.e. guided activation of specific facial muscles or postures [Zajonc et al. 1989]), to simulated realistic social interactions (cf. [Harmon-Jones & Sigelman 2001; Niewiadomski et al. 2016; Harmon-Jones et al. 2007]) and supervised real-life studies (i.e. putting the subjects into a situation in which they experience strong emotions) (cf. [Dikecligil & Mujica-Parodi 2010] and [Healey & Picard 2005]). Especially when applying the later listed methods of simulated realistic social interactions and supervised real-life studies it is of high importance to keep the subject unaware of the actual data purpose to receive spontaneous unbiased emotions.

In this Thesis two of the presented inducement methods are applied in Chapter 3, namely, emotion elicitation through retrospective and through supervised real-life studies. In Section 3.2.3, it was further confirmed that these inducement methods led to a successful inducement of the relevant emotional states.

2.1.4 Generating the Ground Truth

With the correct inducement method and study design, it could be assume that the collected data material meets all the requirement needed for the later emotion recognition task. Nevertheless, with a strong inter-individuality of the human subjects, a ground truth of the data is needed to assure the usability of the data for the desired application domain (i.e. recognition of natural emotions from speech).

As emotions evolve over time and are not continuously expressed by the speaker (cf. [Pell & Kotz 2011]), the speech data needs to be divided into smaller sub-samples, for which the emotional state can be assumed to be constant. In most cases this is done by labeling the speech signal utterance wise or in speech-segments of equal length. Afterwards, there are multiple ways to determine the ground truth and label of the obtained sub-samples. They can be roughly distinguished into subjective and objective measures. While subjective measures are based on the subjective feedback given by the test subject herself/ himself, objective measures are based on externally measurable signals, such as but not limited to changes in bio-physiological signals or annotations provided by independent labelers. To assess the emotional state through changes occurring in bio-physiological signals of the subject an accurate response pattern needs to be available. In case of emotional arousal, a clear relation to an increase or decrease in heart rate, cutaneous blood flow, piloerection, sweating and gastrointestinal motility is observed [Purves et al. 2001]. In [Ekman 1999], the authors state that there exist distinctive patters of the autonomic nervous system for the states anger, fear, disgust and sadness. The automatic generation of the ground truth by mapping a distinct emotion category or dimensions of the emotional space to these response patterns is, however, still challenging. Therefore, the speech data needs to be labeled using either the subjective self-report of the speaker or an objective annotation of independent labelers.

One could assume that a labeling based on a subjective self-report would show the true emotional state of the speaker. This, however, is disproved in [Truong et al. 2012]. Here, the authors show that the agreement between multiple self-ratings of the speaker is lower compared to the inter-rater agreement of multiple independent labelers. They further identify that this does not only affect the outcome of the labeling but also the performance of the speech emotion recognition system trained on this kind of labeled data. From emotion theory it is known that the expression of a certain emotion by a subject is affected by so-called display rules, which are strongly dependent on the individual's cultural, gender and family background [Ekman & Friesen 1975]. This also affects the way a certain emotion is perceived by the subject's counterpart. Another disadvantage of utilizing subjective self-reports is the disruption of the subject while being involved in the experimental scenario at regular intervals. Not only might the speaker not be able to verbalize the actual emotional state felt while conducting the experiment, she/ he might also be distracted from the actual experimental scenario utilized to trigger a certain emotion [Scherer 2005a]. One way to subsequently determine the ground truth would be to conduct post-hoc interviews. This however, is a costly process, as a trained psychologist is needed to perform and analyze these interviews. A more common way to generate the ground truth is therefore to assess the experimental data by a large number of independent labelers and determine a valid emotional label by performing a majority voting. To ensure the reliability of the labeling, the labelers need to be well-trained and familiar with the assessment of emotional speech. Furthermore, it is also advisable to assess the quality of the obtained labels by calculating suitable measures such as the Inter-Rater-Reliability (IRR) (see page 22).

Labeling Methods

In either case, subjective self-report or independent labeling, a suitable labeling method needs to be provided to the labeler. A good overview on relevant emotional labeling methods is given in [Siegert 2015]. I will now only introduce those methods used in the scope of this Thesis, namely *free text input*, *word lists*, *Self Assessment Manikins (SAM)* [Bradley & Lang 1994] and the *Geneva Emotion Wheel (GEW)* [Scherer 2005b; Scherer et al. 2013].

Free Text Input: The labeler can describe the experienced/ perceived emotional state in their own words, without further instructions. She/ He is free to state emotional bullet points or write whole sentences. This on one hand gives the labeler an unrestricted possibility to describe the perceived emotion, but on the other hand is in need of a costly post-evaluation of the provided labels, as the provided text input needs to be mapped into clusters of similar meanings.

Word Lists: A list of emotional words is provided to the labeler. She/ He can choose from these words to label the experienced/ perceived emotional state. To gain reliable results the labeler needs to undergo a training process in which she/ he is introduced to the given labels and their meaning [Morris 1995]. Furthermore, the labeler is strongly restricted in the decision process, by the pre-defined labels. This may lead to an information loss of the labeled data, if several emotions are merged into one emotional term. This, however, may also be intended in some cases, where a fine-grained differentiation is not wanted/ needed. To prevent this information loss, in some cases an additional free text input is provided to the labeler. Commonly used words used in emotional word lists are: basic emotions as provided by Ekman (cf. Section 2.1.2), positive vs. negative or specific task related words (cf. [Lefter et al. 2012; Devillers & Vasilescu 2004] or [Lee & Narayanan 2005]).

Self Assessment Manikins: The SAM-scale (cf. [Bradley & Lang 1994]) is used to annotate the dimensions of valence, arousal and dominance based on a picture-oriented labeling process containing five images for each emotional dimension [Bynion & Feldner 2017]. The original SAM-scale, as described in [Bradley & Lang 1994], consists of a five-point scale (see Figure 2.2). It was first introduced by Lang to self-assess the emotional response to an object or emotion [Bynion & Feldner 2017] (cf. [Lang 1980]). Nowadays, it is also used in case of independent labeling. There also exist larger versions containing a 9-

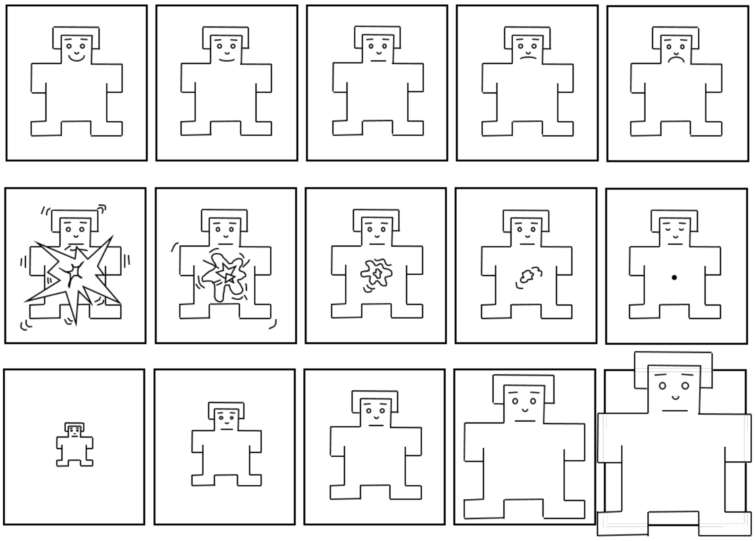


Figure 2.2: Self Assessment Manikins (SAM) (adapted from [Bradley & Lang 1994]).

or 21-point scale with intermediate decision points lying in-between two consecutive images. The upper image row, depicted in Figure 2.2, corresponds to the dimension of valence from left, positive valence, to right, negative valence, the middle row corresponds to the dimension of arousal from left, high arousal, to right, low arousal, and the bottom row corresponds to the dimension of dominance from left, low dominance, to right, high dominance. Depending on the utilized scale the labeler chooses one of the images (or intermediate points) for each dimension. This imagery-based approach is not restricted to a certain language and can be used cross-lingually and cross-culturally [Bradley et al. 1992]. This also enables the usage by children [Lang 1985].

Geneva Emotion Wheel: The GEW (cf. [Scherer 2005b; Scherer et al. 2013]) consists of 20 emotion families arranged in a circle (see Figure 2.3). It is designed to combine the approaches of discrete and dimensional emotion assessment, and aligns the emotion families to the dimensions of valence and control in a circular arrangement, separating the wheel into for quadrants (negative valence - low control, negative valence - high control, positive valence - low control and positive valence - high control). The labeler can choose between the 20 emotion families and additionally rate the intensity of the experienced/ perceived emotion. Furthermore, she/ he has the option to label *none* (no emotion) or *other* (not included emotion). The authors further provide a separate instruction document on their homepage[2], stating three alternatives on how to utilize the GEW. The first alternative instructs the labeler to rate the intensity of the one emotion which best describes the experienced/ perceived emotional state. The second alternative instructs the labeler to rate the in-

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

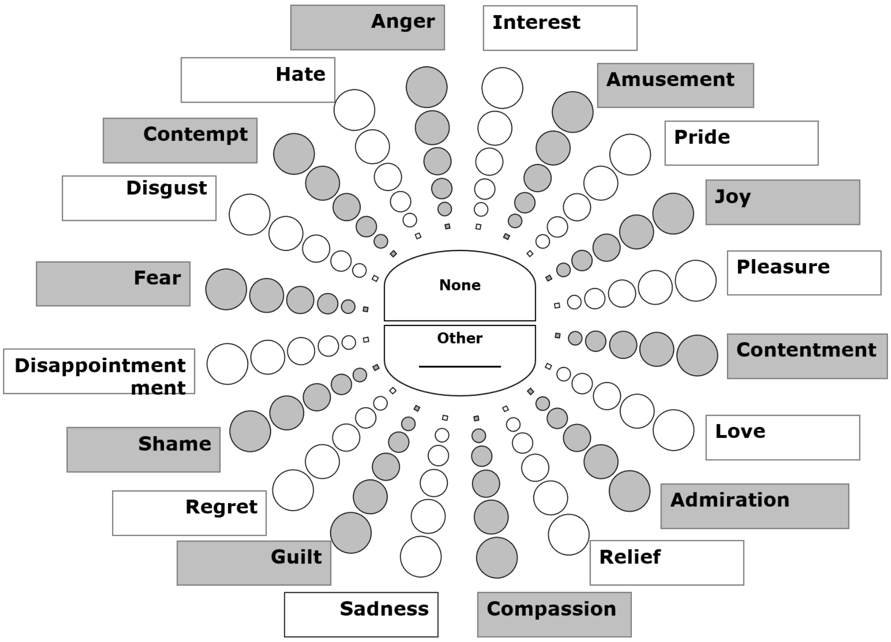


Figure 2.3: Geneva Emotion Wheel (GEW) (cf. [Scherer 2005b; Scherer et al. 2013]).

tensity of those emotions in the wheel, which contribute to the experienced/ perceived emotional state. The last alternative instructs the labeler to rate the intensity of all the emotions in the wheel and rate those emotions, which were not experienced/ perceived at all with the lowest intensity.

These four methods are all applied in the course of this Thesis. Other methods, which were not applied are: the FEELTRACE [Cowie et al. 2000], the Product Emotion Measurement Tool (PrEMO) [Desmet et al. 2007], the AffectButton [Broekens & Brinkman 2009; Broekens & Brinkman 2013], the PANAS [Watson et al. 1988], the 26-item scale Berlin Everyday Language Mood Inventory (BELMI) [Schimmack 1997], the 5-point Differential Emotions Scale (Version 4) (DES-IV) [Izard et al. 1993], and the 18-point bipolar Semantic Difference Scale (SDS) [Mehrabian & Russell 1974]. While the first three methods (FEELTRACE, PrEMO and AffectButton) can be used in case of self-reporting and independent labeling, the latter are developed for self-reporting only.

Inter-Rater-Reliability (IRR)

As stated earlier in this Section, the way humans express and perceive certain emotions is strongly affected by so-called display rules (cf. [Ekman & Friesen 1975]). When employing multiple labelers to annotate the same emotional data, it is assumable that they will not rate the perceived emotion identically but with a variation depending on their individual cultural, gender and family background. Therefore, the reliability of the obtained annotation results needs to be determined by calcu-

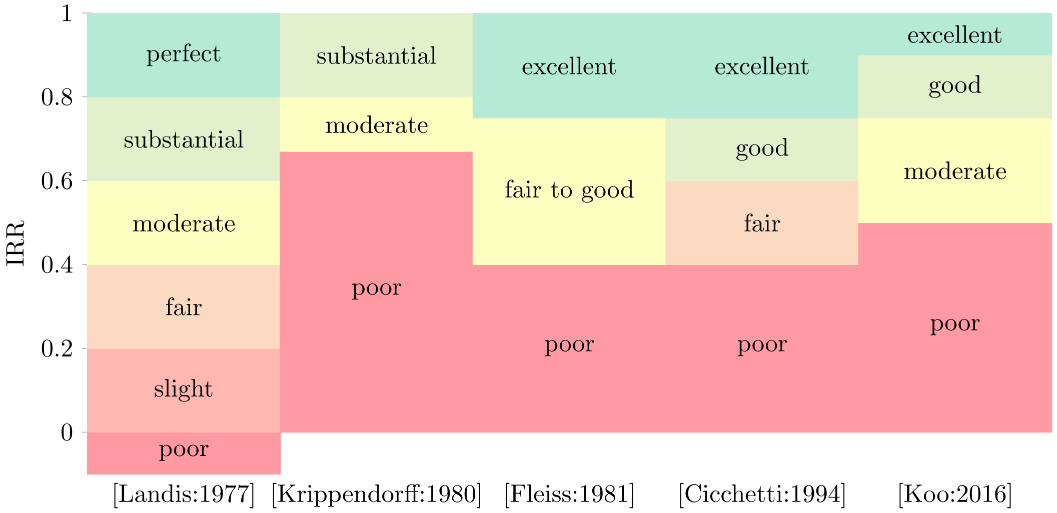


Figure 2.4: Recommendations of interpretation of the IRR, as presented in [Landis & Koch 1977; Krippendorff 2004; Fleiss et al. 2003; Cicchetti 1994] and [Koo & Li 2016]

lating the IRR of the employed labelers/ raters ($r = 1, ..., R$, with R being the total number of employed raters).

The most commonly used IRR-measures are Cohen's κ [Cohen 1960] and κ related measures (e.g. Bennett, Alpert and Goldstein's S [Bennett et al. 1954], Scott's π [Scott 1955], Fleiss' Multi-π [Fleiss 1971], Multi-κ [Light 1971; Davies & Fleiss 1982] or Cohen's weighted κ [Cohen 1968]), Intra-Class-Correlation (ICC) [McGraw & Wong 1996; Shrout & Fleiss 1979], and Krippendorff's α [Krippendorff 2004] (cf.[Hallgren 2012]). Their values all range from -1 to 1, with 1 indicating an identical labeling of the raters, 0 a random labeling and -1 a complete reverse labeling (similar to the correlation coefficient used to measure the statistical relationship between two variables). The interpretation of the values is, however, strongly dependent on the annotation task and there exist several recommendations on how to assess the IRR (cf. [Landis & Koch 1977; Krippendorff 2004; Fleiss et al. 2003; Cicchetti 1994] and [Koo & Li 2016]). Figure 2.4 gives an overview on these recommendations.

Furthermore, all measures have advantages and disadvantages and are therefore more or less suitable depending on the present annotation task (cf. [Hallgren 2012; Artstein & Poesio 2008]). The measure holding the most limitations is Cohen's κ. It is based on the probability of agreement with consideration of the expected agreement obtained by chance. One serious disadvantage of κ is that all disagreements of raters are treated equally (e.g. nominal data) and that it is primarily developed to determine the agreement of two raters only. There do exist adaptations of Cohen's κ (κ-like measures), that can be used in case of non-nominally scaled data or multiple raters. These measures, however, do only consider either the application on non-nominally scaled data or multiple-raters and not both at the same time. A second measure, which has lesser limitations than Cohen's κ is the ICC. The estimator of the ICC is based on the random effect model, which states that the observed value by the rater is a combination of the true label, the deviation of the true label from the mean rating of the to-be-labeled item and the measurement error. It is applicable with two or more raters on ordinal, interval and ratio data, but cannot be applied to incomplete data containing missing values. The most extensive measure is Krippendorff's α, as it can be used with multiple raters, it gives the possibility to define a distinct distance measures for different data types (e.g. nominal, ordinal, interval, ratio, ...) and it is applicable in case of incompletely labeled data. Furthermore, Krippendorff introduces his measure for application in content analysis, like it is the case for speech emotion annotation. Therefore, I opt for Krippendorff's α to determine the IRR of the labeling performed in the scope of this Thesis (cf. Section 5.1) and will hereinafter give a more detailed insight on how Krippendorff's α (α_{Kr}) is calculated.

Krippendorff's α_{Kr}

Unlike Cohens κ, κ-like measures or the ICC, Krippendorff's α (α_{Kr}) origins as a measure of variance and is based on similar assumptions as made for a single factor Analysis of Variance (ANOVA), with each item representing a different factor group (cf. Appendix C on page 272). This leads to the following Equation used to determine α_{Kr}:

$$\alpha_{Kr} = 1 - \frac{s^2_{within}}{s^2_{total}} \tag{2.1}$$

$$= 1 - \frac{D_o}{D_e} = \frac{\text{Average}\,\boldsymbol{\delta}^2_{k_j,k_l}\text{within all units}}{\text{Average}\,\boldsymbol{\delta}^2_{k_j,k_l}\text{within all data}} \tag{2.2}$$

with s^2_{within} being the within item variance, s^2_{total} the total variance of all data, D_o being the observed disagreement and D_e the expected disagreement. In case of a perfect agreement ($D_o = 0$), α equates to 1, indicating an excellent reliability of the labeled data. In case of random agreement ($D_o = D_e$), α equates to 0, indicating the absence of reliability. In case of systematic disagreement or sampling errors α can also equate to negative values ($D_o > D_e$). The term $\boldsymbol{\delta}^2_{c_j,c_l}$ represents the squared distance metrics between any two classes c_j and c_l, with $j, l = 1, ..., C$ and C being the number of classes included in the labeling process. As distance metrics any square difference function can be used [Artstein & Poesio 2008], such that D_o and D_e equate to

$$D_o = \frac{1}{IR(R-1)} \sum_{i \in I} \sum_{j=1}^{C} \sum_{l=1}^{C} \mathbf{n}_{i,c_j} \mathbf{n}_{i,c_l} \boldsymbol{\delta}^2_{c_j,c_l} \text{ and} \tag{2.3}$$

$$D_e = \frac{1}{IR(IR-1)} \sum_{j=1}^{C} \sum_{l=1}^{C} \mathbf{n}_{c_j} \mathbf{n}_{c_l} \boldsymbol{\delta}^2_{c_j,c_l}. \tag{2.4}$$

The term $\mathbf{n}_{i,c}$ corresponds to the number of raters who assigned class c to item i, with $i = 1, ..., I$ and I being the number of items included in the labeling process, and the term $\mathbf{n}_c$ to the number of times class c was assigned by any rater to any item. Most commonly used distance metrics for $\boldsymbol{\delta}_{c_j,c_l}$ are provided in [Krippendorff 2004], namely:

Nominal metric: In case of nominal data, the class labels are not related to any quantitative values. The labels of a considered item either match or mismatch and there exists no better or worse agreement of mismatching labels. This is, for examples, the case when utilizing emotional word lists as labeling method. The distance metric is determined by the following function:

$$_{nominal}\boldsymbol{\delta}^2_{c_j,c_l} = \begin{cases} 0, \text{if } c_j = c_l \\ 1, \text{if } c_j \neq c_l \end{cases} \tag{2.5}$$

In general, the distance between matching values equates to 0, and the distance between mismatching values equated to 1. Hence, all mismatching labels contribute equally to the determination of α in Equation 2.1.

Ordinal metric: For ordinal data the value of the assigned class can be ranked in a distinct order. The class c does not correspond to a numerical value but describes the order of the values with the algebraic difference being unknown (e.g. option 4 is better than option 3, but it is unknown how much better). This is, for example, the case when utilizing the SAM scale as labeling method. The metrics function is based on the number of ranks lying in between the labels of two raters and is determined as follows:

$$_{ordinal}\boldsymbol{\delta}^2_{c_j,c_l} = \left(\frac{n_{c_j}}{2} + \sum_{c_g > c_j}^{c_g < c_l} n_{c_g} + \frac{n_{c_l}}{2} \right)^2. \tag{2.6}$$

The distance of labels that lie only a few ranks apart is generally lower than the distance of labels with a high difference in their rank. It is evident that this distance metric is symmetric, as the rank difference between two ranks is independent of the order of the ranks and always positive (i.e. $\boldsymbol{\delta}_{c_j,c_l} = \boldsymbol{\delta}_{c_l,c_j}$).

Furthermore, in case of matching ranks, the distance should always equate to 0 (i.e. $\boldsymbol{\delta}_{c_j,c_j} = \boldsymbol{\delta}_{c_l,c_l} = 0$).

Another labeling method where the labels are ranked in a two-dimensional circular way is used in the GEW. This two-dimensional representation requires an adaptation of the metric presented in Equation 2.6. In [Siegert et al. 2014] the authors present a novel approach to determine a distance for the circular representation of emotions by calculating the Euclidean distance between the chosen state on the GEW and considering the GEW as coordinate system of two dimensions. The distance metric is then determined as

$$_{ordinal,2D}\boldsymbol{\delta}_{c_j,c_l} = \sqrt{\left(\cos\varphi_{c_j} - \cos\varphi_{c_l}\right)^2 + \left(\sin\varphi_{c_j} - \sin\varphi_{c_l}\right)^2}, \tag{2.7}$$

with the distance from one to another emotion family given as the angle $\phi = 360°/C$.

Interval metric: Contrarily to ordinal data, interval data is labeled using numerical values. In this case, the numerical values also give insight on the algebraic difference of the labels obtained from two or more raters. An example of interval data is the temperature scale in Celsius or Fahrenheit. To determined the distance metric, the simple algebraic difference is calculated:

$$_{interval}\boldsymbol{\delta}^2_{c_j,c_l} = \left(c_j - c_l\right)^2 . \tag{2.8}$$

In case of ranked data of equal frequency, $_{ordinal}\boldsymbol{\delta}^2_{c_j,c_l}$ and $_{interval}\boldsymbol{\delta}^2_{c_j,c_l}$ lead to identical α_{Kr}-values in Equation 2.1.

Ratio metric: The most sophisticated metric is used in case of ratio data. Here, not only the algebraic difference matters but also the relation to their reference point (i.e. how far they lie away from zero). This implies that differences of small values, lying close to zero, are weighted higher than the same differences between large values, lying further away from zero. This is, for examples, the case for age, weight and income. While the difference of one year of age when rating an older persons age corresponds to a remarkable accuracy, the difference of one year when rating the age of a baby correspond to a rather inaccurate result. Considering this relation to zero, the distance metric of ratio data is determined as follows:

$$_{ratio}\boldsymbol{\delta}^2_{c_j,c_l} = \left(\frac{c_j - c_l}{c_j + c_l}\right)^2 . \tag{2.9}$$

With regard to these four distance metrics, Krippendorff introduces a highly conservative interpretation scheme of the corresponding α-values [Krippendorff 2004], which is presented in Figure 2.4. A recommendation of this interpretation, however, is only valid for the presented work of content analysis by Krippendorff and would lead to unsatisfactory results when applied without further adaptation [Hallgren 2012]. Therefore, it is recommended to interpret Krippendorff's α depending on the considered study method and research question. One publication focusing on the IRR of annotated emotional speech data is [Siegert et al. 2014]. The authors apply Krippendorff's α on the annotations of multiple benchmark emotional data sets. The annotations are either provided with the data set or are assessed using multiple labeling methods (Word lists, SAM, FEELTRACE and GEW). Depending on the utilized labeling method they either apply a nominal or ordinal distance metric. In case of the well-known Vera am Mittag (VAM) data set (cf. Section 2.1.5), the authors use the annotations provided with the data set, which were obtained using the SAM scale of valence, arousal and dominance. By applying an ordinal distance metric they achieve α_{Kr} of at most 0.199 (valence), 0.485 (arousal) and 0.443 (dominance), respectively. Considering the original recommendation of Krippendorff, this indicates a poor reliability of the annotations. In case of the SEMAINE Solid-SAL data set [McKeown et al. 2010], containing more natural evoked emotions, the reliability of the provided annotations of the five core dimension (intensity, valence, arousal, power and experience) using the FEELTRACE is even lower. The best IRR is obtained for the dimension of *intensity* with an α_{Kr} of 0.14. For all other dimensions the IRR ranges between 0.09 and 0.12. This indicates a poor reliability of the annotations. From these results, the authors of [Siegert et al. 2014] conclude that emotion or affect annotation will lead to rather low IRR values compared to annotations of more objective measures like gesture, head position or linguistic turns. This supports the previously made assumption that the interpretation of the IRR is strongly dependent on the present annotation task.

Costs and Cost-Reduction

Especially for natural speech data, a complete manual annotation is difficult to obtain, as it is extremely time consuming and in most cases further limited by a fixed budget for employing (expert) labelers. Therefore, it is of high interest to decrease the amount of manually labeled data and consequently the annotation cost. This research topic will also be addressed in Chapter 5 of this Thesis. In general, there are multiple ways to decrease the annotation effort. An overview on relevant approaches and comparable results will be presented now.

One method is to reduce the amount of, by human experts, manually labeled samples, as it is the case in Passive Learning (PL) and Active Learning (AL) (e.g. [Zhang & Schuller 2012], [Han et al. 2013] and [Zhang et al. 2015]). For both approaches it is assumed that there exists some initial labeled data on which a first model of a classifier can be trained. From the remaining unlabeled data only a sub set is re-evaluated by human labelers. The main problem is how to decide which samples of the unlabeled data set should be re-evaluated to obtain a good performance of the classifier. This is done iteratively by selecting a sub set out of the unlabeled data pool, re-evaluating this sub set and re-training the model using all labeled instances. The number of iterations is set by the experimenter. In PL a sub set of samples is picked out at random. For AL the algorithm actively chooses the data from which it learns by selecting those samples from the unlabeled data pool that are intended to be the most informative. The degree of informativeness can be obtained by different query strategies. In [Settles 2010] an overview on the most common AL approaches and query strategies is given. One well established query strategy is called *uncertainty sampling* [Lewis & Gale 1994]. Here the informativeness of the samples is obtained by computing the posteriori probability or confidence values of the class assignment obtained by utilizing the model trained on the existing labeled data. The sample for which the classifier is least certain on how to assign a label is then chosen to be re-evaluated by a human expert. Other variants of uncertainty sampling are *margin sampling*, where the sample achieving the lowest margin between the two most popular class labels is chosen for re-evaluation [Scheffer et al. 2001], or methods based on the information *entropy* of the sample (e.g. [Z. Zhang et al. 2018]). A way to also deal with noisy data is to use a medium certainty query strategy as presented in [Zhang & Schuller 2012]. By using the medium certainty level, data including acoustic distortions or labels obtained from unreliable annotation data are being disregarded for a re-evaluation, as they would most certainly lead to low confidence values but not contain the desired high informativeness. From more recent investigations, as presented in [Abdelwahab & Busso 2019] and [Chen & Hao 2020], it can be seen that the performance of AL is not only dependent on the type of emotional data (categorial/ discrete or dimensional/ continuous) but also on the utilized machine learning algorithm for the classifier.

Another method to reduce the annotation effort is based on eliminating the intervention of a human annotator completely, for example by utilizing Semi-Supervised Learning (SSL) strategies. An overview on most relevant SSL strategies is given in [Zhu 2008]. As for AL, a first model is trained on already existing labeled data samples and iteratively re-trained by the re-evaluated sub set of the unlabeled data pool. In contrast to AL, the re-evaluation of the sub set is not based on human expert labeling but relies completely on the results of the initially trained model. The sub set is chosen by evaluating the certainty value of the predicted class. Only those samples achieving high certainty values are considered in the sub set. This well-established SSL strategy is called *Self-Training*. Another approach is the so-called *Co-Training* [Blum & Mitchell 1998], where the decision is not based on one model, but on two models, initialized based on two independent feature sets which are sufficient to train a good classifier. The sub sets are then chosen independently for each model and the classifiers are re-trained using the obtained sub set of the other model.

The main problem with the presented approaches of AL and SSL is that the sub set they are trained on is only partly based on annotations done by human experts. For the in-vehicle real-world data presented in this Thesis, it can be assumed that AL and SSL will not lead to satisfying results, as the data set comprises not only noisy data but also highly natural low-expressive emotions for which an automatic recognition is highly challenging and of lower reliability (cf. Section 2.6). Therefore, a manual annotation is inevitable to achieve reliable recognition performances.

As SSL relies completely on the performance of the initially trained model, even when the whole unlabeled data pool is labeled by the model, the algorithm would not reach the performance of the classifier trained on the fully manually annotated set. From [Zhang et al. 2015] it can be seen that the performance of the SSL self-training and co-training algorithm never reaches the Unweighted Average Recall (UAR) (cf. Section 2.2.7) of the corresponding baseline Support Vector Machine (SVM) classifier (cf. Appendix A) for discrete speech emotion recognition, but converges towards a much lower UAR with an increased number of labeled instances (difference in-between UARs ranging from 3.3% to 6.7%).

For AL, studies have shown that the number instances which have to be manually labeled can be slightly decreased while receiving a similar UAR as when utilizing the fully manually labeled data set. In [Han et al. 2013], depending on the AL strategy, at least 88% manually labeled samples out of the unlabeled data pool are needed to achieve a recognition performance comparable to the results obtained on the fully manually labeled data utilizing Support Vector Regression (SVR) on continuously labeled values of valence and arousal. In a different study, by applying AL strategies for discrete speech emotion recognition and SVMs as classifiers, a reduction of the number of manually labeled data up to 85% of the unlabeled data pool is achieved (i.e. at least 15% manually labeled data) while the UAR is kept at a similar level as when utilizing the complete unlabeled data pool [Zhang et al. 2015; Zhang & Schuller 2012]. However, the UAR when manually labeling all samples of the unlabeled data pool is never outperformed (depending on the annotation quality this phenomena could also occur). It should be kept in mind that the presented numbers do not include the amount of manually labeled data needed to initialize a first classifier. Taking into account this amount of needed labeled data, 55.9% of data samples have to be manually labeled. Therefore, the goal of the approach presented in this Thesis, was not to limit the number of samples labeled by a human labeler, but to decrease the annotation effort (mostly in time) while receiving a full manual annotation. This approach will be presented in Chapter 5 later in this Thesis.

2.1.5 Relevant Benchmark Data Sets

There exist various well-established benchmark data sets that are commonly used in the field of speech emotion recognition. As already stated earlier in this Section, the data sets can be roughly distinguished regarding their level of naturalness (e.g. acted emotions, scripted emotions, natural emotions), their recording setup (e.g. anechoic chamber, studio, in the wild) and the utilized emotion concept (e.g. categorial emotions, dimensional emotions). An extensive survey article providing information on currently available benchmark data sets is, for example, presented in [Akçay & Oğuz 2020]. I will now introduce the reader only to those data sets which were further utilized in the scope of this Thesis, namely, EmoDB [Burkhardt et al. 2005] and VAM [Grimm et al. 2008]. Table 2.1 gives an overview on the characteristics of the EmoDB and VAM data sets.

Berlin Emotional Speech Database (EmoDB)

The EmoDB data set [Burkhardt et al. 2005] contains 494 emotional speech samples spoken by ten German speaking professional actors (five females) and recorded inside an anechoic chamber. The actors are instructed to simulate ten sentences of emotionally neutral content in seven emotion categories (anger, boredom, disgust, fear, happy, neutral and sadness). This results in a total of 800 emotionally colored speech samples. These recordings were afterwards annotated by 20 independent labelers according to their emotion recognizability and level of naturalness. The final data set only includes those samples with an emotion recognition rate of more than 80% and rated naturalness of more than 60%, resulting in 494 speech samples (127 anger, 79 boredom, 38 disgust, 55 fear, 64 happy, 78 neutral and 53 sadness) of approx. 2 seconds length and a total of 22 minutes of recorded audio material. The audio recordings are obtained using a sampling rate of 48 kHz but are provided to the user in a downsampled version of 16 KHz.

Vera am Mittag Corpus (VAM)

The VAM corpus [Grimm et al. 2008] is a German audio-visual emotional data set containing 946 speech samples (47 minutes of audio recordings) of 47 non-professional speaker (36 females). It consists of recordings taken from the German talk-show *Vera am Mittag*, where the guests of the show perform unscripted spontaneous discussions moderated by the anchorwoman, Vera. The recordings are

Table 2.1: Overview on benchmark emotional speech data sets employed in the scope of the Thesis.

Name	Reference	Samples [#]	Cat.	Dim.	Naturalness	Rec. Environment
EmoDB	[Burkhardt et al. 2005]	494	7	-	acted	anechoic chamber
VAM	[Grimm et al. 2008]	946	-	3	scripted	television studio

recorded inside a television studio with a sampling rate of 44.1 kHz but are later downsampled to 16 kHz. Even though the discussions are named to be unscripted and spontaneous, the discussions often escalate quickly which leads to a strong expressiveness of the emotions. Furthermore, the discussions mostly evolve around negative topics, which makes it assumable that the data set comprises a majority of negative emotions. The audio recordings of 12 broadcasts of the talk shows are divided into utterances (complete sentences, exclamations, affect burst or grammatically incomplete sentences) and afterwards annotated by 17 independent labelers in the three dimensions of valence, arousal and dominance using the SAM scale. A histogram, showing the distribution of samples along the three dimensions, provided in [Grimm et al. 2008] confirms the previously made assumption that a majority of the annotated samples lie in a region of neutral or negative valence. In [Schuller; Vlasenko; Eyben et al. 2009] the authors provide a mapping of the provided labels onto four quadrants (q1, q2, q3, q4) of the valence-arousal space, as depicted in Figure 5.1 on page 142. Considering this mapping, the VAM corpus contains 21 samples in q1, 50 in q2, 451 in q3 and 424 in q4. With q3 and and q4 lying in the negative valence half-space, this indicates a strong bias of the dataset towards negative emotions.

2.2 Modeling a Speech Emotion Recognition System

As soon as suitable data is available, it is possible to build a classification model for the considered speech emotion recognition task. The approach presented in this Thesis is based on supervised data-driven machine learning algorithms. A schematic overview on how to build this kind of systems is shown in Figure 2.5. First, emotion-relevant speech features need to be extracted from the speech signal. Then, those features contributing most to the recognition task need to be identified by applying a feature reduction algorithm. Afterwards, the data is split into an independent train and test set. The independence of the data sets is inevitable to overcome overfitting of the classifier. The train set is then used to train the classification model. Finally, the classification model is optimized (i.e. via hyper-parameter optimization, typically on a held-out evaluation data set) and validated using the reduced feature set of the independent test set. In this Section, I will provide a detailed description on how to proceed in the process of building a speech emotion recognition system. This will be done by clearly differentiating speech emotion recognition from Automatic Speech Recognition (ASR) and giving deeper insight on the feature space used for speech emotion recognition, suitable machine learning based classification algorithms and finally validating the classification model. This information will later be used to realize the classification experiments presented in Chapters 4, 5 and 6.

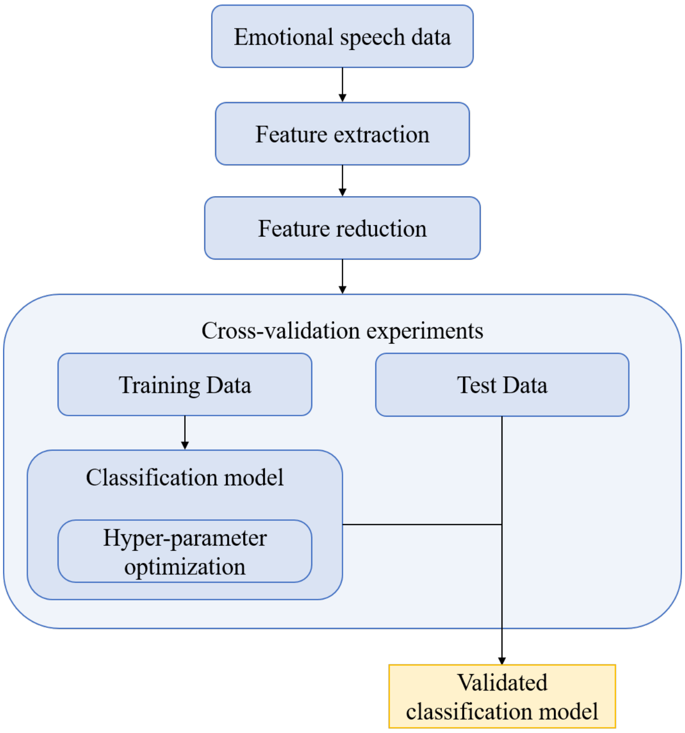


Figure 2.5: Schematic overview of the procedure to build a speech emotion recognition system.

2.2.1 Introduction to Automatic Speech Recognition

To highlight the main differences and similarities between speech emotion recognition and ASR, I will now give a brief introduction in the field of ASR. This introduction includes the general architecture of the recognition system, the utilized features, and most commonly used machine learning approaches, known from literature.

The general classical architecture of an ASR system, as described in [Yu & Deng 2015], consists of four main processing units. First, a pre-processing of the speech signal including the feature extraction is performed. Second, these features are fed to the acoustic model to estimate the likelihood of the recognized phonetic unit. Third, a language model is utilized, which estimates the probability of a hypothesized word sequence. All available words are taken from a dictionary, which can vary depending on the application domain or language the ASR system is designed for. Last, the likelihood of the phonetic unit sequence and hypothesized word sequence are combined and the hypothesis search component renders the word sequence obtaining the highest probability score.

Commonly used acoustic features for speech recognition, as state in the literature (cf. [Benesty et al. 2008; Yu & Deng 2015]), are Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Prediction (PLP) coefficients, and Linear Predictive Coding (LPC) coefficients, together with their deltas and delta-deltas. For speech emotion recognition, features and statistics are calculated frame-wise. While for speech emotion recognition the decision of the feature space strongly influences the performance of the recognizer (cf. Section 2.2.2), this is only to a small degree the case for speech recognition. Here, the performance for a continuous speech recognition system is mostly influenced by the chosen acoustic and language model, and utilized prescribed dictionary. First, the feature vectors are passed to the acoustic model and a phonetic likelihood is estimated. One main problem which needs to be stated at this point is the strong variability of time duration of each utterance. Even when repeating the same sentence multiple times, each utterance will have a unique length [Deng & Li 2013]. This implies that the same sentence will also result in a variable feature vector length. By applying techniques like Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs), this problem is solved [Yu & Deng 2015], by learning the time-alignment of the training and test phrases. A commonly used method to recognize a word sequence is by combining traditional machine learning techniques like Gausian Mixture Models (GMMs), SVMs or Artificial Neural Networks (ANNs) with HMMs in so called hybrid models. Some approaches also utilize ANNs (e.g. Deep Neural Networks (DNNs)) only, these approaches are, however, in need of large training material to obtain reliable recognition results with high recognition performances (cf. [Padmanabhan & Premkumar 2015; Solera-Ureña et al. 2007] and [Deng & Li 2013]). To further simplify the recognition task, the utilized dictionary is used to limit the number of possible entities (phonemes, syllables, etc.) that can be recognized by the acoustic model. Second, the language model is used to gain information on the probability of a hypothesized word sequence ($P(W)$), with regard to the correct, or commonly used, grammar of the sentence (cf. [Yu & Deng 2015]). This information is typically taken from text-based training corpora. By introducing this additional language model to the recognition process, the high degree of freedom of a recognized word sequence by the acoustic model is limited to possible word sequence candidates. This simplifies the process of continuous speech recognition, compared to using an acoustic model only, which would need an horrendous amount of training material. It is assumed that certain word sequences appear with a higher probability than others, and some may not be covered by the language model at all, leading to a considerably low probability of the recognized word sequence in these cases. The word sequence probability is then determined based on the probability of a certain word following the past n words:

$$P(W) = \prod(w_k \mid w_{k-1}, w_{k-2}, ..., w_{k-(n-1)}). \tag{2.10}$$

By utilizing this, so called, n-gram model, the probability of a word w_k, given the preceding $n-1$ words, is estimated [Wendemuth 2004; Benesty et al. 2008]. n-gram models are often described using finite-state models, with one state for every word w_i and a transition between the words forming a grammatically correct word sequence. Each transition between the words is weighted by a transition probability, estimated for the present word sequence [Benesty et al. 2008].

As this Section only gives a very brief overview on ASR systems, I recommend to read [Benesty et al. 2008] or [Yu & Deng 2015] for more detailed information.

2.2.2 Choosing the Feature Space

In the classical approach of speech emotion recognition, the feature space is based on the features used for emotion perception by humans. Here, in contrast to ASR, not the spoken content of the speech itself is of interest to gather information on the emotional state of the speaker, but the prosodic information (paralinguistic features) is needed [Frick 1985; Scherer 1986a]. Early research has already shown that it is not only possible to communicate emotions through prosodic information within one culture, but also in between cultures and languages [Clynes & Nettheim 1982; Davitz & Beldoch 1964; Krauss et al. 1983]. However, there exist differences in the communication of emotion of different cultures [Kramer 1964; Sogon 1975]. Even though the cross-cultural communication of emotion is less applicable for some cultures, a highly effective communication is still noticed [Frick 1985]. For within culture emotion communication, it is even shown that with increasing time of communicating (e.g. roommates sharing the same household) the ability to interpret the prosodic information correctly increases and hence also the ability to perceive the emotion correctly [Hornstein 1967]. It is further shown that with loss of prosodic information, the emotion perception is impaired [Lieberman & Michaels 1962; Knower 1941; Dusenburg & Knower 1939; Pollack et al. 1960; Burns & Beier 1973; Ross et al. 1973; Kramer 1964]. This is especially the case when information on the pitch or loudness of the speech is degraded (cf. [Lieberman & Michaels 1962]) or frequencies above a certain frequency-level are filtered from the signal (cf. [Burns & Beier 1973; Ross et al. 1973]). Some relevant features and their correlation to certain discrete emotional states are presented in [Frick 1985; Scherer 1986a], with a special focus on automatically extracted features in [Scherer et al. 2003]. A broad overview on the past findings in vocal emotions is presented in [Murray & Arnott 1993] and, with a focus on speech emotion recognition, in [Cowie et al. 2001]. It can be summarized that especially the fundamental frequency (pitch), pitch-contour, lower formant frequencies, loudness, intensity, speaking rate and voice quality have great impact on the perceived emotion. Furthermore, [Breitenstein et al. 2001] show that a manipulation of these features can significantly affect the perceived emotional state.

Considering only the paralinguistic features, in [El Ayadi et al. 2011] a grouping into four feature categories is presented: continuous features (e.g. pitch, formants and loudness), spectral features (e.g. MFCCs and LPCs), voice quality features (e.g. tense, harsh and breathy voice) and Teager Energy Operator (TEO)-based features (cf. [Teager & Teager 1990] and [Kaiser 1990]). The latter ones are primarily used for the detection of speech under stress condition [Zhou et al. 2001; Cairns & Hansen 1994]. Stress not per se defines an emotion but rather a transactional process requiring psychological, physiological and/ or behavioral effort by the individual to retrieve her/ his personal well-being related to an event, which is perceived as relevant to the individual [Lazarus & Folkman 1984]. Nevertheless, several well known-researchers in the field of emotion science have investigated the relation between stress and emotion and show that stress is interdependent with the field of emotion, especially considering negative emotions (e.g. [Lazarus 2006; Scherer 1986b]). In the scope of this Thesis, the effect of stress on the emotional state of the driver is not further investigated. Furthermore, in case of voice quality features, there occur two main challenges. Firstly, the voice quality is mostly described using impressionistic labels like tense, harsh and breathy [Cowie et al. 2001]. These are subjective labels, which needs to be judged auditorily by an individual and can only to some limit be extracted from the speech signal automatically. In [Laver 1980] a wide range of phonetic variables is presented, which are assumed to impact the subjective impression of the voice quality. Out of these measures a direct relation to emotions is drawn for the open-to-close ratio of the vocal cords, jitter, harmonics-to-noise ratio and spectral energy distribution [Klasmeyer & Sendlmeier 1995]. A simple way to include information of the voice quality is therefore to utilize spectral properties, which are already assumed to give relevant information on the emotional state, as stated previously. As this information can be seen as redundant, a unique objective measure of voice quality needs to be found. This can be done by recovering the glottal waveform from the speech signal, which is a highly challenging process, as it is neither a measurable signal nor are the vocal track filter parameters known [El Ayadi et al. 2011]. By utilizing inverse-filtering techniques, as presented in [Gobl & Chasaide 2003], it is, however, possible to estimate the glottal waveform. A second challenge of using voice quality as emotional feature is that from literature there exist contradictory statements on the relation between voice quality and the perceived emotion (cf. [El Ayadi et al. 2011; Scherer 1986a; Cowie et al. 2001]). Considering these findings on TEO-based features and voice quality features, the most relevant features when it comes to speech emotion perception are continuous and spectral features. These features can be obtained straight from the speech signal in time or frequency domain and are therefore of high interest when it comes to automatic speech emotion recognition. For further elaboration I want to refer to [El Ayadi et al. 2011; Swain et al. 2018] and [Akçay & Oğuz 2020], who give a good overview on speech features and how they influence the recognition performance of emotions.

A further aspect that needs to be addressed is the fact that speech signals are transient signals with all their features changing constantly over time. To obtain reasonable feature values, the signal from which the features originate needs to be quasi-stationary. This is accomplished by dividing the speech signal into small consecutive overlapping time windows (i.e. frames) of the same length, in which the speech signal is assumed to be short-time-stationary. This process is referred to as windowing and is a common methodology in digital speech processing (cf. [Wendemuth 2004]). The emotion speech features are calculated for each sample of the considered time window. To aggregate the changes occurring within the time window, statistics, such as mean, min, max, range and other variability coefficients, are determined. Consequently, for each speech sample a vector of each feature and statistic is generated, resulting in a high dimensional feature matrix. There exist multiple designated feature sets based on the paralinguistic features presented before, which are commonly used for speech emotion recognition. These features sets are described using so-called Low-Level Descriptors (LLDs) and statistical functionals. While the LLDs correspond to the global feature (supra-segmental), which is extracted directly from the speech signal, the utilized functional correspond to the applied statistics (super-segmental). An overview on benchmark feature sets is given in the next Section.

2.2.3 Feature Sets

In literature there exist multiple feature sets, which are used in the field of speech emotion recognition. The most commonly used feature sets are referred to as baseline feature sets and comprise the *emobase* and *emo large* feature set, provided by the OpenSMILE toolkit [Eyben et al. 2009; Eyben et al. 2010], the *IS'09 Emotion* feature set, provided by the INTERSPEECH 2009 emotion challenge [Schuller; Steidl & Batliner 2009], the *IS'10 Paralinguistic* feature set, provided by the INTERSPEECH 2010 paralinguistic challenge [Schuller et al. 2010], the *ComParE'13* feature set, provided by the INTERSPEECH 2013 computational paralinguistic challenge [Schuller et al. 2013], and the more recently established *GeMAPS* feature set, which is strongly based on features related to the phonetics of emotional speech [Eyben et al. 2016]. The *GeMAPS* feature set is the first approach towards phonetic based features. There exists a minimal and an extended version of the set, with the minimal set containing only features based on the phonetics and the extended set containing further spectral features like the MFCC 1-4 and formant bandwidth. These benchmark feature sets can be automatically extracted from the speech signal/ sample by utilizing the OpenSMILE feature extraction toolkit [Eyben et al. 2010], which provides designated configuration files for each set. OpenSMILE was used throughout this Thesis to extract the utilized feature configurations.

The main difference of the presented features sets is the number of LLDs and functionals, which are applied to the speech signal. Table 2.2 gives an overview on the different feature sets and the number of included LLDs and functionals. It needs to be noted, that not all functionals are applied to each LLD and that not only the LLDs but also their deltas and delta-deltas (i.e. first and second order derivatives or differences) are applied in some cases. For a more detailed description of the feature sets, please refer to the stated references. Further, it can be seen that a majority

Table 2.2: Overview on benchmark speech emotion feature sets sorted by year of publication. All feature sets comprise a different number of LLDs and applied functionals, with not all functionals necessary being applied to each LLD.

Name	Reference	LLDs [#]	applied functionals [#]	$\sum$
emobase	[Eyben et al. 2009; Eyben et al. 2010]	26	19	988
emo large	[Eyben et al. 2009; Eyben et al. 2010]	56	39	6552
IS'09 Emotion	[Schuller; Steidl & Batliner 2009]	16	12	384
IS'10 Paralinguistic	[Schuller et al. 2010]	38	21	1582
ComParE'13	[Schuller et al. 2013]	64	61	6373
GeMAPS	[Eyben et al. 2016]	18	16	62
eGeMAPS	[Eyben et al. 2016]	25	16	88

of the available feature sets comprise large numbers of features. Only the *GeMAPS* contains a low number of features which were designatedly chosen and are assumed to carry the most valuable information regarding speech emotion recognition.

In general, when choosing one of these benchmark feature sets, one needs to consider the *curse of dimensionality* (cf. [Bellman 1961]). The number of utilized LLDs and statistical functionals in the set define the complexity of the considered feature space. Regarding the number of available data samples for each class of the considered classification problem, the feature space, and all its combinatorial possibilities ($\#LLDs^{\#statistics}$), needs to be sufficiently covered by the data samples. With an increasing number of features and hence complexity of the feature space, also the number of utilized data samples needs to be increased dramatically. Without an increase of sample size, the feature space is only sparsely covered and it becomes much easier to find clusters of data samples inside the feature space, which can lead to an overfitting of the classifier [Spruyt 2014; Keogh & Mueen 2017]. However, there exists no simple way to determine a suitable number of feature space dimensions or data samples, as many factors contribute to this problem. One mayor factor is the correlation of the features included in the set. Assuming all features contribute the same amount of information to the recognition performance of the classifier, at least N^l samples are needed to densely cover the entire feature space, with N being the number of samples needed to sufficiently cover one dimension in the feature space and l being the number of dimensions considered [Theodoridis & Koutroumbas 2009]. This assumption is, however, not always valid, as for most cases there exists a correlation in between the features with some features contribute more to the classification task than others. For this case, it can be assumed that a much lesser number of data samples is necessary to overcome the cures of dimensionality. In literature there exist three common ways to cope with this issue (cf. [Clarke et al. 2009; Theodoridis & Koutroumbas 2009; Keogh & Mueen 2017]):

1. Cross-validation:
   The sample set is divided into two independent sets of training and test data. The classifier is then trained on the one set and afterwards applied to the other set. By testing on an independent set of data samples, one can prevent overfitting. If the classifier works well on the training set but does not show a good performance on the utilized test set, overfitting is most probably the case.

2. Feature extraction:
   The feature space gets reduced by combining features, which show a high correlation among each other into one feature. By combining different features, the interpretation of the new feature, as physical parameter of the production system, is, however, not possible anymore.

3. Feature selection:
   Only those features showing high importance for the present classification task and, hence, are most informative are being selected. By utilizing a feature selection those features contributing the most to the classification task are identified and an interpretation of the paralinguistic information is still possible. Features not contributing to the classification task are excluded from the set.

With regard to the problem of high dimensional feature spaces and small data sets, *emobase* shows a good performance on the most prominent benchmark emotional data sets presented in Section 2.1.5 (cf. also [Haider et al. 2021]). To make the results presented in this Thesis comparable among each other and to other investigations presented in literature, I therefore opt for this feature set. Even though the *GeMAPS* feature set comprises a much lower number of features and seems to be of higher suitability in case of small data sets, this set was outperformed by all other benchmark feature sets when applied for classification on the benchmark data sets (cf. [Eyben et al. 2016]). Furthermore, the contribution of different features to the recognition performance can vary dramatically depending on the utilized data set (cf. [Siegert et al. 2017]). By conducting a feature selection, the most relevant features of the present classification task can be identified. Another side effect of performing a feature selection is an increase of the recognition performance, as irrelevant features or redundant features are being excluded from the feature set.

To get a better insight on the utilized feature set, I will now give a brief overview on the features included in the *emobase* feature set. The set contains a large range of prosodic (e.g. pitch, loudness), spectral (e.g. MFCCs, Line Spectral Pairs (LSPs)) and voice quality related features (e.g. voicing probability), which are named to have a high impact on emotion recognition from speech (cf. Section 2.2.2). In detail 26 LLDs are include in the set. Furthermore, a broad number of 19 statistical

Table 2.3: LLDs and applied functionals of the *emobase* feature set provided by the OpenSMILE feature extraction Toolkit (adapted from [Requardt; Ihme et al. 2020]).

LLDs (26 x 2)	Functionals (19)	Features [#]
Pitch, pitch envelope, intensity, loudness, 12 MFCCs, probability of voicing, 8 frequencies of LSPs, Zero-Crossing-Rate (ZCR)	min/ max value and index of position within the signal, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, 3 quantiles and 3 inter-quantile ranges	(26 x 2) x 19 = 988

functionals are applied to each LLD and their deltas. This sums up to 988 features included in the *emobase* set. A detailed listing of the utilized LLDs and functionals is presented in Table 2.3.

2.2.4 Feature Reduction

As already stated, one major problem in using existing feature sets is their pretension on applicability for emotion recognition from speech in general. However, depending on the classification task and utilized data set, some features may be of higher importance than others and some features may show a high correlation in their contribution on the classification task. A second problem, which was stated just previously in this Section, is the curse of dimensionality, and that with an increase of utilized features also the number of data samples needs to be increased exponentially to densely cover the whole feature space. Therefore, it is advisable to perform a feature reduction and identify those features carrying the most importance for the considered classification task or combine features showing a strong correlation. This will on the one hand decrease the number of features and can on the other hand have an impact on the recognition performance (cf. [Özseven 2019] and [Daneshfar & Kabudian 2020]). Feature reduction will play a decisive role in Chapter 6 of this Thesis. In the following, I will refer to the two most commonly used methods of feature selection and feature extraction/ generation, which have been introduced above, jointly as feature reduction.

By conducting a feature extraction/ generation, an information loss of the paralinguistic information occurs, as features with a high correlation are combined and a new feature is generated. In the present Thesis, I focused on feature selection, as I wanted to identify those paralinguistic features contributing the most to a good recognition performance of the present real-world in-vehicle speech samples. It should be noted that in the following I will only give a brief overview on past and present work in feature reduction, as depending on the utilized data and feature set, the effect of the applied feature reduction method on the classifiers performance can largely vary. If the reader is interested in more detailed information on the recognition performance obtained when applying different feature reduction methods, I refer to the publications referenced.

The most popular *feature extraction* methods presented in literature are *factor analysis*, Linear Discriminant Analysis (LDA), Principle Component Analysis (PCA) and Independent Component Analysis (ICA) (cf. [Clarke et al. 2009; Theodoridis & Koutroumbas 2009]). In speech emotion recognition the most prominent feature extraction algorithm is PCA. [Daneshfar & Kabudian 2020] present a comparative study utilizing PCA, probabilistic PCA, LDA and *factor analysis*, as well as a novel approach based on quantum-behaved particle swarm optimization. They further compare their results obtained on the data samples of the EmoDB and IEMOCAP Corpus (cf. [Busso et al. 2008]) with the results of several other publications utilizing these data sets. Further investigations based on PCA are presented in [You et al. 2006a; You et al. 2006b; Sidorov et al. 2015; Siegert et al. 2015] and [Lee & Narayanan 2005; Chen et al. 2012; Xu et al. 2015], with a special focus on noisy speech environments in [You et al. 2006b]. LDA and *factor analysis* is applied in [You et al. 2006a; You et al. 2006b] and [Xu et al. 2015; Wu et al. 2011], and in [Wang et al. 2012; Song et al. 2015] and [Desplanques & Demuynck 2018], respectively. Although most of the investigations presented in literature show promising results, one major disadvantage of feature extraction is the remaining high computational effort, as all features need to be determined to apply the mentioned extraction methods. This is not the case for feature selection, where only relevant features need to be determined for a later application of the designed classifier.

The list of publications with a focus on *feature selection* is even larger. It can be distinguished between so called *wrapper* and *filter* methods (cf. [Theodoridis & Koutroumbas 2009]). Broad survey articles, covering mainly feature selection approaches, are provided by [Chandrasekar et al. 2014; Swain et al. 2018] and [Akçay & Oğuz 2020] and give a good overview on utilized feature selection methods in the field of speech emotion recognition from the past two decades. I will now briefly describe the two approaches of feature selection, with a special focus on *wrapper* methods, as the utilized feature selection method presented in Chapter 6 is based on this approach.

Wrappers are based on the selection of those feature subsets which contribute the most to a good classification performance [Theodoridis & Koutroumbas 2009; Stańczyk et al. 2018]. It is distinguished between greedy algorithms, like forward and backward feature selection, and genetic algorithms. In case of forward selection, features contributing the most to the classification performance are subsequently added to the feature subspace. In backward selection, features contributing less to the classification performance are excluded from the feature set and the remaining subset is kept as the optimal set. Contrary to these greedy methods, the genetic algorithm is based on evolutionary principles like mutation and selection of features [Goldberg 1989]. As the selected features are chosen based on the performance of the applied classifier, the generalizability of the selected feature sets is rather low. Furthermore, the presented methods are highly resource consuming, as an individual classifier needs to be validated for each selected feature or feature group (i.e. trained and tested). However, since *wrapper* methods are designed for one particular combination of classifier and dataset, the selected feature subset will most certainly outperform any other set applied to this constellation regarding their classification performance. Most common *wrapper* techniques used in literature are: Sequential (Floating) Forward/ Backward Search (S(F)F/BS) and genetic algorithms. Investigations based on greedy algorithms are presented in [Kwon et al. 2003; Ververidis et al. 2004; Luengo et al. 2005; Lin & Wei 2005; Özseven 2019; Schuller; Müller et al. 2005; Rong et al. 2009; Pérez-Espinosa et al. 2012; Planet & Iriondo 2012; Egorow et al. 2018] and [Lee et al. 2001; Lee & Narayanan 2005; Wu et al. 2011; Schuller & Lang et al. 2005]. The utilization of genetic algorithms seems to be still less common in the field of speech emotion recognition. Some publications based on these methods are presented in [Schuller; Reiter et al. 2006; Sedaaghi; Ververidis et al. 2007; Sedaaghi; Kotropoulos et al. 2007; Hübner et al. 2010; Sidorov et al. 2014] and [Sidorov et al. 2015].

The *filter* approach is based on statistical analysis of the features included in the feature set and is independent of the recognition performance and utilized classifier. To determine the most relevant features, distance and correlation measures are utilized to determine the class separability (cf. [Theodoridis & Koutroumbas 2009]). With known class labels (i.e. supervised learning), either the distance between the considered classes is determined, or the assumption that a good feature subset must contain features which are highly correlated within one class and uncorrelated among the different classes, is used to obtain information on the class separability. These approaches tend to be much faster and of stronger generalizability than *wrapper* methods, as the feature subset is chosen by only evaluating the intrinsic properties of the data. This implies that by applying a *filler* method it is possible to receive a feature subset which performs good when applied to a broad range of classifiers, but which will most certainly be outperformed by a selected feature subset when utilizing a *wrapper* method on one particular classifier. Most prominent filtering techniques used in literature are Fisher's discriminant ratio (FDR) (cf. [Theodoridis & Koutroumbas 2009]), distance-based, correlation-based and statistical testing based methods. Filter methods based on FDR are utilized in [Sun et al. 2019; Harimi & Esmaileyan 2014; Liu et al. 2018] and [Chen et al. 2012; Wu et al. 2011]. In [Vogt & André 2005; Schuller 2011; Mencattini et al. 2014] and [Özseven 2019] a focus is drawn on correlation-based filter techniques. Feature selection based on distance measures is presented in [Wang & Guan 2004] and [Liu et al. 2018], and statistical testing based methods utilizing ANOVA and the Kolmogorov-Smirnow test are presented in [Sheikhan et al. 2013] and [Ivanov & Riccardi 2012].

It should be noted, that the presented studies on feature reduction only cover investigations based on speech emotion recognition. Furthermore, this listing is only an abstract of the available work and should not be taken as comprehensive.

2.2.5 Feature Normalization

From the field of phonetics it is known that there exist differences in the individual speaker characteristics with regard to form, size and mass of the vocal organs and their state of health [Scherer et al. 2003; Laver 1994]. These differences also affect the characteristics of the features used for speech emotion recognition. A common way to cope with this problem in speech emotion recognition is to normalize the feature values originating from the different speakers. A comparative study on efficient ways to normalize speech features is presented in [Böck et al. 2017]. In this paper, the authors apply different normalization techniques to nine emotional speech data sets. As normalization methods they utilize *standardization* (transformation of the measured values into standard values with an expected value of 0 and variance of 1), *range normalization* (transformation of the measured values into a range interval of [-1, 1], often also referred to as min-max feature scaling), *centering* (shifting the measured values to a mean value of 0; only the value changes not the scale) and *neutral normalization* (determination of the normalization parameters only on samples labeled as neutral and then applied to samples labeled as other emotion [Busso et al. 2011]). By using SVM classifiers (cf. Appendix A) and Leave-One-Subject-Out (LOSO) cross-validation (cf. Section 2.2.7), it is shown that feature normalization can lead to a significant increase of recognition performance compared to when utilizing non-normalized features. It is assumed that by normalizing the feature values, the inter-individual effects of the speaker on the speech signal are adjusted while the effects of the emotion on the speech signal is maintained. For all nine evaluated data sets the recognition performance is increased when applying standardization. Two sets achieve the highest recognition results when applying standardization in combination with natural normalization. Only for one data set the best result is obtained by applying range normalization. Centering never leads to an increase of recognition results. In contrast, this normalization approach even shows a decreasing behaviour for four of the nine data sets. The strongest increase in recognition performance is achieved for EmoDB, where the non-normalized features accomplish 47.4% UAR, while the standardized features reach 77.2% UAR. Overall it is concluded that standardization in most cases leads to substantial improvement of the recognition performance for all evaluated data sets. Considering this finding, for all classification experiments presented in this Thesis a feature normalization using standardization was realized.

2.2.6 What Kind of Classifier to Use

In contrast to ASR, where also the spoken content of the speech is needed, speech emotion recognition can be based on paralinguistic features only. There does exist work, which includes the detection of emotional keywords into the speech emotion recognition system (e.g. [Chuang & Wu 2004]). This, however, will not always lead to an increased recognition performance, as emotional phrases like "I'm happy." or "Everything is Okay.", do not always refer to the intrinsic state of the speaker [Chopade 2015]. The assumption of not considering the spoken content differentiates speech emotion recognition significantly from ASR, as no vocabulary, language model or correct grammar is needed.

In literature it is assumed that the emotional state of the speaker can change rapidly over time, even within one sentence [Pell & Kotz 2011]. Therefore, the classification of the speech signal is mostly performed over small speech segments, for which the emotional state of the speaker is assumed to be constant. Hence, in the field of speech emotion recognition, mostly a focus is draw on the detection of the emotional state in a considered discrete time interval. To also include the evolution of emotion over time it is possible to utilize HMMs or DNNs (e.g. [Schuller et al. 2003; Ntalampiras & Fakotakis 2011]).

In case of data driven machine learning approaches, the classifier will only be as good, as the data it is trained on. This is one of the main challenges considered in this research domain, as the generation of valuable data is costly in resources and time. Therefore, it is essential to make use of classification algorithms, which can cope with the available amount of data. Most commonly used machine learning approaches in speech emotion recognition are SVMs, decision trees (e.g. Random Forests (RFs)), ANNs, DNNs, Long Short-Term Memorys (LSTMs) and autoencoders. Even though a comparatively high number of emotional data was obtained during the data collection realized in this Thesis (cf. Section 5.1.2 on page 140) compared to the number of samples included in benchmark emotional data set (cf. Section 2.1.5), this is not sufficient to obtain reliable results from ANN or DNN approaches. Therefore, I opt for SVMs and RFs as classifiers. A further positive aspect of these classification algorithms is that they can cope with biased-distribution data sets, as it is the case for the present data. There further exists a high number of comparable studies from other researchers, which also employed SVMs and RFs as classifiers, allowing to draw conclusions and comparisons. For more information on the SVM and RF machine learning algorithms, I refer the reader to Appendix A.

2.2.7 Evaluating the Recognition Performance

In the introduction of this Section, I have mentioned the validation of the classification model by utilizing an independent validation set, also referred to as test set. This set is used to determine the classification model performance on unknown data samples. There exist designated performance measures, used in the field of machine learning. Furthermore, when modeling multiple classifiers, it is of high importance to determine if there exist a statistical difference between the models. This can be done by applying suitable statistic analyses. The three steps of validating the model (test set), determining the performance of the model (performance measures) and determining the significance of the performance (statistical analysis) is presented hereinafter and will be used throughout this Thesis.

Validating the Model

There exist different methods to split the data samples into a training and test set. While the training set is used to train the classifier, the test set should solely be used to validate the model and should therefore not include samples already contained in the training process. Commonly used methods are random k-fold cross-validation (cf. [Schaffer 1993] and [Kohavi 1995]), LOSO cross-validation (cf. [Picard & Cook 1984] and [Schuller et al. 2008]) and Leave-One-Subject-Group-Out (LOSGO) cross-validation (cf. [Schuller; Vlasenko; Eyben et al. 2009]).

In case of k-fold cross-validation the data set is randomly split into k equal sample sets. Afterwards, each of the k sets is used to validate a model trained on the remaining $k - 1$ sets. This corresponds to k classification experiments. The performance measures obtained for each experiment are then averaged over all k experiments to estimate the average performance of a classifier trained on completely unknown data samples. A less random method to split the data set into more representative training and test sets, is *stratified* cross-validation. Here, the data distribution of the training and test set is chosen in the same proportion as in the underlying population (e.g. age or gender).

One disadvantage of the k-fold cross-validation is that a speaker dependency of the samples is left out of consideration. This could lead to an overfitting of the classifier, as the classifier could be tested on speech samples originating from a speaker already used during training, implying that the speaker is not unknown to the model. This problem can be overcome by utilizing a LOSO cross-validation. For this method, the data samples are split by their origin, such that the number of subjects/ speakers included in the data set, determines the number of folds and, hence, the number of performed classification experiments, with each fold containing only samples of one speaker. Furthermore, the LOSO validation scheme leads to a generalization of the classification results and takes into account the inter-individual differences of the subjects. As with an increasing number of subjects also the validation effort increases dramatically, another way to maintain speaker independence is by performing a LOSGO cross-validation. Here, the data is split into groups of independent subjects/ speakers, with a maximum limit of 10 groups, such that the number of folds and, hence, classification experiments is limited to 10 [3] (also see [Schaffer 1993] and [Kohavi 1995]). This decreases the validation effort while maintaining a speaker independence.

Determining the Performance of the Model

To determine the performance of the model, the predicted output obtained by the classifier is compared to the true label of the sample. There exist several measures to evaluate the performance of the model (cf. [Olson & Delen 2008; Powers 2007]). Most of these measures are established to evaluate binary classification system with the assumption that the outcome is either positive or negative. A frequent way to represent these results is by utilizing a confusion matrix, as depicted in Table 2.4. Here, it is distinguished between True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) classification results. The TPs entries correspond to the number of samples correctly predicted (green entries) as positive, TNs correspond to the number of samples correctly predicted as negative, FPs correspond to the number of samples falsely predicted (red entries) as positive and FNs correspond to the number of samples falsely predicted as negative.

By utilizing the terms defined in the confusion matrix, it is possible to determine different performance measures. The most frequently used measure is called Accuracy (ACC). This measure represents the percentage of correctly predicted instances and is determined by calculating

$$\mathrm{ACC} = \frac{TP + TN}{TP + FP + FN + TN}. \tag{2.11}$$

Its compliment is referred to as Error Rate (ER) and determined by calculating

$$\mathrm{ER} = 1 - \mathrm{ACC} = \frac{FP + FN}{TP + FP + FN + TN}. \tag{2.12}$$

Especially in case of unbalanced data set distributions the interpretation of the ACC and ER can be misleading, as the dominant class will bias the measure towards

Table 2.4: Confusion matrix of a binary classification problem.

		Predicted	
		Positive	Negative
True	Positive	TP	FN
	Negative	FP	TN

[3] This number of folds is at least needed to obtain a minimal statistic and can be seen as rule of thumb.

its detectability. Measures, which take into account such imbalances are Recall, Precision and F1-Measure. The Recall, also referred to as Sensitivity, determines the percentage of the correctly predicted positive samples out of all true condition positive samples:

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{2.13}$$

Its compliment is referred to as Specificity and determines the percentage of the correctly predicted negative samples out of all true condition negative samples. The Precision determines the percentage of the correctly predicted positive samples out of all predicted positive samples, i.e. the probability of the true prediction to be correct:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.14}$$

Its compliment is referred to as negative predictive value and determines the probability of the false prediction to be correct.

In general, Recall and Precision are correlated with each other and, to achieve a satisfactory performance of the classification model, a good trade-off between both measures needs to be achieved. A measure considering this correlation is called F-measure, which combines Recall and Precision using its harmonic mean and controls the trade-off by introducing the constant β. In most cases, however, β is chosen to be 1 and the corresponding F1-measure is determined as:

$$\begin{aligned} \text{F1-measure} &= \frac{(1+\beta^2)\cdot\text{Recall}\cdot\text{Precision}}{(\beta^2\cdot\text{Recall})+\text{Precision}}, && (2.15)\\ &= \frac{2\cdot\text{Recall}\cdot\text{Precision}}{\text{Recall}+\text{Precision}}. && (2.16)\end{aligned}$$

In case of multi-class classification problems, the terms in Table 2.4 can be directly adapted, by considering each class separately in a “one vs. all” manner, as exemplary shown in Table 2.5 for *class 1* of a three class classification problem. The performance measures can then be computed as described in Equation (2.11)-(2.15). In case of multi-class problems the performance measures are often stated as macro-average over all considered classes. These averaged values are then denoted as Unweighted Average Recall (UAR) and Unweighted Average Precision (UAP). The corresponding macro F1-measure is then determined by calculating Equation (2.15) using the UAR and UAP. In case of cross-validation results, the UAR, UAP and F1-measure, are also used to represent the average measures per class, macro-averaged

Table 2.5: Exemplary confusion matrix of a multi-class classification problem, representing the TP, TN, FP and FN of *class 1*.

		Predicted		
		class 1	class 2	class 3
True	class 1	TP_1	FN_1	
	class 2	FP_1	TN_1	
	class 3			

over all cross-validation experiments. In this Thesis, these results will be indicated accordingly, to avoid confusion.

Statistical Analysis of the Performance

Whenever multiple classification models are trained to solve identical classification problems, it needs to be determined how the models differ in their performance and if this difference can be seen as significant. This is done utilizing so called statistical hypothesis tests (cf. [Bortz & Lienert 2008] and [Bortz & Schuster 2010]). Depending on the type of data (parametric and non-parametric, paired and unpaired), number of factor groups and number of factors in each group, different types of statistical tests need to be utilized. These are, for example, t-tests, in case of parametric paired and unpaired data samples of two factor groups, and single factor ANOVAs, in case of parametric unpaired (one-way ANOVA) and paired (repeated-measures ANOVA) data samples of multiple factor groups. Furthermore, the data samples need to fulfill certain requirements before correctly applying said statistics. A more detailed description of relevant hypothesis testing methods and their requirements is presented in Appendix C of this Thesis.

2.2.8 Hyper-parameter optimization

In this Thesis a focus is drawn on the classification of highly natural and low expressive emotional data. It can be assumed that the recognition performance on this kind of data is comparatively low, unlike for a classification performed on well-established baseline data sets (e.g. EmoDB [Burkhardt et al. 2005] and DES [Engberg et al. 1997]) [Siegert; Lotz; Egorow; Böck et al. 2016]. To increase the performance of the classifier and push it to its limits, a hyper-parameter optimization is inevitable and will play a decisive role in Chapter 6 of this Thesis.

There are multiple ways to perform a parameter optimization. In this Section a brief overview on the most relevant parameter optimization methods is given. This includes *random search* as special case of the well-established *grid search*, *cross-validation* and *heuristic choices*. Of course there are also mathematical optimization methods known from non-linear programming, e.g. Simplex or Gauss-Newton algorithm [Grimme & Bossek 2018], which can be used to solve the given optimization problem. These approaches however are accompanied with high computational costs and, depending on the initial point and complexity of the problem, do not guarantee to find the most optimal solution in a reasonable computing time [Steinwart & Christmann 2008]. Because of the nature of the utilized data set, an implementation of this approach would be too time consuming and is therefore left unconsidered.

Grid Search (Random Search)

In *Grid Search*, for each of the to-be optimized hyper-parameters a search vector is defined in which the parameter values are distributed in an equidistant or geometrical way, such that they form a grid in the search space of all hyper-parameters [Steinwart & Christmann 2008]. Afterwards, for each combination of hyper-parameters classification experiments are conducted and the best performing parameter combination is chosen as optimal hyper-parameter set. To evaluate the different parameter combinations, this approach is often combined with cross-validation (described in Section 2.2.7 on page 44). One major problem of this approach is the *curse of dimensionality*, as the number of parameter combinations increases exponentially with the number of hyper-parameters [Bellman 1961]. Additionally, the narrowness of the grid plays an important role on whether it is possible to find the optimal combination of hyper-parameters. In case of a widely distributed grid, it might occur that only a local optimum or no optimal solution at all is found, as the performance of the classifier would not change significantly with respect to the other parameter combinations (cf. [Scheibner 2012]). By tightening the grid, the chance to find a solution close to the global optimum increases but at the same time the computational costs of the performance evaluation increases dramatically (cf. curse of dimensionality). Two simple measures taken to overcome this problem are:

1. starting with a widely distributed grid and tighten the grid only in the region of the search space where an increase of performance occurred (only possible if a significant change in performance was observed) or
2. changing the parameter distribution in the search space from equidistant/ geometric to random, also referred to as *random search* [Bergstra & Bengio 2012].

By randomly distributing the parameter combinations in the search space, the subspaces of the parameters are covered in a more sufficient way. Especially in parameter combinations for which one parameter is less important than the others, random search can lead to better results, respectively.

Cross-Validation

A cross-validation is performed as described in Section 2.2.7 on page 44. The hyper-parameters are optimized for each of the conducted experiments separately (e.g. by performing a grid search or choosing heuristically). The optimal hyper-parameters are chosen by selecting those hyper-parameters, which performed best on average over all the experiments. By averaging the performance measures of the individual models over all experiments it is possible to give an estimate of the classifier's predictive performance on unknown data. One disadvantage of this approach is that the sub sets are generated on a random basis. This implies that samples originating from the same subject can occur in both training and validation set of the classifier which can lead to overfitting, as described earlier. This problem can be avoided by performing a Leave-One-Subject-Out (LOSO) or Leave-One-Subject-Group-Out (LOSGO) cross-validation instead [Schölkopf & Smola 2003].

Heuristic Choices

The method of heuristic choices is based on educated guessing. In [Schölkopf & Smola 2003] several methods are listed for SVM hyper-parameter optimization ranging from using parameter settings, which worked well for similar problems, up to using theoretical approaches like considering the Vapnik-Chervonenkis Bound (cf. [Vapnik 2000]). A different approach based on theoretical considerations and empirical results is presented in [Cherkassky & Ma 2004]. For RF, [Hastie et al. 2009] give recommendations on which default values to use to obtain satisfactory classification results. However, they also state that the right parameter choice is strongly dependent on the considered classification problem and should therefore always be treated as tuning parameter. From only considering heuristic choices and making educated guesses it is most unlikely to find an optimal hyper-parameter combination.

Overall, it can be stated that each of the presented approaches can be used to tune the classifier's hyper-parameters with more or less success. Depending on the determining factors of the parameter optimization, certain approaches are more suitable than others. Especially with an increase of data, the use of *grid search* leads to extremely high computational costs. Therefore, the number of experiments which can be conducted in an appropriate time span is very limited. To be able to cover the search space in the best possible way while keeping the computational cost in an acceptable region, I opt for *random search* in combination with a LOSO cross-validation. To determine the search space of the hyper-parameter optimization, heuristic choices as well as predefined search intervals taken from literature were utilized in the scope of this Thesis.

2.3 Recent Findings in Speech Emotion Recognition

In the previous Sections I have already given an introduction to the concept of emotions in general (cf. Section 2.1.2) and emotions in speech in particular (cf. Section 2.2.2). It was stated that the first mentioning of *passion* and *mood* was already made by the philosopher Socrates and his student Aristotle in the ancient Greece, while the *classical* definition of emotions received increased attention from the 1960th onwards. It took another 35 years until Rosalind Picard founded the field of *Affective Computing*, also referred to as *emotion AI*, along with the same titled publication [Picard 1997]. Since then the research on automatic emotion recognition from different modalities (e.g. bio-physiological signals, facial expressions, gesture or speech) has evolved quickly.

The field of speech emotion recognition has gained increasing attention in the past 20 years of research [Schuller 2018]. Starting with emotion recognition from acted emotional speech in the early 2000s, the research focus is now drawn to emotion recognition in the wild. This leads to many challenging situations such as but not limited to, non-optimal recording conditions, non-stationary environmental conditions and low-expressiveness of emotions in natural everyday communication. A first large scaled comparative study with an attempt towards emotion recognition in natural communication environments is presented in [Schuller; Vlasenko; Eyben et al. 2009]. It is shown that speech samples obtained under non-optimal recording conditions of non-professional speakers can already lead to a decrease of recognition performance by 61.1% compared to similar classification experiments realized with speech under optimal recording conditions of professional actors (from 84.5% ACC for EmoDB down to 23.5% ACC for SmartKom [Schiel et al. 2002]). Several survey articles have followed since then, which emphasize the importance of a suitable emotion concept, data set, feature set and classifier, specially designed for the considered application domain (cf. [El Ayadi et al. 2011; Swain et al. 2018; Akçay & Oğuz 2020] and [Chen et al. 2012], only to name some). The term emotion recognition *in the wild* is, however, not used consistently. While in some publications the authors define *in the wild* as emotion recognition in a natural everyday spontaneous speech environment (e.g. in [Albanie et al. 2018; Avila et al. 2021] or as presented in the stated survey articles), others present cross-corpus evaluations on a mix of acted, scripted and spontaneous emotional speech, to achieve a high variability of the employed speech data and generate a general emotion classification model. In [Kaya & Salah 2016] it is shown, that by utilizing a cross-corpus evaluation the performance on the spontaneous speech data could not be increased (ACC = 34.20%, for a six-class classification problem). As *in the wild* data they employ the AFEW 4.0 data set (cf. [Dhall et al. 2014], which contains video and synchronous audio clips of emotional movie scenes annotated using the movies' transcripts for the visual impaired [Dhall et al. 2011]). However, their classifier is trained and tested on samples originating from this data set. In a comparable study, the authors of [Avots et al. 2019] draw a distinct line between data sets used for training and testing of the classifier. As test set, only samples of the AFEW 4.0 are utilized, while the training set contains samples originating from three benchmark data sets (eNTERFACE'05 [Martin et al. 2006], SAVEE [Haq & Jackson 2010] and RML [Wang & Guan 2008]). With this approach they achieve an ACC of 27.1%, utilizing the speech signal only. In [Kim et al. 2017] a cross-corpus evaluation among six data sets containing acted and natural emotional speech is presented. By conducting baseline within-corpus experiments on each data set (training and test data originate from the same data set), the authors achieve high recognition accuracies in case of acted emotional speech (range: 53.7% - 95.3%, depending on data set and applied classifier). In case of natural emotional speech, the recognition accuracies are noticeable lower (range: 40.9% - 56.9%, depending on data set and applied classifier). The corresponding cross-corpus experiments show an even stronger decrease in the recognition performance. Independent of naturalness of the emotional speech data the accuracies range from 32.7% to 49.8%, which corresponds to a decrease of performance for all data sets.

It can be summarized that by concentrating on the present application domain (within-corpus evaluation), the performance of a speech emotion recognizer is increased compared to cross-corpus evaluation experiments. Even though the cross-corpus evaluation takes into account the high variability of the speech material, to obtain high recognition results it is recommended to use data originating from the considered application domain. This is also reasonable, as not only additive environmental noises (e.g. background noises), but also manipulations of the speech signal through sound propagation, as described in Section 2.5.1, can degrade the speech signal and its features used for speech emotion recognition. With an increasing number of available emotional data sets originating from various application domains, a cross-corpus evaluation could, however, in the long term, lead to classification models of higher generalizability. This goal has not yet been reached, and with a special focus on the in-vehicle noise environment considered in this Thesis, an extensive domain dependent analysis is not available. The goal of this Thesis and especially Chapter 6 is to give novel insights on this field of research.

2.4 Speech Quality

As concluded in the previous section, especially for *in the wild* emotion recognition from speech, the speech quality plays a decisive role, as the environmental conditions influence the quality of the recorded audio signal. This includes the presence of background noise or other disturbances (cf. Section 2.5) and non-optimal recording setups. But what exactly is speech quality? Is it the pure potential to understand the speech content in terms of speech intelligibility of the signal, or is it the listening quality as level of pleasantness? At this point it needs to be stated, that speech quality and speech intelligibility are not the same. While intelligibility measures how well comprehensive the speech content is, the speech quality focuses on how well the speech signal is produced by the speaker in terms of naturalness, clarity, pleasantness, brightness, etc. [Loizou 2011]. Some investigations have shown that there exists a correlation between speech quality and speech intelligibility (cf. [Taal et al. 2009; Chiaramello et al. 2015; Prodeus et al. 2018] and [Ma et al. 2009] only to name some). Nevertheless, there exist distinct subjective and objective measures to evaluate speech intelligibility (e.g. Articulation Index (AI) [French & Steinberg 1947], Speech Transmission Index (STI) [Houtgast & Steeneken 1973; Steeneken & Houtgast 1980] and its successor Speech Intelligibility Index (SII) [ANSI/ASA 1997]) and speech quality (e.g. Mean Opinion Score (MOS) and Signal-to-Noise Ratio (SNR), cf. Section 2.4.1). In some investigations speech intelligibility is synonymously used with the word error rate [%], which describes the percentage of correctly understood speech content. This, however, is not the definition of speech intelligibility in general (cf. [ANSI/ASA 1997]).

A majority of investigations in the field of disturbed speech and in-vehicle noises utilize speech intelligibility as a measure to evaluate the effect of signal processing steps on the speech signal (e.g. speech enhancement or speech coding). This is reasonable, as these investigations focus on how well comprehensive the speech content is compared to the unprocessed speech signal. In case of speech enhancement, the intelligibility should be increased compared to the original noisy speech signal, whilst for speech coding a high intelligibility of the compressed speech signal should be sustained. It is assumable that for speech recognition, where the correct recognition of the textual information is of interest, a good speech intelligibility also leads to high recognition performances. Research conducted in medical science in the field of speaking disorders, where the pronunciation of words by the speaker is disturbed, confirms this relation (cf. [Maier et al. 2009] and [Schuster et al. 2006]). In [Gallardo et al. 2017] the authors show that for compressed speech, utilizing several audio codecs, a low word error rate is correlated with a high speech intelligibility. As the work presented in this Thesis focuses on the detection of the emotional state of a driver based on paralinguistic cues only, and the textual/ linguistic information is left unconsidered, I assume that speech quality measures are more relevant when it comes to speech emotion recognition, as they describe the quality of the signal itself. Therefore, I will now give a brief overview on speech quality assessment methods and introduce two speech quality measures (MOS and SNR), which are the most relevant for the investigations presented in this Thesis (cf. Section 2.4.1). Afterwards, a focus will be drawn on the speech quality of in-vehicle speech (cf. Section 2.4.2). Here, I will also address the relation between the subjective speech quality measure MOS and the objective measure SNR. Furthermore, a first insight on how the quality of noisy speech affects the speech emotion recognition performance will be given in Section 2.4.3.

2.4.1 Relevant Speech Quality Measures

To investigate the speech quality, a description of two quality assessment methods of speech audio recordings will now be given, and will be later applied in Chapter 4. This includes a common assessment method designed to evaluate the listening quality of speech (MOS) as subjective measure, and one method describing the quality from the signal side (global SNR in time-domain) as objective measure. One major problem when assessing the speech quality is the lack of a clear definition of speech quality. Good overviews on common subjective and objective measures for speech quality assessment are given in [Loizou 2011] and [Kondo 2012]. In these book chapters, the authors distinguish between subjective listening tests and objective quality measures. While the subjective listening tests are based on the subjective rating of human labelers asked to rate the quality of the original and processed speech signal, the objective quality measures are based on the numerical distance between the original and processed signal. Depending on the application domain (i.e. compressed speech, noisy speech) different approaches are recommended. The most commonly used subjective speech quality measure is MOS, which can be utilized for compressed speech, while for noisy speech an adaption of the listing test needs to be made (cf. [ITU-T 2003b]). As subjective listening tests are highly time and resource consuming, a realization, in the scope of this Thesis, was not possible. However, there exist possibilities to predict the test results regarding the MOS by utilizing a Perceptual Objective Listening Quality Assessment (POLQA) (cf. [ITU-T 2018]). As second objective speech quality measure, I opt for the SNR, which is widely used in the speech processing community. The original aim of the SNR is to calculate the ratio between the clean speech power and the power of the present background noise (cf. Equation (2.17) on page 54). There are possibilities, as described in [Kondo 2012], to apply the SNR also to compressed speech, where no increase in the signal power will occur. In this case, the power of the noise signal is defined as the absolute power difference between both signals and can, therefore, not be compared to SNR values obtained for noisy speech signals. Because of this difference in the definition of the SNR, when applied to compressed or noisy signals, the SNR will only be used in case of present background noise.

Mean Opinion Score

The Mean Opinion Score (MOS) is a subjective measure to evaluate the quality of a speech sample in terms of listening quality as defined in the ITU-T recommendation P.800 (08/1996) [4] [ITU-T 1996] for audio compression. It is based on a 5-point category-judgment scale ranging from 1 *bad* to 5 *excellent*. One way to obtain the MOS-value is by letting subjects rate the quality of the speech on the presented 5-point scale. This approach corresponds to a subjective measure obtained by human

[4] https://handle.itu.int/11.1002/1000/3638

raters and should be conducted as described in ITU-T P.800. Another way to assess the MOS-value is by utilizing a POLQA with regard to the ITU-T recommendation P.863 [5] [ITU-T 2018] (Version used in the scope of this Thesis: P.863 (09/2014) [ITU-T 2014]). It is designed to predict the overall listening speech quality of degraded speech samples compared to their uncompressed high quality samples as perceived by the user in a listening test, as defined in ITU-T P.800. This predicted value is referred to as MOS - Listening Quality Objective (MOS-LQO). However, it should be noted that it does not replace subjective testing. POLQA can be applied in two operating modes: narrowband (NB) (300 - 3400 Hz) and super-wideband (SWB) (50 - 14000 Hz). For the investigations presented in this Thesis, the SWB-mode was utilized. For this mode the MOS-LQO saturates at $MOS\text{-}LQO_{swb} = 4.75$. The MOS-LQO can be utilized for both, compressed and noisy speech signals.

Signal-to-Noise Ratio

The Signal-to-Noise Ratio (SNR) is defined as the unit-less ratio between the power of a speech signal and the power of the background noise [Benesty et al. 2009],

$$SNR = \frac{P_s}{P_n}, \tag{2.17}$$

where $P_s = \sum_{n=-\infty}^{\infty} |x(n)|^2$ denotes the total signal power of the clean speech signal $x(n)$ and $P_n = \sum_{n=-\infty}^{\infty} |d(n)|^2$ the total signal power of the background noise $d(n)$ under similar recording conditions. When considering the SNR in the logarithmic decibel scale,

$$SNR_{dB} = 10 \cdot \log_{10}(\frac{P_s}{P_n}) = 10 \cdot \log_{10}(P_s) - 10 \cdot \log_{10}(P_n), \tag{2.18}$$

the SNR_{dB} corresponds to the signals' log power difference of the clean speech and the present background noise. A schematic visualization of the SNR_{dB} is depicted in Figure 2.6. A positive SNR_{dB} value indicates that the total power of the speech signal is higher than the total power of the present background noise and vice versa for a negative value. This SNR is also referred to as *global* SNR, as it is based on the total power of the speech signal.

To get a rough feeling on how the SNR affects the speech intelligibility I will now briefly introduce the SII, which describes the intelligibility of speech in terms of audibility and usability (e.g. the clearness and comprehensiveness of the spoken content) (cf. [ANSI/ASA 1997]) not to be confused with the listening quality (MOS). The SII ranges between 0 and 1, with 0 implying that none of the speech information is available to the listener and 1 implying that all information is given to the listener.

[5] https://handle.itu.int/11.1002/1000/13570

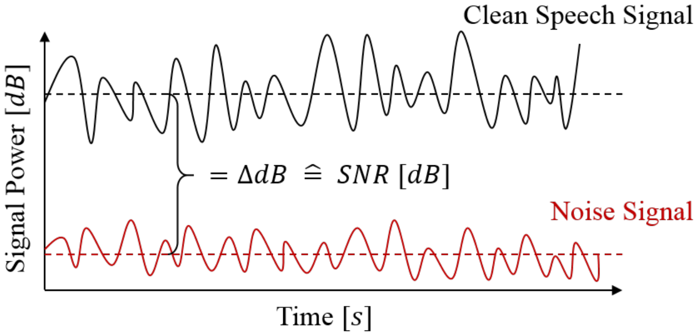


Figure 2.6: Schematic visualization of the SNR in the logarithmic decibel scale [dB].

In [Hornsby 2004] it is stated that with an SII of 0.5 a none-impaired listener is able to understand up to 100 % of the spoken content correctly. An increase of the SII is only given for SNRs up to 30 dB [Sauert et al. 2006]. Afterwards, there exists no audible difference for the listener in the correct understanding of the spoken content. A good communication system is indicated by an SII of 0.75 and higher. Poor communication systems show an SII of 0.45 and lower. However, the SNR level at which a speech signal is still clearly intelligible is strongly dependent on the acoustic characteristics of the room (i.e. vehicle interior) [Dong & Lee 2018] and the type of noise signal [Goli & Karami-Mollaei 2016]. Therefore, the SNR at which a certain SII-value is reached can strongly vary depending on the present background noise. From [Goli & Karami-Mollaei 2016] it can be seen, that for traffic noise an above poor communication is possible with an SNR of -5 dB and higher. A fairly good communication is possible from 0 dB onwards. In contrast to traffic noise, for factory, white and babble noise an above poor communication is scarcely reached at an SNR of 0 dB. The SNR itself can, additionally, be influenced by the recording conditions (e.g. distance speaker to microphone, reverberation times) [Hodgson & Nosal 2002], as well as the environmental conditions (i.e. road surface, vehicle velocity and/ or weather conditions). In [Botinhao & Yamagishi 2017] the influence of the road type and velocity of a passenger car on the SNR is addressed. Depending on these factors the SNR ranges from -25 dB under highway driving conditions up to 5 dB while parking. It can be assumed, that the SNR-values presented in this Thesis lie in comparable regions.

2.4.2 In-Vehicle Speech Quality

I will now address the relation between speech quality measures with a special focus on the in-vehicle noise environment. This field of research will also be addressed later in Section 4.3 in my own work.

While most of the presented work is based on out of the shelf car noises added with different SNRs, a small share of investigations is based on real recordings of in-vehicle noises under different driving conditions. For this kind of real in-vehicle noise data, a natural high variation of speech quality is present, due to changes in the velocity, road surface or vehicle type. This results in a natural Gaussian distribution of the SNR-values. By adding real noises with different SNRs this natural information is lost. A focus is drawn on these naturally occurring quality measure distributions.

A majority of the publications presented in this Section are based on additive car and traffic noises. Investigations based on real in-vehicle noises distributions, which are later added to well-established benchmark data sets, in this case with a focus on emotional speech, are presented in [Grimm; Kroschel; Schuller et al. 2007; Grimm; Kroschel; Harris et al. 2007; Schuller; Rigoll; Grimm et al. 2007; Schuller 2008; Tawari & Trivedi 2010a] and [Tawari & Trivedi 2010b]. One investigation based on speech recorded inside a real driving environment, is presented in [Botinhao & Yamagishi 2017].

In [Grimm; Kroschel; Schuller et al. 2007; Grimm; Kroschel; Harris et al. 2007; Schuller; Rigoll; Grimm et al. 2007] and [Schuller 2008] the same in-vehicle noise recordings are utilized. The noise recordings are obtained inside four different vehicle types (BMW 530i (Touring), 645Ci (Convertible), M5 (Limousine) and Mini Cooper (Convertible)) under three different driving conditions (city road, highway and big cobbles) and afterwards added to three benchmark emotional data sets (EmoDB [Burkhardt et al. 2005], VAM [Grimm et al. 2008] and eNTERFACE'05 [Martin et al. 2006]). The authors show that the SNR differs significantly depending on the utilized vehicle type, driving condition and data set. However, independent of the vehicle type and added emotional speech material, the lowest SNRs are obtained under big cobble road surface and the highest SNRs in case of the city driving condition. The SNRs obtained under highway driving condition lie in-between these two cases. It is further shown that the SNRs of the in-vehicle noises recorded inside the BMW 645i (Convertible) and BMW M5 (Limousine) are higher than the ones obtained for the remaining two vehicle types.

The authors of [Tawari & Trivedi 2010a] and [Tawari & Trivedi 2010b] focus on three different driving conditions (highway, parking lot and city street) but give no further insight on the vehicle they are recorded in. A detailed SNR distribution is only given in case of the highway and parking lot driving condition. After superimposing the recorded noise segments to the LISA-AVDB (cf. [Tawari & Trivedi 2010a] and Section 2.6.3), the SNR for highway driving, ranges from -10 dB up to 15 dB. The SNR obtained for the parking lot driving condition ranges from 0 dB up to 25 dB. This distinct difference is evident, considering the perceived noise level when driving on a highway compared to the more silent parking lot condition.

While the previous investigations are based on additive noise on emotional benchmark speech data, in [Botinhao & Yamagishi 2017] real SNR-distributions of in-vehicle speech are presented. The authors replay emotional speech data inside a running hybrid compact car under three different driving conditions (city road, highway and parking) and re-record the data using a head and torso mannequin. The replay volume of the speech material is chosen in a way that it is similar to real in-vehicle communication, with a constant volume in each driving condition. Considering this recording setup, in case of the highway driving condition the SNR ranges from -25 dB up to -10 dB. For the city route and parking driving condition the SNR follows a bimodal distribution of two overlaying Gaussian distributions. This is reasonable, as for the highway driving condition the speed of the vehicle shows only little variation, while for the city route condition, the vehicle is accelerated and slowed down at traffic lights and other traffic incidents. In case of the parking lot driving condition, the hybrid car would switch at low velocities to electric mode, which causes a much lower noise level compared to the petrol mode used at higher velocities. The bimodal Gaussian distribution of the city route condition ranges from -27 dB up to -3 dB and from -5 dB up to 0 dB, respectively. For the parking lot condition the highest SNRs are ranging from -14 dB up to 0 dB and from -5 dB up to 5 dB, respectively. This represents a similar tendency as the results presented in the previous paragraph.

As MOS describes the subjective listening quality of the speech, most investigations analyze the relation between the objective SNR and the subjective MOS. This is for example done in [El-Solh et al. 2007]. Here, the authors investigate the listening quality of car noises added to a clean speech signal with different SNRs. They identify that with an increasing SNR also the MOS-LQO increases. Similar findings were made in [Sharma et al. 2010] and [Gelderblom et al. 2019]. In [Sharma et al. 2010] the authors further identify that the MOS-LQO converges towards a saturation for noises of 30 dB onwards. It can be summarized that for all the presented investigations the relation between MOS and SNR is striking, meaning that an increase of SNR is always accompanied by an increasing MOS. This information will be used later in Section 4.3 to validate the performance of the novel Compression Error Rate (CER) (cf. Section 4.1.2).

2.4.3 Relation between Speech Quality and In-Vehicle Emotion Recognition Performance

The most commonly used speech quality measure when it comes to emotion recognition from speech in noisy environments is the SNR. The number of publications covering the relation between the listening quality MOS and recognition performance is, however, still limited. One more recent investigation, not considering in-vehicle noises, is presented in [Avila et al. 2018]. Here, the authors investigate the effect of speech enhancement on speech quality and the ability to automatically recognize the speaker's emotional state. As data the RECOLA data set (cf. [Ringeval et al. 2013]) containing airport babble noise with different SNRs is utilized. They identify a clear increase of recognition performance for arousal with increasing MOS and SNR of the unprocessed and speech enhanced data samples, respectively. For the recognition of valence a much lesser increase of recognition performance is achieved for the unprocessed samples. For the speech enhanced samples no increase of recognition performance (of value) is observed. For some cases even a slightly reverse behavior is noticed (decrease of recognition performance of valence with increasing MOS and SNR), but this has not further been analyzed. In the previous Section, I stated that there exists a clear relation between MOS and SNR. This is in line with the findings made in [Avila et al. 2018]. Therefore, it can be assumed that both measures also have a similar impact on the speech emotion recognition performance.

First results based on in-vehicle noises are presented in [Grimm; Kroschel; Schuller et al. 2007; Grimm; Kroschel; Harris et al. 2007; Schuller; Rigoll; Grimm et al. 2007] and [Schuller 2008]. They all utilize the same in-vehicle noise recordings (cf. Section 2.4.2), which are added to the clean speech signal. As emotional speech data the EmoDB [Burkhardt et al. 2005], VAM [Grimm et al. 2008] and eNTERFACE'05 [Martin et al. 2006] data sets are utilized. When training the classifier on clean speech samples and testing it on unknown noisy speech samples (cross-validation), a clear increase of recognition performance with increasing SNR is observed for the VAM and eNTERFACE'05 speech samples. In case of the EmoDB data samples, an almost random behavior, without a clear relation between recognition performance and SNR, can be observed. This, however, could be attributed to the nature of the data samples, which contains data recorded under ideal recording conditions inside an anechoic chamber.

In [Tawari & Trivedi 2010b] similar classification experiments, utilizing the EmoDB speech samples but different in-vehicle noise recordings, are presented. In contrast to the previously presented investigations, not the natural SNR distributions of the noise recordings are evaluated but the recordings are added to the clean speech signal with predefined SNRs (5 dB, 10 dB and 15 dB). For these experiments a clear increase of recognition accuracy is observed with an increasing SNR. The authors further evaluate a data set of real in-vehicle recordings and scripted emotions (LISA-AVDB, see [Tawari & Trivedi 2010a] and Section 2.6.3). However, they only collect speech data inside the vehicle without driving and other in-vehicle disturbances. Afterwards, the same noise recordings as utilized for the EmoDB data set are added to simulate a real driving environment. As for the experiments performed on the EmoDB samples, a clear relation between SNR and recognition accuracy is noticed.

Another investigation based on additive in-vehicle noises in different SNRs (0 dB, 5 dB, 10 dB and 15 dB) is presented in [Chenchah & Lachiri 2016]. Here, the authors add different pre-recorded noise types to the well-established IEMOCAP data set (cf. [Busso et al. 2008]). Considering the experiments carried out using in-vehicle noises, the worst recognition performance is achieved when adding noise with a 0 dB SNR. An increase of recognition performance is shown up to 10 dB SNR. With an SNR of 15 dB the performance decreases slightly, but never below the performance obtained with 0 dB. The authors of the work, however, did not comment on these findings. Nevertheless, the presented results could be interpreted as a saturation of the recognition performance.

A more detailed insight on the classifiers performance, in relation to the recognition of emotions, obtained in the stated publications is presented in Section 2.6.1. Furthermore, these findings are used as requisite of the investigations presented in Chapter 4 to identify statistical evidence of the relation between the speech quality and the speech emotion perception/ recognition performance.

2.5 Disturbed Speech

To get a better insight on the nature of disturbed speech, a brief overview on different noise types and their occurrence will be given. In the remainder of this Thesis, I will focus on two types of external disturbances. These are *acoustic disturbances* where the speech signal is affected by external factors (i.e. acoustic characteristics of the room and additive environmental noises) and *signal degradation* based on the manipulation of the speech signal by a signal processing unit (i.e. speech coding, speech enhancement). Internal disturbances (e.g. interference or crosstalk[6]), influenced by the recording equipment, will not be further addressed. As the main focus of the Thesis lies in the evaluation of emotional speech in noisy in-vehicle environments, I will first describe the effect of acoustic noises and their influence on speech emotion understanding and recognition. Realistic emotional speech data in an in-vehicle environment is, however, hardly available (cf. Section 2.6.3) and difficult to establish (cf. Section 2.1). Therefore, I will further give insights on the signal degradation occurring in speech coding technologies. Contrarily to noisy speech, this data is easy to generate by applying the desired audio codec on existing emotional speech data and will be exploited in Section 4.2 later in this Thesis. Furthermore, the degradation of the speech signal, defined as changes in the power spectrum and waveform (i.e. canceling certain frequencies out of the spectrum), can be assumed to have a similar effect on the performance of a speech emotion recognizer as changes occurring through other disturbances.

2.5.1 Noisy Speech

Acoustic disturbances can be roughly distinguished between disturbances caused by the acoustic characteristics of the room and additive environmental noises. While the first describe the manipulation of the signal through sound propagation inside the room, the second is defined as an additional sound signal disturbing the primary speech signal.

[6] electromagnetic interference of transmission channels [Slavik 2008]

In general, the sound propagation from point A to point B inside a room can be described by introducing the impulse response $h(n)$. The term $h(n)$ takes into account all the manipulations of the sound signal arising from being produced at point A to being observed at point B. This includes reflections of the sound signal on the walls of the room, absorption of sound through the walls and diffraction of sound around obstacles. Depending on the dimensions, material and interior and hence the acoustic characteristics of the room, the impulse response can differ significantly. With the assumption of superposition and time-invariance of the sound signal, the observed signal $y(n)$ at point B can now be calculated as

$$y(n) = x(n) * h(n) \tag{2.19}$$

with $x(n)$, as the excitation signal produced at point A (i.e. clean speech signal), convoluted with the impulse response $h(n)$.

In case of additive environmental noises a second sound source $d(n)$ (disturbance signal) is added to $x(n)$. Now, the room acoustic influences both sound sources and the impulse response is, therefore, convoluted with both, $x(n)$ and $d(n)$:

$$y(n) = x(n) * h(n) + d(n) * h(n). \tag{2.20}$$

This equation is further simplified to

$$y(n) = x(n) + d(n), \tag{2.21}$$

with $x(n)$ and $d(n)$ already including the convolution with the impulse response of the room, if not indicated differently. Common background noises, evaluated in literature are: babble noise, factory noise, airplane noise and vehicle noise (**NOT** in-vehicle noises!). The definition of Equation (2.20) plays a decisive role when it comes to the simulation of specific noise types and will be further discussed in Section 2.6.1 and 2.6.2. Furthermore, this information emphasizes the necessity of a real-world data collection (cf. Chapter 3), as the specific in-vehicle environment strongly influences the impulse response in Equation (2.20).

Disturbances in the In-Vehicle Environment

I will now give a brief overview on factors influencing the recording of in-vehicle data to highlight how important a suitable data collection is to receive reliable and reproducible results of the investigations presented in this Thesis. As described previously in this Section, the impulse response $h(n)$ is influenced significantly by the material, the dimension and the interior of the room. This is also the case for the in-vehicle acoustic characteristics. Depending on the type of vehicle (e.g. convertible car, station wagon, truck, ...), the size of the cabin can already differ significantly and, hence, influence the sound propagation through the room. Furthermore, the material of the body and interior lining plays a decisive role, especially when it comes to absorption of sound and transmission of environmental noises, as their absorption factors can vary largely. For example, soft-cover convertible cars have much lower absorption factors than hard-cover convertible cars or station wagons and, therefore, the environmental noises from outside of the vehicle are transmitted much stronger.

Typical in-vehicle noises which arise while driving in a convenient gasoline powered vehicle are engine sounds, exhaust noise, noises of the wheels on the road surface, noises of the body flow around and noises of the surrounding traffic [Lerch et al. 2009]. While the noise of the wheels on the road surface and the body flow around are of higher relevance while the vehicle is moving with a speed of over 30 to 40 km/h, the engine sound and exhaust noise also occur while the vehicle is at stillstand or moving slowly. When driving with a higher speed the most observable noise originates from the body flow around, which increases with the sixth power of the speed [Lerch et al. 2009]. The noise of the wheels on the road surface is also strongly dependent on the road surface itself. Uneven road surfaces will more certainly lead to an increased noise pollution than even road surfaces.

Speech Emotion Perception/ Recognition from Noisy Speech

In this Thesis I primarily focus on the effect of in-vehicle noises on speech emotion recognition. Therefore, a more in-depth consideration on the in-vehicle noise environment is presented in separate Sections of this Chapter. In Section 2.6 the focus is drawn to speech emotion recognition in noisy in-vehicle environments and Section 2.4 emphasized already the relation between speech quality and speech intelligibility and the effect of speech quality/ intelligibility on speech emotion recognition, especially in the field of in-vehicle speech. Speech intelligibility and quality, should, however, not be confused with speech emotion perception/ intelligibility, as emotion intelligibility refers to the correct understanding of the emotional content and is hardly related to the correct understanding of the speech content. Even without a clear understanding of the spoken content a correct perception of the emotional content is possible (cf. Section 2.2.2).

With a focus on in-vehicle speech, speech emotion intelligibility is still a less investigated field of research. To my knowledge, there only exist a few publications focusing on speech emotion intelligibility in general. In [Parada-Cabaleiro et al. 2017] and [Malik et al. 2020] the authors evaluate the ability of emotion perception by humans employing the GEMEP data set (cf. [Bänziger et al. 2006]) and three different noise types (white, pink and brownian noise). The noise is added in three different SNR levels (+3 dB, +1 dB, -0.5 dB and -1 dB). In [Parada-Cabaleiro et al. 2017] it is shown that the best emotion intelligibility is achieved when no noise is added to the signal, independent of the considered emotional state (cold anger, elated happiness, hot anger, panicked fear, pleasurable happiness and sadness). For a majority of the evaluated emotional states, pink noise achieves the lowest perception accuracy by the human labelers followed by white noise. The best results are always received under brownian noise. For this perceptual study 50% of the subjects are native German speakers (13 subjects). The rest of the subjects belong to seven other nationalities (India, Tunisia, Spain, Iran, Mexico, Russia and United Kingdom). Even though utilizing the same data set under comparable noise conditions and a similar experimental setup of the perceptual study, the results obtained in [Malik et al. 2020] differ significantly. These differences can be attributed to the fact that different participants are employed to participate at the study. Unfortunately, no detailed information on the nationality of the participants is given. Surprisingly, the authors state that the best results are not always achieved when listening to the clean (no noise added) emotional speech sample. This fact is, however, hardly addressed in the discussion of the publication. A third study investigating the effect of background noise on speech emotion intelligibility is presented in [Dmitrieva & Gelman 2012]. In this journal article, it is shown that the ability to detect a certain emotion correctly (joy, neutral and anger) decreases when background noise is added to the clean speech signal. However, it is very unfortunate that the authors do not state what kind of background noise was added to the clean signal.

A more deeply investigated field of research is automatic speech emotion recognition from noisy speech (cf. [Schuller; Arsic et al. 2006; You et al. 2006b; Pao et al. 2007; Schuller; Rigoll; Grimm et al. 2007; Schuller; Seppi et al. 2007; Yeh & Chi 2010; Yang et al. 2014; Zhao et al. 2014; Chenchah & Lachiri 2016; Satt et al. 2017; Avila et al. 2018; Avila et al. 2021; Bashirpour & Geravanchizadeh 2018], only to name some). In these publications the authors utilize speech from different languages (e.g. German, English, Danish, Chinese mandarin and Persian speech) and add different kind of noises to the clean speech samples (mostly babble and white noise with different SNRs). In some cases the noise is added to the clean speech signal, while for some investigations the speech signal is already recorded inside the noisy environment (e.g. by employing the SUSAS [Hansen & Bou-Ghazale 1997; Schuller; Arsic et al. 2006] or AIBO corpus [Batliner et al. 2004; Schuller; Seppi et al. 2007]). The importance of this differentiation is addressed in Section 2.6.1 and Section 2.6.2. Furthermore, the ground truth used to evaluate the reported emotion recognition tasks is based on the original emotion labels obtained under clean speech conditions. A comparison to the ability of a human to recognize the emotional state under the considered noise conditions is hardly addressed. In general, without applying any further speech enhancement, it is shown that for a majority of the presented investigations the accuracy of the recognition system increases with an increased SNR and that a recognition on the noisy speech samples is mostly outperformed by the recognition on the corresponding clean speech samples.

2.5.2 Denoising - Speech Enhancement

One major challenge of noisy speech, especially in case of additive background noises, is their negative impact on the speech quality and intelligibility (cf. Section 2.4) [Kunche & Reddy 2016]. It affects the ability of humans to communicate in a noisy environment and consequently also the ability to perceive the speaker's emotional state from the communicated prosodic information (see Section 2.2.2). To increase the speech quality and intelligibility of the degraded speech signal, speech enhancement algorithms, also referred to as noise reduction algorithms, are applied. This reduces the effect of noise for speech communication and can improve the performance of speech applications such as speech compression and speech recognition [Kunche & Reddy 2016]. Nevertheless, it is not possible to reduce the noise without also affecting the original clean speech part included in the noisy speech signal. Hence, the applied speech enhancement algorithm should consider the right tradeoff between speech distortion and noise reduction [Boll 1979]. In Section 5.2 of this Thesis, I will focus on this research question and investigate, if speech enhancement has a positive impact on the present speech emotions recognition task. In the following paragraphs I will therefore introduce the reader to the most relevant speech enhancement methods.

There are multiple factors which influence the performance and, hence, decision on the optimal speech enhancement systems. These are, for example, the number of available noise sources and the number of speech sources corrupted by said noise [Banchhor et al. 2013]. The different enhancement systems can be distinguished by the number of available input channels, the domain of processing (i.e. time or frequency domain) and type of algorithm (i.e. adaptive or non-adaptive) [Kunche & Reddy 2016]. In literature it is typically only distinguished between single-channel and multichannel speech enhancement systems [Loizou 2007]. While single-channel systems only employ one audio-channel (i.e. one single microphone), multichannel systems are based on employing multiple microphones either separated in space, or combined into one microphone array. I will now give a short introduction into single-channel and multichannel enhancement systems and their application domain.

Single-channel speech enhancement systems are commonly used for speaker and speech recognition, mobile communication and hearing aids [Kunche & Reddy 2016]. In these application domains a second microphone is usually not available. This single-microphone system is, compared to multichannel systems, easy to build and less expensive. Nevertheless, they have some major disadvantages, as no dedicated microphone to continuously pick-up the noise signal only is available and, therefore, no adaptive noise cancellation is possible. The main assumption made by single-channel systems is that the present noise signal is stationary over the speech intervals. This approach does not take into account the natural variation of the noise signal over time. The noise only audio parts are used to determine the noise signal, which is afterwards subtracted from the noisy speech audio signal to estimate the clean speech signal. A good overview on speech enhancement methods applicable for this kind of setup is given in [Loizou 2007].

For adaptive noise cancellation at least two microphones need to be integrated into the system (i.e. multichannel enhancement system), one primary microphone, which receives the noisy speech signal, and one reference microphone receiving the noise only signal, which is uncorrelated with the speech signal. This noise only signal is put through an adaptive filter, which is adjusted automatically by feeding the system's output (i.e. denoised speech signal) back into the filter [Widrow et al. 1975; Chhikara & Singh 2012; Kunche & Reddy 2016]. The most commonly known application of this kind of adaptive noise cancellation algorithm is used in noise canceling headphones, where it is easy to pick up the noise signal only. Whenever it is not possible to integrate a reference microphone, which is uncorrelated with the speech signal, it is recommended to utilize a microphone array. In general, the larger the number of microphones inside the array, the easier the speech enhancement becomes [Loizou 2007]. Typically, the design of the microphone array is conditioned by the used enhancement method and restrictions given by the application environment. One commonly used speech enhancement method is based on broadband beamforming [Nordholm et al. 2014]. The general approach of beamforming is based on the delay in the signals' time of arrival at each microphone inside the array. The synchronized signals' components form the desired direction of arrival (i.e. direction of source/ speaker regarding the microphones beampattern) are then summed together to obtain an enhanced speech signal (delay-and-sum beamformer). However, this approach is mostly based on the assumption of a far-field source, for which it can be assumed that the source's direction of arrival is approx. the same for all microphones inside the array [Benesty et al. 2016]. This can either be achieved by increasing the distance between array and source or by decreasing the distance between the microphones inside the array. Adaptions of the algorithms to be used in the near-field or mixed near- and far-field are possible, but are not in the scope of this Thesis [Doclo & Moonen 2003]. For speech and audio signals, the bandwidth can range from $60Hz$ to $20kHz$. As the beampatterns of a microphone varies with different frequencies, it is not possible to design a simple delay-and-sum beamformer. Here, a so-called broadband beamformer needs to be utilized. This can be realized by using nested microphone sub-arrays, where each sub-array is designed for one specific frequency range [Benesty et al. 2016].

Speech Enhancement and its Impact on Speech Emotion Perception/ Recognition

The effect of speech enhancement on speech emotion perception and recognition is still a scarcely investigated field of research. Some investigations focus on emotion recognition from enhanced speech and compare these results to the results obtained under noisy speech conditions. Depending on the utilized data set, type of noise, feature set and speech enhancement algorithm the results of these studies differ significantly. While in [Tawari & Trivedi 2010b] speech enhancement always leads to an increase in recognition performance, in [Triantafyllopoulos et al. 2019] the recognition results obtained from the clean speech signal always outperform those from the enhanced signals. In [Pohjalainen et al. 2016; Avila et al. 2018] and [Triantafyllopoulos et al. 2019] the same data set is employed, namely RECOLA (cf. [Ringeval et al. 2013]). However, all studies utilize different speech enhancement algorithms, feature sets, noise types and classification algorithms, hindering a reasonable comparison of their results. Nevertheless, a clear consistency in the recognition performance among the experiments is shown, with a majority of the experiments performed on the clean speech signal outperforming the other classification experiments. Only in [Pohjalainen et al. 2016] the results obtained when applying cepstral noise reduction outperform the results of the clean speech experiment. In [Chenchah & Lachiri 2016] the authors do not compare their results with the results obtained from the clean speech signal. Depending on the utilized noise type the results differ significantly when applying different speech enhancement algorithms. One investigation, also stating an emotion dependent recognition performance, is presented in [Xiaoqing et al. 2017]. When utilizing the EmoDB data samples (cf. [Burkhardt et al. 2005]) with additive white noise in different SNRs and SVMs as classifiers, the authors show that depending on the considered emotional state the recognition performance can increase (anger, neutral), decrease (fear, happy) or stay approx. the same (sad) when using the speech enhanced data samples. They further apply a feature selection filter method (cf. Section 2.2.4) on the utilized feature set and evaluate the recognition performance dependent on the utilized number of features. The authors state that with an increasing number of features also the emotion recognition performance tends to increase.

However, none of these investigations utilize the effect of the applied speech enhancement algorithm on the feature space, which was used to automatically recognize the emotional state. This is a highly risky approach, as speech enhancement is based on the manipulation of the noisy speech signal and provides only an estimate of the original clean speech signal. This implies that the signal may sound well intelligible, but has significant differences when it comes to the actual features characterizing the signal, compared to the original clean speech content. In [Chenchah & Lachiri 2016] the authors investigate the effect on the utilized features when applying speech enhancement. It is shown that when employing Gammatone Frequency Cepstral Coefficients, the recognition performance obtained under enhanced speech condition decreases compared to the results obtained from the noisy speech signal, while for MFCCs the results always increase with enhanced speech. This, however, only gives a marginal insight on the effect of speech enhancement on the feature space. Therefore, this issue will be addressed in the own work in Section 5.2 of this Thesis. Furthermore, the results presented in the current Section are only based on noisy speech obtained by artificially adding different noise types to benchmark data sets. I will later explain why this artificially added noise is not per se comparable to speech data obtained under noisy recording conditions (cf. Section 2.6.1 and Section 2.6.2).

2.5.3 Compressed Speech

As stated earlier in this Section, the generation of compressed speech is, in contrast to real-world noisy speech, simple to generate. Therefore, first investigations, presented in Section 4.2 of this Thesis, were done utilizing compressed speech signals. Depending on the application domain, compression is either used to reduce the space needed to store a certain amount of data, or to reduce the transmission bandwidth and latency of a certain amount of data. The first aspect is of large interest, especially, in the era of "Big Data", where more and more data is produced and stored [Shen et al. 2016], and, more classically, when only having a limited amount of data storage available (e.g. storing the maximum amount of music on a CD while maintaining a high listening quality). Depending on the data type, a compression can lead to a significant decrease in needed storage space. While text does not need a large amount of space to be stored uncompressed, this is different for video or audio data, where already for a small amount of high quality recordings a large space is required [Salomon 2007]. The second aspect - reducing the transmission bandwidth and latency - is primarily used in the field of mobile communication. In this field, a low transmission latency is needed to experience a transparent interaction [ITU-T 2003a]. Acceptable transmission latencies, as defined by the ITU-T in recommendation G.114[7], range from 150 ms up to 400 ms, depending on the communication type (end-to-end or one-way). To maintain this limitation, the amount of data communicated via the network in the desired time needs to be limited. This is achieved by compressing the original audio signal and thereby reducing the data size. With regard to in-vehicle speech emotion recognition, the second application domain is of greater interest, especially when it comes to online machine learning solutions [Fontenla-Romero et al. 2013] where the existing classifier is continuously updated with new incoming training material such that it dynamically adapts, for example, to the current user. These solutions are mainly realized using vehicular cloud computing, which is based on a wireless communication between the vehicle sensors/ electronic control units and a cloud platform providing computing power and data storage [Whaiduzzaman et al. 2014]. For this approach, it is reasonable to not only transmit the processed signal values to the cloud platform, but to transmit the raw sensor signal, as it enables an adaptation of the signal processing components (e.g. feature extraction, signal enhancement) inside the cloud without the vehicle being physically present. Therefore, it is of great interest to investigate how much the audio data can be compressed while maintaining a high usability for a later application in machine learning with a focus on speech emotion recognition.

[7] http://handle.itu.int/11.1002/1000/6254

Audio Coding Technologies

In this Section, the most prominent audio coding technologies used in everyday applications, such as Voice over LTE (VoLTE), Voice over Internet Protocol (VoIP), Skype, Spotify, Amazon Music and Apple Music, are presented. Standard audio codecs can be grouped in three categories based on different compression technologies and their application domain (music vs. speech). These are Analysis-by-Synthesis (AbS), perceptual coding and hybrid coding. Depending on the application domain (i.e. data storage as used in music compression, and audio transmission as used in mobile and internet telephony) certain compression technologies are more or less suitable. In case of music storage a focus is drawn on maintaining a high listening quality while reducing the file size. For these applications perceptual coding, based on the human auditory system, is used. Signal parts, which the human ear is unable to perceive, are being discarded in the compressed signal. In case of mobile and internet telephony a focus is drawn on low transmission bandwidths and latencies in context with high speech intelligibility. Codecs applied in this research area mostly use AbS, which is based on a closed-loop optimization technology. Newer audio codecs can also operate in a, so called, hybrid mode based on both AbS and perceptual coding and distinguish between speech and non-speech audio parts automatically or make use of both techniques at the same time. In the following a broad overview on the technical functioning of AbS and perceptual coding will be given. The presented algorithms are mainly based on the descriptions presented in [Chen & Thyssen 2008], [Herre & Lutzky 2008] and [Sinder et al. 2015]. All information based on different references will be indicated as such. A more detailed description of the most relevant audio codecs that utilize the introduced coding technologies is given in Appendix B.

Perceptual coding is motivated by the perceptual properties of the human auditory system and is mostly used in codecs designed for music compression. It aims to represent the original audio signal in a more compact way while maintaining the original perceived sound quality. Consequently, a focus is drawn on the optimization of the perceived speech quality. This is done by exploiting irrelevances in the signal and discarding those signal parts, which are supposed to be beyond the resolution of the human auditory system, in the compressed audio signal. This is realized by utilizing filter bank-based audio coding technologies. They consist of an analysis filter bank that maps the speech signal to a spectral representation. The most commonly used filter banks are Modified Discrete Cosine Transforms (MDCT) and hybrid structures of other filter banks used in combination with MDCT [Brandenburg et al. 1992]. Furthermore, a perceptual model is utilized, which estimates the signal's time- and frequency dependent threshold of perceptibility, by considering psycho-acoustic effects like masking in frequency domain. The spectral values obtained from the filter banks are then quantized and coded with regard to the threshold of perceptibility obtained by the perceptual model. The result is packed into a bit stream and transmitted to the decoder. The size of the bit stream is defined by the chosen bit rate, which describes the number of bits transmitted in one second ([bit/s]). Hence, with increasing bit rates also the compression ratio decreases and the compressed sound contains more information of the uncompressed sound signal. The decoder consists of a decoding of the bit stream and inverse quantization. Finally, a synthesis filter bank maps the spectral values back into time-domain. Most prominent codecs, relevant for this Thesis, that utilize perceptual coding are MPEG-1/MPEG-2 Audio Layer-3 (MP3), Advanced Audio Coding (AAC) and Windows Media Audio (WMA) (cf. Appendix B).

In contrast, AbS coding is based on waveform-approximating coding (waveform coders) and is mostly used in codecs designed for speech communication. An AbS standard which utilizes waveform coders is Codec-Excited Linear Prediction (CELP), which uses closed-loop optimization for en- and decoding of the speech signal. To understand the AbS waveform coding, first the basic linear predictive speech waveform coding will be introduced. This approach is then adapted to the basic AbS coder. During the encoding process of linear predictive coding, the speech signal is compared to a predicted version of the signal. The residual between the two signal is calculated and quantized sample by sample. This quantized residual is then added to the predicted speech signal resulting in the quantized speech signal. The linear predictor, represented as a transfer function ($P(z)$), uses this quantized signal as its input and produces a new predicted speech signal. This predicted speech signal is then, again, compared to the input speech and produces a "new" predicted residual. In the process of quantization the quantizer produces a signal codebook which is transmitted to the decoder as a compressed bit stream. During decoding this codebook is applied and the quantized residual is added back to the predicted speech signal and output as quantized speech. This feedback loop can be regarded as a synthesis filter with a transfer function of $1/[1 - P(z)]$. In linear predictive coding, the major task of the encoder is to identify the model parameters of this synthesis filter. In case of AbS the residual signal is not quantized sample by sample but block by block with blocks of sample size K and bit rate r. This results in $2^{K \cdot r}$ residual candidates, also called excitation candidates, which are each passed through the synthesis filter and result in a synthesized speech signal, which is then compared to the input speech and produces a new excitation signal (former predicted residual). In this case, the task of the encoder is to choose the excitation candidate that minimizes the error between the synthesized and input speech. The model parameters of the synthesis filter are derived directly from the input speech signal itself. The most prominent audio coding technologies to date, based on AbS, are Algebraic CELP (ACELP) [Laflamme et al. 1990], Embedded ACELP [ITU-T 2009], Forward Backward linear predictive coding (FB-LPC) [Andersen et al. 2002; Andersen et al. 2004] and Two-Stage Noise Feedback Coding (TSNFC) [Chen 2006]. Well-known audio codecs based on AbS are Adaptive Multi-Rate (AMR), AMR Wideband (AMR-WB) and Speex (SPX) (cf. Appendix B).

So called hybrid codecs are a mixture of both AbS and perceptual coding and can be used more universally for both speech communication and music compression. Most codecs switch between the two coding modes frame by frame in case of speech presence or absence in the current audio frame (e.g. AMR-WB+). Others utilize both coding modes at the same time, for example, by running a filter bank-based audio coding on top of the underlying AbS coder (e.g. MPEG-4 Scalable Audio Coding, not further considered in this Thesis), or by coding high and low frequency bins utilizing different coding modes (e.g. OPUS, cf. Appendix B).

Speech Emotion Perception/ Recognition from Compressed Speech

Until now, the focus of both application domains (i.e. reduction of storage size and transmission latency) is the reduction of data size while maintaining a good listening quality [Salomon 2007]. This, however, is not enough when it comes to speech emotion understanding and recognition, as the decrease in data size is mostly accompanied with an information loss. With regard to speech quality and listening quality, this information loss can be neglected from certain compression bit rates upwards, depending on the type of audio codec. From the signal point of view, this information loss corresponds to differences in the waveform itself and differences in the spectral power of the signals (cf. Figure 4.1 on page 111, depicting an exemplary power spectrum of an uncompressed and compressed speech sample and the resulting error). These differences in the speech signals are barely investigated, as the scope of audio compression desires mainly a good listening quality. Especially for emotion understanding, however, this information loss is of high importance, as, without regarding textual features, the emotional content is taken mainly from paralinguistic cues (cf. Section 2.2.2). In [Labelle et al. 2016] and [Lahaie et al. 2017], the authors state that the ability to perceive the emotional content from speech decreases when an audio codec or bandwidth limitation is applied to the speech signal. As for automatic speech emotion recognition the relevant speech features are automatically extracted directly from the speech signal in time and/ or frequency domain, it is assumable that a similar decrease in the recognition performance will also occur. One of the first studies where the effect of audio compression on speech based emotion recognition is investigated is presented in [García et al. 2015]. Here, the authors investigate the impact on the accuracy of detecting fear-type emotions of the EmoDB (cf. [Burkhardt et al. 2005]) and eNTERFACE'05 (cf. [Martin et al. 2006]) data sets by applying several speech codecs and GMMs as classifiers. As evaluation scheme a LOSGO cross-validation is performed utilizing the corresponding clean and compressed speech samples for training and testing (matching conditions). Depending on the considered data set and whether voiced or unvoiced speech segments are employed, some audio codecs lead to an increase of accuracy. In case of voiced speech segments, an increase occurs when applying the AMR-WB codec with low bit rates and the OPUS codec with an average bit rate of 64 kbit/s on the EmoDB data samples. For the eNTERFACE'05 data samples, this is only the case when the OPUS codec is applied. When utilizing the unvoiced speech segments only, the EmoDB data samples never show an increase in accuracy when applying an audio codec. In case of the eNTERFACE'05 data samples, a clear increase in accuracy is observed with the AMR-WB coded samples at all bit rates. In [Albahri & Lech 2016] and [Albahri et al. 2016] the authors investigate the effect of four audio codecs (AMR, AMR-WB, AMR-WB+ and MP3) on the accuracy of a speech emotion recognition of the EmoDB data set when utilizing GMMs, considering three different acoustic feature types (MFCCs, TEO features, and glottal time and frequency domain features). As evaluation scheme a 15-fold cross-validation is performed utilizing 80% of samples for training and 20% for testing of the classifier (matching condition). When applying MFCCs or glottal frequency domain features, the accuracy obtained on the uncompressed speech samples always outperforms the results obtained on the compressed samples. In case of the TEO features, this is not the case when utilizing the AMR-WB codec with a bit rate of 6.6 kbit/s. When applying glottal time features it further needs to be distinguished between female and male speaker. For female speakers, the best results are always achieved when using the original speech samples. In case of male speakers, the AMR-WB coded samples when applying various bit rates and the MP3 coded samples with 8 kbit/s show an increase in accuracy compared to results obtained from the original uncompressed samples. The, to my knowledge, newest investigation on speech emotion recognition from compressed speech is presented in [Oates et al. 2019]. Here the authors, present a broad range of audio codecs and utilized bit rates for three different speech emotion data sets (EmoDB [Burkhardt et al. 2005], Polish-EMO [Staroniewicz & Majewski 2009] and eNTERFACE'05 [Martin et al. 2006]). As classifier a simple SVM with linear kernel is chosen and four feature sets, designated for speech emotion recognition, are utilized (*IS'09 Emotion*, *ComParE'13*, *emo large* and *eGeMAPS*, cf. Table 2.2). As validation scheme they utilize a LOSO cross-validation with matching and mismatching training and test conditions. Depending on the utilized feature and data set the results of the emotion recognition experiments differ significantly. Therefore, a clear statement on the influence of the feature set and applied audio codec on the recognition performance cannot be made. By specifying the performance of the classifier as UAR, these results are comparable to the results, which will be presented in Section 4.2 of this Thesis.

2.6 Speech Emotion Recognition in Vehicles

We have now come to a point of this Chapter where I have introduced all building blocks of speech emotion recognition, speech quality and disturbed speech to the reader. Based on this foundation, I will now focus on a more specific introduction to the research field of speech emotion recognition in in-vehicle environments. The research results presented in this Section serve as basis for the study design of the data collections in Chapter 3, the quality assessment and first recognition experiments on simulated in-vehicle emotional speech data in Section 4.3, and in-depth recognition experiments on real-world in-vehicle speech data in Chapter 6.

2.6.1 Simulated Driving Environment

Most of the work on in-vehicle emotion recognition presented in literature focus on the evaluation of speech emotion recognition systems in a simulated driving environment. This includes recordings obtained inside a real car body, but also more simplified simulation environments, where the simulator consists of a car seat and a steering wheel. Others do not use a simulator environment at all and add in-vehicle noises to benchmark data of emotional speech. As mentioned in Section 2.5.1 and defined in Equation (2.20) on page 60, in real-world scenarios the speech signal and noise signal are both convoluted with the impulse response defined by the present room acoustics. This implies that by simply adding in-vehicle noises to a speech signal, the in-vehicle acoustics are left completely unconsidered and only the original recording conditions effect the signal, meaning that the utilized speech and noise signals are convoluted with the impulse response of the original recording setup. Based on Equation (2.20), this would result in the following equation for the obtained observed signal:

$$y(n) = x(n) * h_1(n) + d(n) * h_2(n), \tag{2.22}$$

with $h_1(n)$ representing the impulse response of the original clean speech recording setup and $h_2(n)$ representing the impulse response of the original in-vehicle noise recording setup. For other investigations, where real in-vehicle noises are replayed inside a simulation environment, it has to be distinguished between a real person talking inside the simulator and a data set being replayed. In case of a real person talking inside the simulator, $h_1(n)$ represents solely the impulse response of the simulator and $x(n)$ the raw speech signal. For the second case, where recorded speech samples are replayed, $x(n)$ also contains the speech signal, but this signal is already convoluted with the impulse response of the original recording setup. The same holds for $d(n)$, when replayed inside a simulation environment. These assumptions are quite different compared to real-world driving scenarios, where $h_1(n) = h_2(n)$ and $h_1(n)$ represents the impulse response of the in-vehicle environment, and $x(n)$ and $d(n)$ represent the raw excitation signals of the speech and noise source. Investigations evaluating the effect of the in-room acoustics of university lecture rooms have shown that this effect should not be neglected with regard to speech emotion recognition ([Höbel-Müller et al. 2019]). Furthermore, a replay of benchmark emotional speech data sets does not mirror the communication inside a real-world driving situation and does not take into account the Lombard-effect, occurring in natural human communication inside noisy environments.

Emotion Recognition in a Simulated Driving Environment

I will now give an overview on hitherto published work on speech emotion recognition in a simulated driving environment. It can be distinguished between four types of simulation: 1) Additive in-vehicle noises in different SNRs on benchmark emotional speech data sets, 2) Additive in-vehicle noises with natural SNR distribution on benchmark emotional speech data sets, 3) Additive in-vehicle noises in different SNRs on real in-vehicle audio recordings of acted emotional speech, and 4) Replayed in-vehicle noises with natural SNR distribution on real in-vehicle communication data. All these simulation types differ in the definition of the observed signal in Equation (2.22) and will now be described in more detail.

1) Additive in-vehicle noises in different SNRs on benchmark emotional speech data sets

For this case $h_1(n)$ and $h_2(n)$ correspond to the impulse response of the original recordings' setup of the emotional speech data set and noise data recording setup, respectively. This approach is used in [Chenchah & Lachiri 2016]. Here, the authors utilize the IEMOCAP data set (cf. [Busso et al. 2008]), containing English emotional speech in four emotion categories (anger, happy, neutral and sad) with four artificially added noise types (car, babble, train, and airport noises) in four different SNRs (0 dB, 5 dB, 10 dB and 15 dB). Furthermore, the influence of three different speech enhancement methods is evaluated. By utilizing HMMs with MFCCs as prosodic features and a cross-validation scheme (training on clean speech, testing on noisy/ enhanced speech), a recognition accuracy ranging from 57.71% to 58.66% for noisy speech and 56.13% to 59.73% for enhanced speech is reached for the car noise condition. A more natural data set is utilized in [Weninger et al. 2011], where babble and street noise of the Aurora noise database (cf. [Pearce & Hirsch 2000]) is added to the close-talk microphone recordings of the German FAU Aibo Emotion Speech corpus (cf. [Batliner et al. 2004]). The different noise recordings are artificially added in four different SNRs (-5 dB, 0 dB, 5 dB and 10 dB). Afterwards, classification experiments are conducted using a cross-validation scheme. The validation schemes, relevant for in-vehicle emotion recognition, are based on a training of the classifier on clean, babble noise speech, street noise speech and a multicondition (clean + babble + street) training and testing on the street noise speech samples. As classifier a SVM with linear kernel and hyper-parameter optimization is utilized to classify the two emotion categories, *negative valence* and *neutral*. This results in UARs ranging from at most 60.60% to up to 67.20% for the different training conditions and feature sets. Furthermore, a strong deviation of the recognition performance is noticed for the different SNRs. In case of a two class classification problem, where the chance level lies at 50%, these results are rather unspecific. However, in [Weninger et al. 2011], the utilized street noises correspond to observations made from the outside of the vehicle on the traffic occurring on the street. This is only partly related to real noises observed inside a vehicle while driving. A different study, utilizing self-recorded noise samples, is presented in [Tawari & Trivedi 2010b]. Here, the noise recordings are obtained inside a real vehicle while driving. The recordings are obtained under highway, parking lot and city street driving conditions and are artificially added in three different SNRs (5 dB, 10 dB and 15 dB). The EmoDB data set, containing German acted speech samples in seven emotion categories (anger, boredom, disgust, fear, happiness, sadness and neutrality) (cf. Section 2.1.5) is utilized and a SVM with linear kernel and a 10-fold cross-validation (training on clean speech, testing on noisy/ enhanced speech) is applied. As features 1054 acoustic features are used and a 10-fold selection procedure is applied to select the most relevant features out of the set. The resulting recognition accuracy ranges from 16.4% to 37.0% under noisy test conditions and 37.5% to 63.0% under speech enhanced test conditions.

2) Additive in-vehicle noises with natural SNR distribution on benchmark emotional speech data sets

As for case 1), $h_1(n)$ and $h_2(n)$ correspond to the impulse response of the original recordings' setup of the emotional speech data set and noise data recording setup. Contrarily to case 1), here, the real SNR distribution of the noise signal is employed, without artificially manipulating the SNR. Before adding the noise signal to the emotional speech samples, it is normalized to fit the loudness of the clean speech signal. First investigations are presented in [Schuller; Rigoll; Grimm et al. 2007]. In this publication, self-recorded noise samples are added to the EmoDB data set. The noise samples are recorded inside a simulator, with a microphone mounted in the middle of the instrument panel, where the microphone of the in-vehicle communication system is commonly installed. The in-vehicle acoustics of four different car types (BMW 530i (Touring), 645Ci (Convertible), M5 (Limousine) and Mini Cooper (Convertible)) and three road types (big cobbles at 30 km/h, smooth city road at 50 km/h and Highway noise at 120 km/h) are emulated by the simulator. As classifier a SVM with linear kernel and 1406 acoustic features is utilized. Further, the acoustic features of each speaker are normalized and a LOSO cross-validation is conducted. Two different validation strategies are presented: training on clean speech and testing on noisy speech and training and testing on noisy speech. This results in error rates ranging from over 25% (ACC = 75%), for strong noise disturbances and clean training condition, to below 19% (ACC = 81%), for weak noise disturbances and noisy training conditions. In [Schuller 2008] the presented classification experiments are applied to the more ”natural“ speech samples of the eNTERFACE’05 data set (cf. [Martin et al. 2006]). This data set comprises scripted English speech samples in six emotion categories (anger, disgust, fear, happiness, sadness and surprise). Additionally to the experiments presented in [Schuller; Rigoll; Grimm et al. 2007] a feature selection is performed for both data sets, EmoDB and eNTERFACE’05. This results in an increase of accuracy for the EmoDB data set up to 83.0%. For the eNTERFACE’05 data set the recognition accuracy ranges from below 49%, for strong noise disturbances and clean training condition, to over 63%, for weak noise disturbances and noisy training conditions. By conducting the feature selection it is even increased to 65%. Similar classification experiments, utilizing the same noise samples but a different emotional speech data set, are presented in [Grimm; Kroschel; Schuller et al. 2007] and [Grimm; Kroschel; Harris et al. 2007]. Here, the VAM data set is used, comprising emotional speech in the three dimensions of valence, activation and dominance (cf. Section 2.1.5). By utilizing SVR with a Radial Basis Function (RBF) kernel and 20 prosodic features, which were selected by sequential forwards selection, correlation coefficients ranging from 0.10 to 0.45 for valence, 0.40 to 0.81 for activation and 0.40 to 0.79 for dominance are obtained. The lower values are obtained under strong noise disturbances and clean training condition and the higher values under weak noise disturbances and noisy training conditions, respectively. The comparatively low recognition measures obtained under clean training conditions are, however, expectable, as the clean speech data differs significantly from the noisy data the classifier is tested on. In real-world applications it can be assumed that a clean version of a speech signal is not available. Therefore, the results obtained from training and testing on the noisy speech signal are of higher relevance for the investigations presented in this Thesis.

3) Additive in-vehicle noises in different SNRs on real in-vehicle audio recordings of acted emotional speech

For this case $h_1(n)$ corresponds to the impulse response of the real in-vehicle recording setup, while $h_2(n)$ corresponds to the impulse response of the noise data recording setup. This approach is presented in [Tawari & Trivedi 2010a] and [Tawari & Trivedi 2010b], respectively. The authors utilize a self-recorded data collection of English in-vehicle acted emotional speech, the LISA-AVDB (cf. Section 2.6.3) in three emotion categories (positive, neutral and negative). The speech data is recorded in both stationary and moving car environments. The authors, however, only utilize the recordings obtained in the stationary mode and artificially add noise recordings to these speech samples with three different SNRs. The noise recordings correspond to the ones utilized in [Tawari & Trivedi 2010b] and presented in case 1). Similar classification experiments as presented in [Tawari & Trivedi 2010b] are performed (SVM with linear kernel and a 10-fold cross-validation). The recognition results range from 53.6% to 81.7% under noisy test conditions and 62.5% to 82.1% under speech enhanced test conditions, for the three-class classification problem.

4) Replayed in-vehicle noises with natural SNR distribution on real in-vehicle communication data

This case is the most natural way to simulate the in-vehicle environment, as $h_1(n)$ and $h_2(n)$ correspond to the impulse response of the in-simulator recording setup. The term $d(n)$, however, already includes the convolution of the raw noise signal with the impulse response of the noise recording setup. Depending on the type of simulator and how close it is to a real driving experience, more or less natural in-vehicle driving situations can be simulated. A first investigation on simulated in-vehicle communication scenarios is presented in [Jones & Jonsson 2005]. Here, the authors present a data collection of real in-vehicle Human-Computer Interaction (HCI), where a system initiates a conversation between the car and the driver. The simulator, however, only consists of a car seat, steering wheel, accelerator and brake pedals inside a not further specified room. From the pictures presented in [Jones & Jonsson 2005], it can be assumed that the room does not comply with any in-vehicle circumstances. An initial ANN is trained on English clean speech data in five emotion categories (boredom, sadness/ grief, frustration/ extreme anger, happiness and surprise), utilizing the deltas of 10 acoustic features. This classifier is then applied to the simulator recordings and achieves a recognition accuracy of 60% to 70%. In [Jones & Jonsson 2007] the authors further distinguish between different gender groups and could thereby increase the recognition accuracy to 64% for female and 67% for male drivers. In [Jones & Jonsson 2007] a special focus is drawn on older drivers, for which a recognition accuracy of approx. 70% is achieved. A more recent publication by Cevher et al. is focusing on real in-vehicle communication scenarios between the driver and co-driver [Cevher et al. 2019]. The presented driving scenarios used to induce emotions are based on conversational driven and task oriented emotion induction. This approach is closely related to the approach utilized in this Thesis and presented in Section 3.2 (cf. [Lotz; Ihme et al. 2018; Requardt et al. 2018]). It can be assumed that the conversational driver emotion inducement method is the most natural compared to real in-vehicle Human-Human Interaction (HHI). Unfortunately, the authors do not give any detailed information on the classification approach and utilized feature set. A focus is drawn on the detection of insecure, annoyed and happy drivers. For this three-class classification problem a macro averaged F1-score of 29% is achieved. This corresponds to a recognition below chance level (∼33%).

2.6.2 Real-World driving Environment

In contrast to Section 2.6.1, where speech emotion recognition in a simulated driving environment is evaluated, I will now focus on the current state of the art in speech emotion recognition inside a real-world driving environment. A comparable real-world data collection will be presented in Section 3.2. For this real-world scenario, the observed signal inside the running vehicle is calculated as stated in Equation (2.20), with $h(n)$ representing the impulse response of the real in-vehicle recording setup, and $x(n)$ and $d(n)$ representing the excitation signal of the speech and the noise source. Unfortunately, the work done on real-word in-vehicle speech data is strongly limited. One of the very few publications taken into account this natural recording setup is [Abdić et al. 2016]. Here, the authors focus on the recognition of frustrated drivers. The utilized data is recorded inside a 2013 Chevrolet Equinox and a 2013 Volvo CX60 while driving on a highway (cf. [Mehler et al. 2015] and Section 2.6.3). It contains only recordings of HCI and no natural interaction between humans. To detect the driver frustration from speech, a SVM with a linear kernel is utilized. As feature set the GeMAPS feature set is used. The classification problem consists of two classes: high and low frustration, with the ground truth extracted from the driver's self-report and not obtained through an expert labeling. This approach results in a recognition accuracy of 77.4%. A second investigation based on a real in-vehicle driving environment is presented in [Bořil et al. 2010]. The authors utilize the *UTDrive* data set, as presented in Section 2.6.3. By employing a GMM-based maximum likelihood classifier and utilizing a speaker/ gender-independent split of the training and test set, they achieve equal accuracy rates ranging from 66.4% to 69.3% for the two-class classification problem (neutral vs. negative).

2.6.3 Available Modalities and Data Sets

As described in Section 2.6.1 and Section 2.6.2, most investigations in the field of in-vehicle speech emotion recognition are based on well-established benchmark emotional speech data sets with additive in-vehicle noises. The reason for this circumstance is that the number and amount of real-world in-vehicle speech data is still limited. An open source data set on natural in-vehicle emotional data is, to my knowledge, not yet available. All presented publications not based on benchmark data sets, utilize their own designated audio recordings, which in most cases are not publicly available. The realization of these designated data collections is highly resource and time consuming and cannot be established without a certain financial and technical support. An Overview on available speech emotion data sets utilizing an in-vehicle environment is given in Table 2.6. In the following, a more detailed description is given for each set separately.

One early project considering in-vehicle emotions is the *Emotive Driver Project* presented in [Jones & Jonsson 2005]. Here, emotional speech data is collected inside

Table 2.6: Overview on in-vehicle emotional speech data sets. Empty entries were not made available in the referred publication.

Name	Samples [#]	Cat.	Dim.	Naturalness	Rec. Environment	Interaction
Emotive Driver project [Jones & Jonsson 2005]	-	6	-	natural	simulated	HCI
LISA-AVDB [Tawari & Trivedi 2010a]	224	3	-	acted/ natural	real world	HHI
AMMAR [Cevher et al. 2019]	288	6	3	natural	simulated	HCI/ HHI
TUDrive [Angkititrakul et al. 2007; Bořil et al. 2010]	-	2	-	natural	real world	HCI
[Malta et al. 2011]	-	2	-	natural	real world	HCI
[Mehler et al. 2015; Abdić et al. 2016]	596	2	-	natural	real world	HCI

a simulator consisting of a screen, car seat, steering wheel with haptic feedback and a brake and acceleration pedal. The emotions are elicited by communicating with an in-car information system, which informs the driver on topics related to the road conditions, traffic and driving conditions, and engages the driver into a conversation based on interview-like questions. It is not specified what kind of driving scenarios are employed in the simulator and if it only contains straight driving or also other disturbances. No specific emotions are induced and the data is afterwards labeled by trained experts in the categories boredom, sadness, anger, happiness, surprise and neutral. The disadvantages of this approach are evident. First, the utilized simulator is not similar to a real vehicle in various aspects (e.g. acoustics, driving behaviour, etc.). Second, the emotions are not induced. For driving without further disturbances it is rather unlikely to observe an emotion different from neutral or boredom. This is also confirmed by the labeling results which mostly correspond to a neural and bored state of the driver. The full data set comprises data from 41 English speaking participants (21 females) in the same age group (18 to 25 years).

In [Tawari & Trivedi 2010a] a more realistic driving environment is utilized and a multimodal collection of audio and video data, called the *Audio-Visual Affect Database (LISA-AVDB)*, is presented. This data set is collected inside a real operating vehicle as well as inside a stationary vehicle. The vehicle type and other environmental conditions, as street type, vehicle average speed or road surface are not introduced to the reader. To elicit a certain emotion in the driver, two different approaches are utilized. First, the driver is prompted by a computer system to express a specific emotion, giving example sentences on how to express the said emotion. Second, free conversations between the driver and a passenger are recorded. Here, no emotion is induced. The emotional speech data is then labeled into the categories positive, neutral and negative expressions. In total 224 speech samples (82 positive, 82 negative and 60 neutral) from four English speaking participants (two females) are obtained. No further information on the segmentation of the speech signals for the annotation is given.

A more recent collection of in-vehicle emotional data is presented in [Cevher et al. 2019]. The, so called, *AMMAR* data set comprises multimodal data of audio, video and bio-physiological recordings. Furthermore, the audio signal is transcribed to also obtain the textual information. The data collection is performed inside a fixed-base driving simulator consisting of a car cabin (only frontal area of the car body) and a wide screen. As driving environment, ”everyday driving situations“ on highway, rural roads and city streets are utilized and a focus is drawn on the interaction between the driver and a virtual agent as well as a co-driver. To withhold the actual goal of the driving task (i.e. collection of emotional data) from the driver, a cover story is used. The drivers are told to evaluate and improve a given driving assistant system. Furthermore, to reinforce the emotional state, the drivers are told to reach the desired destination as fast as possible, observing the traffic rules and speed limits. The emotions themselves are induced by several events occurring during the driving task. First, the car is cut off by another car and blocked by trucks on both lanes. Second, a skateboarder appears unexpectedly on the street, and finally, the driver is pressured to reach the desired destination first, compared to other participants. The situations are reinforced by the virtual agent, asking situation related questions. Further, the drivers are involved in conversations with the co-driver on rather positive experiences (i.e. last vacation, dream house and perfect job). As ground truth an emotion self-rating of the participants is used, in which the participants listen to their own audio recordings and label every utterance in emotion categories (annoyance, insecurity, joy, relaxation, boredom and no emotion) and dimensions (valence, arousal and dominance). In total 288 speech utterances (90 joy, 26 annoyance, 49 insecurity, 9 boredom, 111 relaxation and 2 no emotion) from 36 German speaking participants in between the age of 18 to 64 years are obtained. It can be stated, that the general study design is closely related to the approach presented in [Lotz; Ihme et al. 2018] and Section 3.2. Nevertheless, it can be assumed, that the secondary task of reaching the desired destination first, will most likely have a negative effect on positive emotions. And strong negative emotions may be reinforced. Furthermore, positive emotions are only induced by highly natural conversation with the co-driver and not intensified by the driving scenario. This may also explain the low recognition rates obtained on this data, which were reported in the previous Section.

While these three data sets comprise speech in multiple emotional states, the following publications focus on the inducement on negative emotions/ frustration of the driver only. In [Bořil et al. 2010] the *dialogue systems scenario* of the *UTDrive* data set [Angkititrakul et al. 2007] is annotated in two categories (neutral and negative). This data base contains audio and video recordings inside a real vehicle in a residential area and business district. During the dialogue system scenario the driver needs to call an airline's flight connection system to get information on the arrival/ departure gates of a particular flight, and a voice portal to obtain information depending on the driver's personal interests. An inducement of emotions is not per se a goal of the data collection. The dialogue system scenario, however, induces negative emotions in the driver, because of the high number of speech recognition errors occurring while communicating with the automated systems. In total 68 English speaking drivers (33 females) of the UTDrive data base are annotated by one expert annotator.

A second data set on frustration only is presented in [Malta et al. 2011]. The authors present a multimodal data collection on audio, video and bio-physiological recordings obtained in a real vehicle under city street environmental conditions. To induce frustration, the experimental route is chosen in a way that the number of frustrating environmental factors (e.g. high traffic density) is increased. Furthermore, the drivers need to perform a secondary task where they have to retrieve and play as many songs as possible, within a certain time limitation using an automatic speech recognition system. This system is highly prone towards speech recognition errors and regularly misunderstands the driver's commands. Except for the secondary task, the drivers are not encouraged to speak during the experiment. An annotation of the data is performed by utilizing the video recordings only and is carried out for the two categories of neutral and non-neutral. The true label of frustration is taken from the participants' self-report in the two categories frustrated and non-frustrated. The results are stored in a continuous stream of binary information. In total 30 Japanese speaking participants (10 females) in between the age of 20 to 58 years are included in the data set. One major concern on the experimental setup is that the secondary task will not only increase the frustration level of the driver but may also lead to cognitive overload due to the time limitation given for the task. This, however, is not further taken into consideration by the authors.

In [Abdić et al. 2016], the authors utilize a subset of the data collection presented in [Mehler et al. 2015]. This data collection is originally designed to evaluate the effect of voice interfaces in embedded vehicle systems on the driver's visual and manual distraction. It comprises multimodal data of audio, video and bio-physiological recordings. The effect of voice interfaces is evaluated by performing three different secondary tasks per participant inside a real vehicle under highway environmental driving conditions. These tasks include entering an address into a navigation system via voice control, manually phone calling, and calling a person via voice control. After each task, the participants self-report their work load (no scale provided). This also includes an assessment of the frustration level on a scale from 1 *not frustrated* to 10 *very frustrated* after each driving task. All tasks with ratings of 4 to 6 are assumed to indicate a *neutral frustration state* and are excluded from the data set. Everything above 7 is assumed to indicate frustration and everything below 3 satisfaction of the driver. Out of the 80 initially collected English speaking participants, only 20 participants are used in the frustration data set, equally distributed in age and gender. This subset of the original data set comprises 596 audio samples of different duration with a majority of samples labeled as frustrated.

As stated earlier, speech data is highly affected by the environment it is recorded in (cf. Section 2.5.1). Therefore, it is assumed that a classifier designed for a certain application domain should be also trained on data originating from this domain. It is a rather vague assumption that a classifier trained on a random set of emotional speech data will later also work in its designated application environment. It can be summarized that an open source benchmark data base on in-vehicle emotional speech data would be highly conductive for an unspecific evaluation of emotional speech in an in-vehicle environment, especially for research institutions with small budgets. Nevertheless, for real world application either a designated data set is needed or a much wider range of data comprising a variety of in-vehicle emotional speech data. By the second approach, a more general data set is generated representing the entity of in-vehicle emotional speech. This, however, seems rather challenging with regard to the low number of available data sets focusing on different languages, naturalness, interaction and recording environments. With a higher number of available in-vehicle data it would further be possible to adapt the features extracted in a non-driving environment. This, however, is not possible so far.

2.6.4 In-Vehicle Emotion Recognition from different Modalities

In Section 2.1.2, I presented the concept of emotions and appraisal theory. Considering the measures presented in this Section, there are different ways to recognize a change in a person's emotional state, which are more or less convenient when it comes to the detection of the driver's emotions. Until now, a focus was drawn on the detection from vocal expressions only. To complement this information, I will now give a brief overview on drivers' emotion recognition considering the measurement of the changes occurring in the five components of an emotion episode.

1. Cognitive component: A measurement of the neural activity of the driver is highly obtrusive. The attachment of electrodes on the scalp of the driver can effect the human-vehicle interaction and disable a natural driving behavior. Even though there has been improvement in dry and portable measurement systems in the past years (e.g. [Zander et al. 2011] and [Volkening et al. 2018]), this approach can be seen as rather inconvenient in the driving context.

2. Neurophysiological component: There exist different methods to measure the drivers' physiological parameters. These range from more obtrusive methods, as utilizing body attached electrocardiogram (ECG) measurement sys-

tems (e.g. ECG-belts, finger sensors or ECG-electrodes attached to the chest of the driver), to less obtrusive contactless systems, such as the ECG-steering wheels (cf. [Gomez-Clapers & Casanella 2012; Lourenço et al. 2015]) or the smart driver seats (cf. [Vetter et al. 2017]).

3. Motivational component: In contrast to everyday behavioral changes, which are rather challenging to assess [Harrigan et al. 2005], the changes occurring in the driver's behavior are comparatively well measurable. They can be assessed through changes occurring in the interaction of the driver with the vehicle, either by directly measuring the vehicle speed, acceleration or time headway, or from input devices in the vehicle, such as the brake, the throttle, or the steering wheel. However, especially in the driving context, these changes are not used by humans to communicate there emotional state to the environment, but can be seen as the consequence of an emotional incident [Scherer 2005b].

4. Motor expressive component: The facial expression, body movement and vocal expression of the driver, can be easily assessed by integrating cameras and microphones into the vehicle cabin. This can be done in a non-disruptive way, not leading to disturbances in the human-vehicle interaction. While many vehicles are equipped with hands-free speaking systems and a microphone array being already integrated inside the vehicle-cabin, the availability of a camera system inside an ordinary vehicle is yet uncommon.

5. Subjective feeling: The assessment of the subjective feeling while driving is highly inconvenient, as the driver needs to provide explicit information on their internal state. It can be assumed, that in a natural driving environment it is very unlikely that the driver will provide this information without being prompted to do so. This is a common way used in psychological investigations, but is highly unsuitable in a driving context.

With regard to the above stated assessment methods, some research is already available in the field of automatic in-vehicle emotion recognition. While using behavioral changes to detect the emotional state of the driver is still under-researched (e.g. in [Shafaei et al. 2019]), there is more work available investigating automatic emotion recognition based on bio-physiological measures and video signals (cf. [Nasoz et al. 2004; Katsis et al. 2008; Malta et al. 2011; Gao et al. 2014; Verma & Choudhary 2018b; Verma & Choudhary 2018a; Ihme; Unni et al. 2018] and [Zepf et al. 2019]).

2.7 Summary of this Chapter

This Chapter of the Thesis provided the reader with the necessary information needed for the investigation of the three main research hypotheses, as introduced in Section 1.4, and understand the research content presented in the following Chapters. Furthermore, this Chapter provided information on how to generate emotional speech, model a speech emotion recognition system, in the wild emotion recognition, speech quality and disturbed speech.

The next Chapter will focus on the collection of own simulated and real-world in-vehicle emotional speech data.

Chapter 3

Realized Simulated and Real-world In-Car Data Collections

Contents

FROM literature it is known, that in machine learning the recognition performance of a classifier is strongly dependent on the quality of the data it was trained on, as discussed in Section 2.2.6. Not only a high standard of the technical equipment (e.g. microphones, audio interface) is needed, as also the recording characteristics themselves can strongly affect the recognizer's performance. This includes the level of naturalness of the data (acted vs. spontaneous) and the recording setup (e.g. room acoustic characteristics, noise conditions, etc., cf. Section 2.5.1). For emotion recognition from speech, recognition rates can drop considerably from over 80% for acted emotions under clear recording conditions, to below 25% for naturalistic emotions under moderate recording conditions [Schuller; Vlasenko; Eyben et al. 2009] for comparable four class classification problems. Therefore, it is of high importance to operate with data, which is comparable to the data used during the later real-world application. If this condition is neglected, the obtained test results of the emotion recognizer cannot be taken as valid.

With regard to this information, it should be clear that a superimposition of in-vehicle noises to well-known emotional speech data sets is not enough to receive a trustworthy statement on the performance of a speech emotion recognizer under realistic disturbed recording conditions, as it is the case for most of the investigations presented in Section 2.6. An accessible real-world data collection, comprising spontaneous affected speech in in-vehicle environments, is, to my knowledge, yet unknown (cf. Section 2.6.3). Furthermore, the in-vehicle acoustics are influenced by many factors. These are, for example, the cabin size, the cabin interior or its material, which have an impact on the impulse response of the room (cf. Section 2.5.1). These differences can already have a considerable effect on the performance of the speech emotion recognizer. This implies, that not only a similar recordings setup inside a real vehicle with realistic in-vehicle noises is needed, but also the vehicle itself is important. While for different vehicle types like the passenger car or motor truck, the difference in the acoustic characteristics is striking (cf. cabin size and interior), for different passenger car models (e.g. limousine and station wagon) this difference is less obvious, but can still lead to noticeable differences in the quality of the audio recordings (cf. motion dampening and sound absorption).

To obtain realistic in-vehicle emotional speech data, with regard to the above mentioned requirements, two data collections were realized in the scope of this Thesis. As reliable data collections of naturalistic emotional speech are highly resource consuming, especially when it comes to the development of a reproducible test-scheme and the generation of a ground truth, the first data collection comprises re-recorded benchmark emotional speech data inside a fixed-based driving simulator. A detailed description of these simulated in-car recordings is presented in Section 3.1 and is based on [Lotz; Faller et al. 2018]. The data was later used for the investigations presented in Section 4.3 to get a first insight on the influence of speech quality in speech emotion recognition, and in Section 5.2 to describe the effect of speech enhancement on the recognition performance. Further advantages of this setup are the relative ease of recording even extensive speech material, and the comparability to results in other environments, as the originating data is a well researched benchmark data set.

The results obtained from the simulator recordings, and presented in Chapters 4 and 5, however, are not sufficient to describe the performance of a speech emotion recognizer in a real-world application to the full extent. From the simulated data, it is possible to analyze the effect of in-vehicle noises on the speech quality, the features used for emotion recognition and other speech processing artifacts (e.g. speech enhancement), and give first insights on the feasibility of speech emotion recognition in a driving environment. Nevertheless, the naturalness of the in-vehicle speech is neglected. Therefore, a second data collection was realized, comprising realistic highly natural emotional speech samples as they would occur in real-world driving situations inside a designated test vehicle. The details of this data collection are presented in Section 3.2 and are based on [Lotz; Ihme et al. 2018]. Furthermore, the usability of this data collection was evaluated and validated in Section 3.2.3 of this Chapter and is based on the results presented in [Requardt et al. 2018]. The recorded data set was later used in Section 5.1 to obtain a ground truth of said data, and in Chapter 6 to evaluate the ability to detect the driver's emotional state from in-vehicle speech.

As the approaches and results presented in this Chapter were already published in [Lotz; Faller et al. 2018; Lotz; Ihme et al. 2018] and [Requardt et al. 2018], several phrasings are taken literally from these publications.

3.1 Simulated In-Car Recordings

The first data collections were conducted by re-recording two well-known benchmark emotional speech data sets inside a fixed-based driving simulator consisting of a real car body inside a simulated environment as presented in [Lotz; Faller et al. 2018]. It was used to get an insight on the influence of real-world in-vehicle speech data on the speech quality and the ability to automatically recognize the drivers emotional state under non-ideal recording conditions. The databases comprised speech samples of different naturalness (acted and scripted emotions) and recording conditions (inside an anechoic chamber and a television studio) of the EmoDB and VAM data sets (cf. Section 2.1.5). An advantage of this approach is that the recognition results obtained from the original clean data samples serve as baseline for the evaluation of the non-ideal re-recordings. This makes it possible to discuss the effect of different signal processing steps, such as speech enhancement, and of noise conditions on the feature space and consequently on the recognition performance, as presented in Chapter 4 and Section 5.2. Furthermore, the conducted experiments, presented later, are also comparable to other state-of-the-art studies based on the evaluated data sets (cf. Section 2.6). This data collection was accomplished with the help of colleagues from the Continental Automotive GmbH. All evaluations on this data, presented in this Thesis, were done by myself.

3.1.1 The Simulator and Simulation Environment

The simulator is located inside a workshop shed at the premises of the Continental Automotive GmbH in Babenhausen, Germany. To dampen the noises coming from the workshop it is additionally placed inside a semi-anechoic chamber. The simulator itself consists of a BMW 5-series chassis, connected to the simulation environment and placed in front of a wide screen (due to copyright restrictions, please refer to [Lotz; Faller et al. 2018] for a picture of the simulator). The screen is used to simulate the driving environment and to give visual feedback to the driver. Additional acoustic feedback is generated to simulate environmental noises and engine sound. The environmental noises (e.g sound of tires on road surface, passing vehicles) are replayed by three speakers located in each front door and the rear window shelf of the vehicle. This placement of the speakers generates a listening experience of surround sound for the driver. To generate a realistic engine sound, an actuator is placed underneath the engine hood using it as a resonator. For technical reasons the dashboard of the cabin is replaced by a structure of strut profiles, simplifying the integration of various sensors into the car.

To enable playback and recording of the EmoDB and VAM data samples, similar to manual driving situations, the simulator was operating in automated mode. This allowed the placement of an active loudspeaker at the position where the driver's head would commonly be located while driving manually. The driving scenario of the simulator was designed as a two-lane highway driving task with a varying traffic density to generate diverse environmental noises during the experiment. Additionally, the driving behavior of the vehicle was manipulated manually from the simulator's control room. Changes in the vehicle's velocity and the traffic lane were initiated at certain points of time and logged onto a separate log file tracking the course of the simulation.

3.1.2 Microphone Integration

Two high resolution directional shotgun microphones (Sennheiser ME66) were integrated onto the strut profile, corresponding to a placement on the dashboard at both A-pillars of the chassis. In the following, the placement will be denoted as *left* and *right* microphone. The inlets of the microphones were directed towards the loudspeaker. An illustration of the hardware setup can be taken from Figure 3.1. To synchronize the audio channels of both microphones, an audio interface (Steinberg MR816CSX) was connected to the recording laptop using FireWire. The audio streams were recorded using the recording software Cubase[1]. The gain settings of the microphones and loudspeakers were set to a subjective volume comparable to passengers talking insider a running car. This setting was identical for all conducted recording setups.

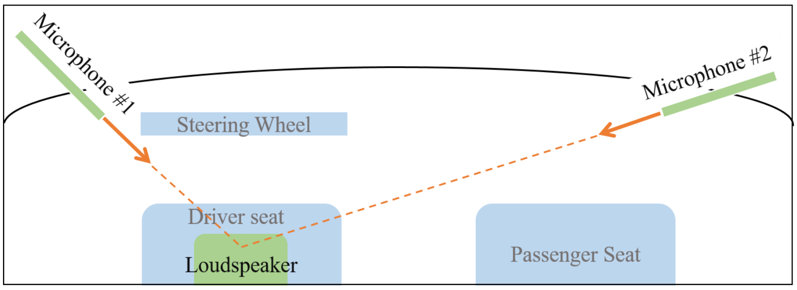


Figure 3.1: Schematic top view on the frontal area of the simulator's cabin.

[1] MIDI-sequencer and digital audio workstation developed by Steinberg Media Technologies.

3.1.3 Recording Setup

All samples of the EmoDB and VAM data set were re-recorded using two different recording setups. For the first recording, the simulator was turned off and only the in-car acoustics of the vehicle were influencing the recording. The recordings obtained under this setup are further referred to as *re-recording under silence condition*. For the second setup, the simulator was turned on and operating in the simulation environment, as described in Section 3.1.1. This led to a distortion of the obtained audio recordings by environmental noises and engine sound. The obtained recordings are further referred to as *re-recording under disturbed condition*. In the following chapters of this Thesis, the re-recorded data-samples under silence and disturbed conditions will be referred to as *re-recorded EmoDB under in-car recording conditions (EmoDB-Car)* and *re-recorded VAM under in-car recording conditions (VAM-Car)*.

3.2 Real-World In-Car Recordings

To also gain real-world data of emotions while driving, additionally to the simulator recordings, a second data collection with naturalistic in-car emotions was realized (cf. [Lotz; Ihme et al. 2018]). The goal of this data collection was to generate reliable, reproducible, multimodal and highly natural emotional in-car data, as they occur in everyday driving situations. Therefore, a strong focus was drawn on the comparability of the data with real-world driving situations. This was accomplished by a well-elaborated experimental setup and study design. To receive a more detailed insight on these important aspects, a detailed description of the data collection will be provided in Sections 3.2.1 and 3.2.2. To verify the usability and quality of the collected data, a validation of the data set, utilizing the driver's subjective self-reports and their peripheral physiological data, was conducted. The results are presented in Section 3.2.3 and are based on the work presented in [Requardt et al. 2018]. The recorded audio data will be used later, in Section 5.1, to generate the ground truth of the driver's emotional state, and in Chapter 6 to evaluate the ability to detect the driver's emotional state from in-vehicle speech.

The data was collected as part of my work for the project ADAS&ME[2]. While this Thesis focuses on the recognition of emotions from speech, other project partners focused on a recognition of emotions considering different modalities. Therefore, these modalities were also recorded during the data collection. The integration of the sensors, except for the microphone system, were not part of my work. The presented experimental setup and study designs, however, was a main contribution of myself in co-operation with the German Aerospace Center (DLR). In the later

[2] Funded by the European Union's Horizon 2020 research and innovation programme (grant agreement No 688900)

realization of the data collection, I only gave support regarding the audio recordings, whereas the evaluation of the recorded speech material was solely done by me.

3.2.1 Experimental Setup

Test Ground and Driving Environment

The real-world in-vehicle data collection was conducted at the compound of the DLR in Braunschweig, Germany, which is a designated test ground for driving experiments. The traffic density on the compound is comparable to a quiet residential area ensuring a realistic driving experience and driving environment. On the test ground, driving is allowed with a maximum speed of 30 km/h. For the experiment a fixed driving round course of around 900 meters was determined, which is depicted in Figure 3.2. One round on the course took approx. 2.5 minutes and was driven by each participant 20 times, interrupted by small pauses of approx. 5 minutes after every 5th round, used to collect the self-reported measures and as recreation of the driver. All recordings were conducted during daytime, to ensure optimal lighting conditions for the video recordings, and under similar and constant weather conditions, to ensure comparability among the recordings of the different participants. Because of the high sensitivity of the in-vehicle audio recordings towards environmental disturbances, the termination criteria of the data collection were strong rain and/ or thunderstorm. During the driving experiment, three persons were seated inside the car: The test subject/ driver on the driver seat, one investigator on the passenger seat, and one technician on the rear bench behind the driver. The investigator was leading the driver through the experimental course and the technician was responsible for the supervision of the sensor data recordings. The whole driving experiment took approx. 2 hours (1 hour of driving) per driver. This included the equipment of the participants and all further tasks like briefing, debriefing and answering provided questionnaires.

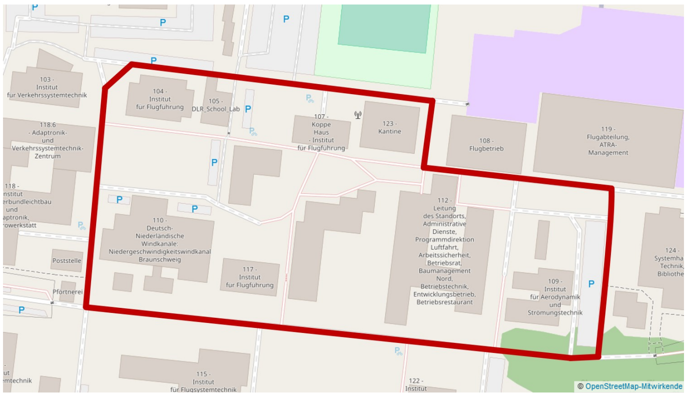

103 - Institut für Verkehrssystemtechnik
104 - Institut für Flugführung
105 - DLR_School_Lab
107 - Koppe Haus - Institut für Flugführung
123 - Kantine
108 - Flugbetrieb
119 - Flugabteilung, ATRA-Management
118.6 - Adaptronik- und Verkehrssystemtechnik-Zentrum
118 - Institut Verbundleichtbau und Adaptronik, Prowerkstatt
110 - Deutsch-Niederländische Windkanäle: Niedergeschwindigkeitswindkanal Braunschweig
117 - Institut für Flugführung
112 - Leitung des Standorts, Administrative Dienste, Programmdirektion Luftfahrt, Arbeitssicherheit, Betriebsrat, Baumanagement Nord, Betriebstechnik, Entwicklungsbetrieb, Betriebsrestaurant
109 - Institut für Aerodynamik und Strömungstechnik
124 - Systemha Technik, Bibliothe
Poststelle
Pförtnerei
115 - Institut für Flugsystemtechnik
122 - Institut
P
© OpenStreetMap-Mitwirkende

Figure 3.2: Round course at the DLR compound in Braunschweig, Germany (taken from [Lotz; Ihme et al. 2018], map taken from `https://www.openstreetmap.de/`)

Ethical Statement

The study procedure was reviewed and approved by the ethics committee of the Otto-von-Guericke University in Magdeburg, Germany (reference number 173/17). Furthermore, it was in accordance with the regulations and guidelines of the DLR. Before the start of the experiments, all participants had to provide a written informed consent to participate in the study.

Involved Participants

In total 30 drivers (seven females) were employed in the presented driving experiment. They were on average 30.5 years ± 5.0 years old and, hence, of the same age group (25 - 40 years). Furthermore, all participants were native standard German speakers without any speaking disorders, to guarantee that the differences occurring in their manner of speaking would not be influenced by their idiom or accent. Before the appointment of participants, they all had to answer a socio-demographic questionnaire and a general questionnaire regarding their driving experience. For safety reasons, only participants which were in possession of a valid driver's license and confirmed an annual mileage of at least 5000 km were taken into consideration for conducting the experiment. Furthermore, pregnant, physically impaired, heart and/ or neurologically disordered people were excluded from the study.

As the experiment took place on the DLR site in Braunschweig, all participants were employees of the DLR by the time of the experiment. To compensate their time effort, each participant received 30 € as reimbursement. At the beginning of the experiment, the participants were further asked to fill out the ATI-scale used to measure their attitude towards technology [Franke et al. 2017] and the Big Five Inventory (BFI-10) [Rammstedt & John 2007] to assess the big five OCEAN personality traits (i.e. Openness on experience (O), conscientiousness (C), extraversion (E), agreeableness (A) and neuroticism (N)). From literature it is known, that the personality can affect the ability to perceive/ utter certain emotions [Revelle & Scherer 2009]. This, however, was not further evaluated in the scope of this Thesis.

Test Vehicle

For the data collection the research vehicle FASCar II provided by the DLR was utilized [Fischer et al. 2014] (see Figure 3.3). The FASCar II is developed for testing driver assistance systems and automated driving. It is equipped with a unique steer-by-wire system to support innovative haptic feedback and intervention strategies. To meet the standard safety regulations an additional brake pedal is available at the co-drivers side. It is used for the safety driver to intervene in critical driving situations but can also be used to conduct Wizard-of-Oz (WoZ)-like driving experiments [DLR 2011; DLR 2019].

Figure 3.3: Research vehicle FASCar II (taken from `https://www.dlr.de/ts/desktopdefault.aspx/tabid-1236/1690_read-13097/`)

Sensor Integration

Three different sensor systems were integrated into the car to collect emotional data from three modalities: speech, facial expressions and physiology. The sensor systems were mounted onto the dashboard of the vehicle or were directly attached to the driver's body. Other hardware, necessary for the data collection, was mounted in the trunk of the car, not being visible to the driver.

A microphone system was used to record the speech of the driver. The system consisted of three microphones (two shotgun, one headset microphone) and an audio interface. The two highly directional shotgun microphones (Shure VP 82) were mounted on the dashboard of the car using elastic mounting to dampen the car's movement, one behind the steering wheel in front of the driver (not impairing her/his visual field), and one at the right A-pillar (cf. Figure 3.4). Both microphones were directed towards the mouth of the driver to suppress as much as possible surrounding noises. Additionally, a headset microphone (Sennheiser HSP-4 EW-3) was worn by the driver to collect high quality reference recordings without further disturbances of the driver's speech signal. These recordings were later used to obtain the ground truth of the data, by conducting a manual annotation (cf. Section 5.1). The different microphone tracks originating from the three microphones were recorded synchronously using an USB audio interface (Steinberg UR44) and the recording software Cubase[3] with a sampling rate of 44.1 kHz and a bit-depth of 16 bit.

The facial expressions of the driver were extracted from the video images captured by a Smart Eye Pro (SEP) Multi Camera System[4]. It consists of two high resolution infrared cameras with active infrared illumination. The system was attached to the dashboard on both sides of the steering wheel (cf. Figure 3.4). The recordings were obtained using a frame rate of 60 Hz, which is highly resource consuming (compu-

[3] MIDI-sequencer and digital audio workstation developed by Steinberg Media Technologies.

[4] Smart Eye AB, Gothenburg, Sweden, `www.smarteye.se`

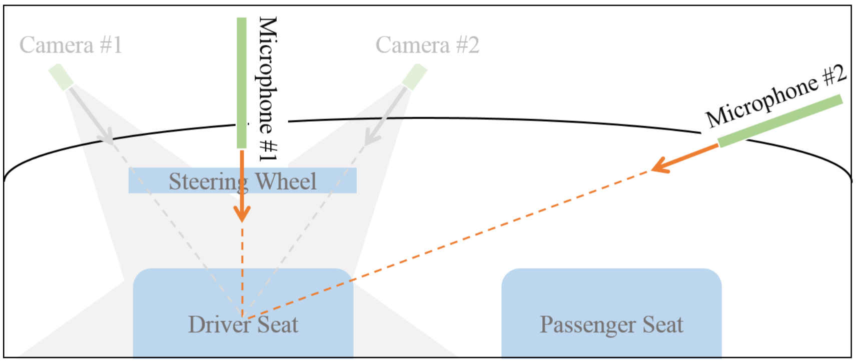


Figure 3.4: Schematic top view on the frontal area of the FASCar II test vehicle. Sensors with high relevance regarding this Thesis are denoted in bright color.

tation time and storage space). Therefore, an additional computer was mounted in the trunk of the car, dedicated to record the video images only.

The peripheral physiological data was recorded using the wireless sensor system Heally[5]. It consists of a standard 3-lead system and a finger sensor. The 3-lead system was used to measure the electrocardiogram (ECG) with a sampling rate of 500 Hz, the finger sensor measured the finger temperature and skin resistance at the index finger of the non-dominant hand of the participant with a sampling rate of 50 Hz. Both sensors communicate via Bluetooth to a computer stored in the trunk of the car.

All sensor systems were synchronized using a trigger signal coming from the SEP system. While the recordings of the peripheral physiological data were directly triggered by the SEP system, a manual synchronization was needed for the audio recordings. The SEP provided an isochronous impulse signal, which was fed to the audio interface as a separate input signal. This impulse corresponded to a flash occurring in the video images. The time stamps of these impulses and flashes were afterwards overlain to synchronize both signals.

3.2.2 Study Design

Target Emotional States

As already stated in the introduction of this Thesis, the most frequent emotions occurring while driving in a car are positive, anger and fear. Most of these emotions are highly expressive and do not occur in everyday driving situations but are triggered by challenging driving situations or defining experiences from the past (e.g. serious accidents)(cf. [Plutchik 1980]). For this data collection a special focus was drawn on milder versions of these emotions with less intensity , which are of frequent occurrence in everyday driving scenarios and can elicit stronger versions of

[5] SpaceBit, Eberswalde, Germany, `http://spacebit.de/html/body_heally.html`

said emotions. These states are denoted as *neutral*, *positive*, *frustration* and *anxiety*. In this Thesis, the positive state includes all relevant positive emotions occurring in the driving context like joy, amusement, contentment or happiness. From literature it is known that especially negative emotions can strongly influence the driving behavior in a negative way and lead to aggressive driving and distraction of the driver. Therefore, a more detailed focus is drawn on the detection of frustration and anxiety. A distinction between frustrated and angry driving as well as anxious and fearful driving is, however, a highly challenging task in speech emotion recognition, as frustration and anger show a significant correlation in their subjective emotion ratings (cf. [Liscombe et al. 2003]), and in [Schmidt-Daffy 2013] the author states that the symptoms of fear and anxiety are barely distinguishable from each other. Therefore, a distinction between these states will not be addressed in the scope of this Thesis.

More important, to design appropriate driving scenarios, a clear definition of the emotional states is necessary. By using the circumplex model of emotion concepts by Russell and Lemay, the target emotional states were mapped onto the dimensions of valence and arousal (cf. Figure 3.5) [Russell & Lemay 2000]. This model defines the neutral state as the region around the origin of the valence and arousal space with a moderate level of arousal and neutral valence. Furthermore, all expressions having a positive valence define the positive state. For the states of frustration and anxiety, no clear distinction regarding valence and arousal was possible, as both, anger and fear, are defined in a region of negative valence and high arousal. Therefore, additionally definitions, obtained from literature, were consulted, which defined frustration as the unpleasant feeling, which occurs in situations in which a person is detained from reaching a desired outcome/ goal and anxiety as the unpleasant feeling of dread over anticipated negative events [Lazarus 1991; Schmidt-Daffy 2013]. The corresponding driving experiments were designed such that the emotion elicitation would meet these definitions.

To sum up, the target emotions comprise *neutral*, *positive*, *frustration* and *anxiety*, which were mapped onto the valence-arousal-space as depicted in Figure 3.5. In the reminder of this Thesis both, the four emotion categories, as well as the dimensions of valence and arousal will be used to classify the drivers emotional state.

Driving Experiment

In the following, a detailed description of the different driving scenarios used to induce the four target emotions will be presented. The scenarios were designed such that the driving itself would only minimally influence the driver's emotional state. To hide the actual goal of the driving experiment from the participants, a cover story was provided. This cover story mainly focused on the evaluation of different *newly* developed driving assistant systems by the participant. For each driving scenario five rounds on the round course described in Section 3.2.1 had to be driven. The

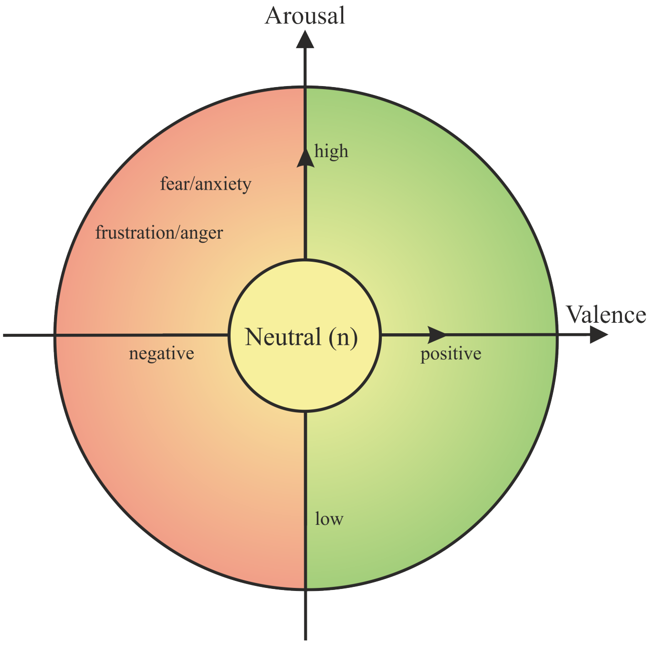


Figure 3.5: Defined target emotional states mapped onto the circumplexmodel of emotions concept (cf. [Russell & Lemay 2000], adapted from [Lotz; Ihme et al. 2018]).

inducement of the target emotions was based on two different approaches, by conducting a secondary task and by emotional elicitation through recalling memories of emotional significance (cf. Section 2.1.3). By utilizing emotion induction methods the emotion is verifiably reflected in the participant's facial expression, speech and physiological data (cf. Sections 2.1.2 and 2.1.3). Most of the recent benchmark emotional data sets are based on acted, scripted and task-induced natural emotions (cf. Section 2.1.5). The approach of emotion elicitation by recalling memories of emotional significance is hardly used. One benchmark data set utilizing this specific kind of emotion inducement is presented in [Martin et al. 2006] as the eNTERFACE'05 audio-visual emotion database. In this Thesis, for each target emotion a distinct driving scenario was designed, which was split into three phases:

1. Baseline driving (1st round):
   Driving without further disturbances of the participant. During this round, baseline measures of the peripheral physiological data were recorded. Furthermore, the participants could take the time to acclimatize to the driving situation and diminish influences of the previous scenario.

2. Secondary task (2nd & 3rd round):
   Inducement of the emotional state by conducting secondary tasks while driving. The task was designed in a way that it would induce the considered target emotion. This was mainly accomplished by realizing so-called WoZ experiments, where the user believes that she/ he is testing an autonomously working technical system, while the system is actually controlled by a human.

This information was withheld from the participants but later revealed during a debriefing.

3. Conversation-driven emotional recall (4th & 5th round):
Stimulating the recall of the considered target emotion by the participants' themselves by asking questions and starting a conversation on related topics while driving. The investigator initiated a conversation starting from the just experienced situation (i.e. just accomplished secondary task). A list of pre-defined questions and topics was available to the investigator to keep the conversation alive. Nevertheless, the interviewer was briefed to individualize the conversation, to sustain a natural-like interaction with the participant, and to prevent the conversation from developing in the wrong direction.

In between each of the four driving scenarios the participants had at least 5 minutes of recess to fill out provided self-reports (cf. next Section). This took on average 2 minutes. The rest of the time the participants could relax to get back to the neutral state. The order of the four driving scenarios was kept constant for all participants.

A detailed description of the driving scenarios is given in the following:

The participants started with the **neutral driving scenario**. As for all other scenarios, this scenario began with one round of baseline driving. In this scenario no secondary task was used to induce the driver state. During the 2nd and 3rd round of the course, the investigator initiated a conversation on neutral topics (e.g. educational background, commute to work, weather, etc.). During the 4th and 5th round of driving, another baseline driving was conducted. This was presented to the participants as a training phase.

Second, the participants underwent the **positive driving scenario**. They were told that before starting with the actual evaluation of the driver assistant systems, a testing of the audio setup was necessary. This was done by replaying a sound file via the loudspeakers during the 2nd and 3rd round of the driving course. As sound file two episodes of the funny radio podcast “Wir sind die Freeses” of the radio station NDR2 was used [Altenburg 2017]. The topics of the chosen episodes were based on, at that time, recent public events (i.e. personal assistants like Amazons Alexa and Bitcoin mining). Starting from these topics and depending on the participant's reaction on the radio show, the investigator initiated a conversation. This was done during the 4th and 5th round of driving.

Next, the **frustration driving scenario** was conducted. For this scenario, a WoZ-based navigation system was utilized. During the 2nd and 3rd round of driving, the participants were told that they should evaluated this *newly* developed navigation system. The system would only respond to speech comments and would not react to other modalities such as touch or gestures. The participant's task was to enter a specific address and start the routing. The system, however, was controlled by the technician seated behind the driver and would regularly misunderstand or not understand the participant's comments to induce frustration. During the 4th and 5th round, the investigator initiated a conversation on similar frustrating experiences also based on technical systems.

The **anxiety driving scenario** was the last one to be conducted for the participants. This scenario was also based on a WoZ setup, where the participants were told to evaluate the usability of a brake assistant. This task was conducted during the 2nd and 3rd round of driving. By utilizing the thinking aloud technique the participants were encouraged to use speech based feedback. The participants were told that the system was able to detect traffic cones at the side of the street, which were used to represent a person approaching the vehicle, and stop automatically in a sufficient distance to the obstacle. To inform the driver of the upcoming braking, a warning signal was replayed through the loudspeaker. The brake, however, was controlled by the investigator with the additional brake pedal at the passenger seat. To induce anxiety, the warning signal would sometimes occur without an actual braking of the car and vice versa. During the 4th and 5th round, the investigator initiated a conversation on similar alarming experiences in driving situations, eliciting anxiety.

The effective operation of the individual driving scenarios was validated in [Requardt et al. 2018] using the subjective self-reported feedback forms (cf. measures presented in the next Section) and the peripheral physiological data. A detailed description of the validation results is presented in Section 3.2.3.

Assessing the Driver's Subjective Emotional State

To assess the driver's subjective emotional state, three self-report measures were employed, namely, the Geneva Emotion Wheel (GEW) [Scherer et al. 2013], Self Assessment Manikins (SAM) [Bradley & Lang 1994] and free text input (cf. Section 2.1.4). The GEW contains 20 discrete emotion terms which can be rated in five intensity levels (cf. Figure 2.3 on page 22). The participants were asked to give feedback on all the emotions stated in the wheel. This corresponds to *alternative 3* presented in Section 2.1.4. For the SAM rating a 5-point Likert-scale was utilized to assess the emotional dimensions of valence (negative to positive) and arousal (low to high) (cf. Figure 2.2 on page 21). The free text input allowed the participants to describe the experienced emotional state in their own words.

To get a first insight on the quality of the emotion inducement methods used in the four driving scenarios, the different measures were inquired before, during and after the driving experiments. Before the experiment, all of the self-report measures were utilized to assess the participant's emotional baseline state. In between the different driving scenarios, the GEW was provided to the driver, accompanied by additional dummy scales to camouflage the actual purpose of the experiment. These scales included the Karolinska Sleepiness Scale [Akerstedt & Gillberg 1990], the

Stress Scale [Dahlgren et al. 2005] and the Scale of Thermal Sensation [Gagge et al. 1967]. After the experiments, a more detailed questionnaire focusing only on the experienced emotions during the drives was filled out by the participants. For this questionnaire the SAM scales were inquired separately for the conversation and task phase of the emotion induction. Furthermore, a free text input was possible to rate their experienced emotion during the secondary task.

3.2.3 Validating the Collected Data

In total 27.49 hours of audio data were recorded during the data collection. For each participant on average 54.99 minutes ± 4.89 minutes of audio material were obtained. To verify the usability and quality of the recorded data, the data was validated using the evaluation results of the drivers' subjective self-reports (i.e. GEW, SAM and free text input) and the peripheral physiological data (i.e. heart rate, finger temperature and skin conductance level). The presented results were achieved with the help of Dr. Klas Ihme, who is a post-doctoral researcher at the DLR, with a scientific background in psychology and cognitive science. As not all measures were available for each participant, only those participants providing a complete data recording (i.e. self-reported measures and peripheral physiology) were utilized for the validation. A full data recording was available for 28 of the 30 participants.

Evaluating the Drivers' Self-Reports

As first indicator for the quality of the recorded data, the results of the self-reported questionnaires obtained for each driver were evaluated. This was done by comparing the received outcome of the drivers' subjective self-reports of the GEW, SAM and free text input for each driving scenario with the induced target emotion. In order to perform this comparison, a pre-processing of the measures needed to be realized. Afterwards, these pre-processed results were compared to the induced target emotion of the considered driving scenario. A detailed description of this process will be given now.

Geneva Emotion Wheel:

For the GEW the items best describing the target emotions in the GEW were selected. For frustration and anxiety, these were the items *anger* and *fear*, respectively. In case of the positive target emotional state, a composite of the items *amusement*, *joy*, *pleasure* and *contentment* was formed, which was further referred to as positive affect scale. A description of the neutral state was not possible by considering certain items of the GEW, as the design of the GEW will always lead to a selection of emotions with low intensity (cf. Section 2.1.4). Whenever the GEW was applied during the driving experiment, the drivers were asked to give feedback on all the emotions they experience and their intensity in the present driving scenario. If a certain emotion was

Table 3.1: Mean and standard deviation of the descriptors of the GEW for the target emotions in the four driving scenarios (adapted from [Requardt et al. 2018]).

Scenario	Positive affect	Anger	Fear
Neutral	3.4 (0.9)	0.0 (0.2)	0.1 (0.6)
Positive	3.6 (1.0)	0.1 (0.3)	0.1 (0.4)
Frustration	2.9 (1.3)	0.6 (1.2)	0.1 (0.4)
Anxiety	3.1 (1.1)	0.2 (0.8)	0.3 (0.9)

not perceived at all during a drive, this emotion was rated with the intensity level “0”. The other intensity levels were rated increasing from “1” (low) to “5” (high). The intensity values, obtained for the items described above, were averaged over all drivers for each driving scenario and are stated in Table 3.1. By applying repeated-measures Analysis of Variances (ANOVAs) on each descriptor of the GEW (i.e. positive affect scale, anger and fear) the following was noticed: Even though, the positive affect scale showed considerably high values for all driving scenarios, a significant effect of the scenarios was observed (F(2.1,56.7) = 9.5, $p < 0.05$, Greenhouse-Geisser-corrected). Post-hoc t-tests revealed that the neutral and positive scenarios were rated significantly higher than the frustration scenario (all p’s < 0.05, Bonferroni-corrected). Furthermore, the positive scenario received significantly higher ratings than the anxiety scenario ($p < 0.05$, Bonferroni-corrected). For the item anger it was shown that there exists a significant difference regarding the driving scenarios (F(1.5,42.2) = 4.7, $p < 0.05$, Greenhouse-Geisser-corrected). Even though the item anger showed the highest intensity during the frustration scenario, non of the post-hoc t-tests showed a significant difference (all p’s $>$ 0.05, Bonferroni-corrected). The item fear was the only descriptor that did not show a significant difference in the driving scenarios (F(1.1,31.2) = 1.1, p = 0.305, Greenhouse-Geisser-corrected). However, the highest intensity of the item fear was present in the corresponding anxiety scenario.

Self Assessment Manikins:

In case of an emotion inducement based on a task and a conversation, the SAM was assessed twice, for each inducement method once. This was the case for the positive, frustration and anxiety scenario. For these scenarios, the values of the valence and arousal scale were averaged over both approaches to obtain one value per driving scenario. For the neutral scenario no task based induction was performed. Therefore, there also existed only one value of valence and arousal for this scenario.

The valence and arousal values averaged over all drivers are stated in Table 3.2. The highest valence value (high positive valence) was obtained for the positive

Table 3.2: Mean and standard deviation of the valence and arousal rating obtained through the SAM scale in the four driving scenarios (adapted from [Requardt et al. 2018]).

Scenario	Valence	Arousal
Neutral	4.1 (0.6)	2.1 (1.0)
Positive	4.4 (0.6)	2.1 (1.1)
Frustration	2.9 (0.8)	2.5 (0.9)
Anxiety	3.0 (0.6)	2.6 (1.0)

scenario and the lowest value (slight negative valence) for the frustration scenario. For arousal the highest value was received during the anxiety scenario and the lowest value for the neutral and positive scenarios. However, all arousal values obtained for the different driving scenarios indicate a moderate arousal. A repeated-measures ANOVA revealed that the driving scenarios strongly significantly affected the valence level of the drivers (F(2.3,63.2) = 61.9, $p < 0.001$, Greenhouse-Geisser-corrected). By conducting post-hoc t-tests it was shown that the valence level obtained for the positive and neutral scenario was significantly higher compared to the frustration and anxiety scenario ([neutral, positive] vs. [frustration, anxiety], all p's < 0.05, Bonferroni-corrected). A significant effect of the driving scenarios was also noticed for the arousal rating of the drivers (F(2.2,57.6) = 6.2, $p < 0.01$, Greenhouse-Geisser-corrected). Post-hoc t-tests revealed that the arousal experience in the anxiety scenario was significantly higher compared to the neutral and positive scenario (all p's < 0.05, Bonferroni-corrected).

Free Text Input:

During the experiment the drivers were asked several times to describe their current emotional state using a free-form input. This was done before the start of the actual driving task, and after the driving experiments were completed for each performed secondary task (i.e. radio show, navigation system and brake assistant). To get comparable results among the different drivers, the text input was analyzed in three steps: First, the text was digitized. Afterwards, as some of the participants did not stick to the provided examples of text input, but wrote whole sentences or gave general feedback on the situation or task, the text was reduced to only contain content related to their current experience. This was done by excluding non experience-related words and transforming all remaining words into adjectives. For example, frustration [German: Frustration] was transformed into frustrated [frustriert] and phrasings like “it was amusing” [“es war lustig”] were transformed into amused [belustigt]. Additionally, repetitions of words per secondary task and participant were excluded. In the third step the occurrence of each adjective was

Table 3.3: English translation of the free text input regarding the experienced emotion before the driving experiment (baseline) and during the secondary tasks of the positive, frustration and anxiety scenario. The word "amused" is mentioned twice, as the uttered German words "belustigt" and "amüsiert" are both translated with "amused" (adapted from [Requardt et al. 2018]).

	Words (count > 2)	
Baseline	excited (9) curious(4) happy (4)	interested (6) neutral (4) expected (3)
Radio show	amused (8) irritated (4) distracted (4)	relaxed (4) entertained (4) amused (3)
Navigation	irritated (11) upset (4) amused (3)	frustrated (6) misunderstood (3) uncertain (3)
Brake assistant	insecure (5) interested (4) surprised (3)	puzzled (4) excited (4) uncertain (3)

counted over all participants for each secondary task and the baseline inquiry separately. Furthermore, the resulting list of adjectives was translated into English. The translated results, containing words mentioned more than two times by different participants, are presented in Table 3.3. The original German words can be taken from [Requardt et al. 2018].

For the baseline survey, the participants were biased towards a positive state, which was reasonable, as the participants were full of expectation towards the upcoming experiment. This was reflected by emotional words describing the excitement and interest towards the upcoming driving experiment. Only four participants stated to be in a neutral state. However, "neutral" was only stated in the baseline survey and never during the report on the experiences during the secondary tasks. For the positive secondary task (listening to a funny radio show), the participants described to have experienced predominantly positive emotions. Nevertheless, also words related to negative emotions were expressed (irritated and distracted). This was expectable as the radio show targeted a specific type of humor, which is not necessarily perceived as funny by different participants. During the frustration task (evaluating a hands free navigation system) the participants expressed mainly words related to negative emotions (e.g. irritated, frustrated and misunderstood). However, three participants stated to be amused, which may be seen as a grim sense of humor. The final task, conducted to induce anxiety (evaluating an automatic brake assistant), was perceived as negative and positive. The expressed negative words were mainly related to an uncertainty of the participant, while the positive words "interested" and "excited" were elicited by the novelty of the conducted task.

Evaluating the Drivers' Peripheral Physiological Data

A more objective validation was obtained by evaluating the drivers' peripheral physiological data. The heart rate of the driver was determined by counting the number of R-waves per minute from the ECG signal. The finger temperature was directly taken from the raw signal obtained by the finger sensor. This sensor also provided the skin resistance, which was inverted to calculate the skin conductance level. As the physiological activity is strongly affected by inter-individual variability [James 1884], a reference value for each participant and driving scenario needed to be determined. This reference value was calculated by averaging the raw value of the considered physiological measure over a certain time span of the 1st round of each driving scenario (1 minute after start till end of 1st round). This part of the experiment was selected, as during the first round of driving for each scenario, the driver was unaffected by any further disturbances except the driving task itself. The obtained reference value was then subtracted from the average raw value obtained while a certain emotion was induced (2nd to 5th round of each driving scenario). These reference-corrected values of the four driving scenarios averaged over all participants are stated in Table 3.4.

For the heart rate, it was noticed that all reference-corrected values showed an increase compared to the reference value. By conducting a repeated-measures ANOVA it was shown that the driving scenario significantly affected the reference-corrected heart rate values ($F(3,84) = 3.52$, $p < 0.05$, $\varepsilon = 1$, no correction needed). Post-hoc t-tests revealed that the heart rate was significantly higher for the positive scenario compared to the neutral scenario ($p < 0.05$, Bonferroni-corrected). For anxiety, a clear increase of heart rate compared to the neutral scenario was observed. This difference was, however, not significant. A strongly significant effect of the scenarios was present for the finger temperature ($F(2.6,71.8) = 5.46$, $p < 0.01$, Huynh-Feldt-

Table 3.4: Reference-corrected mean values of the Heart Rate (HR), Finger Temperature (FT) and Skin Conductance Level (SCL) for the four driving scenarios. Brackets denote standard deviation (cf. [Requardt et al. 2018]).

Scenario	HR [bpm]	FT [°C]	SCL [$10^{-4}\mu$S]
Neutral	0.70 (2.72)	-0.12 (0.79)	5.33 (8.08)
Positive	2.85 (3.12)	0.23 (0.73)	2.87 (7.43)
Frustration	1.59 (3.74)	-0.22 (0.75)	3.17 (17.90)
Anxiety	2.24 (3.44)	-0.30 (0.73)	0.53 (12.40)

corrected). Except for the positive scenario all other scenarios showed a decrease in finger temperature compared to the reference value. A post-hoc t-test revealed that the increase of finger temperature during the positive scenario was even significant compared to the decrease of finger temperature occurring during the anxiety scenario ($p < 0.05$, Bonferroni-corrected). For the skin conductance level, no significant effect of the driving scenarios was noticed $F(1.7,49.5) = 1.12$, $p = 0.326$, Greenhouse-Geisser-corrected). Nevertheless, the skin conductance level was highest during the neutral scenario and lowest for the anxiety scenario.

3.2.4 Findings on the Real-World Driving Scenarios

From the results obtained through the drivers' self-report and peripheral psychological data the following was noticed for the four driving scenarios:

For the **neutral driving scenario** the GEW showed no significant differences compared to the positive scenario. Regarding the positive affect scale a significant decrease of intensity was observed for the frustration scenario. A similar observation was made for the valence and arousal values obtained through the SAM scale, where no significant differences between the results of the neutral and positive scenario were identified. Compared to the two negative scenarios the participants experienced a significantly higher valence and lower arousal. With respect to the physiological data it was shown that the heart rate was significantly lower compared to the positive scenario. The low heart rate obtained for the neutral scenario indicates a successful inducement of the neutral state. The results of the self-report, however, indicate that there is no significant difference compared to the positive scenario.

Regarding the **positive driving scenario**, the GEW indicated that the participants felt more positive compared to the frustration and anxiety scenario (significantly higher positive affect scale). This observation was also made for the valence scale, where significantly lower values were obtained for the negative driving scenarios. For the arousal scale a significant increase was noticed in case of anxious driving. From the free text input it was revealed that the participants felt amused, relaxed and entertained. However, also words related to negative emotions were expressed. As the utilized radio show targets a specific kind of humor, it can be assumed that perceived emotions differ strongly between the participants. The evaluation of the physiological data revealed that the heart rate was significantly higher compared to the neutral scenario. Furthermore, a significantly higher finger temperature compared to the anxiety scenario was observed. This is in line with the observations made by [Kreibig 2010], where it is stated that happiness comes along with an increased heart rate and finger temperature. More recent investigations hypothesize that skin temperature can be seen as a measure of control over the situation, associating higher finger temperature with a higher control (cf. [Fontaine et al. 2007] and [M. Zhang et al. 2018]). Overall, it can be concluded that the inducement of a positive emotional state seemed to be successful for a majority of the participants.

For the **frustration driving scenario** the GEW showed a significant lower intensity in the positive affect scale compared to the positive and neutral scenario. A similar observation was made for the obtained valence rating. For the arousal no significant difference compared to the other scenarios was observed. This is in line with the assumption of frustration being related to a rather negative valence and an only moderate arousal compared to anger having a moderate negative valence and a rather high arousal (cf. [Russell & Lemay 2000; Ihme; Unni et al. 2018] and [Ihme; Dömeland et al. 2018]). This also explains that, regarding the results obtained from the physiological data, no difference compared to the other scenarios was observed. Furthermore, for the item anger, obtained from the GEW, no significant effect regarding the frustration scenario was noticed. This indicated that the participants did not experience anger, but rather a mild negative emotion. This assumption is backed up by the results obtained through the free text input, where the participants stated to have felt *irritated*, *frustrated*, *upset* and *misunderstood*, which is closely related to frustration. None of the participants stated to have felt angry. As already indicated, some participants also mentioned words related to positive emotions. This, however, may be interpreted as a grim sense of humor. It can be concluded, that the induction of frustration worked very well for the presented driving scenario.

The interpretation of the results obtained for the **anxiety driving scenario** was more challenging. From the GEW it was shown that the experience of positive affects was lower compared to the positive scenario. The anger and fear items, however, showed no significant differences to other scenarios, but comparatively lower and higher intensity values, respectively. Regarding the SAM scale, the anxiety scenario obtained significantly lower valence and higher arousal values compared to the neutral and positive scenario. This is in line with the general classification ability of anxiety in the valence/ arousal-space as observed in [Fontaine et al. 2007]. Furthermore, the measured physiological parameters showed a significant decrease in finger temperature compared to the positive scenario. As stated above, the skin temperature is assumed to be related to the control over the situation, indicating that for the anxiety scenario a loss of control was experienced by the participants caused by the unforeseeable reaction of the tested brake assistant system. The results obtained from the free text input, however, showed that the participants did not experience pure anxiety but a rather mild form indicated by the words *insecure*, *puzzled*, *surprised* and *uncertain*. The word *anxious* was only mentioned once. It can be assumed that we did not accomplish to induce strong anxiety, but a milder state which is more related to uncertainty or insecurity. Therefore, this scenario will further be referred to as *mild anxiety scenario*.

3.3 Summary and Discussion

In this Chapter I presented two data sets, which were recorded in the scope of this Thesis to evaluate the ability to detect emotions from speech in an in-vehicle setup. A first data collection was performed inside a fixed-base simulator utilizing well-known benchmark data samples of the EmoDB and VAM data sets, re-recorded under silence (simulator turned off, only recording setup and in-vehicle acoustics influencing the recording) and disturbed (simulator turned on) conditions. This data collection served to evaluate the influence of in-vehicle noises on speech quality (cf. Section 4.3) and to analyze the effect of speech enhancement in speech emotion recognition (cf. Section 5.2). These effects can only be investigated when an information of the *clean* speech signal (original EmoDB and VAM) is available. The re-recorded data samples are further refered to as EmoDB-Car and VAM-Car.

From these simulated data, it is now possible to analyze the effect of in-vehicle noises on the speech quality, the features used for emotion recognition and other speech processing artifacts (e.g. speech enhancement), and give first insights on the feasibility of speech emotion recognition in a driving environment. Nevertheless, the naturalness of the in-vehicle speech is neglected. Therefore, a second data collection was realized. Specifically selected emotions, typically occurring in everyday driving situations (i.e. neutral, positive, frustration and anxiety), were induced without informing the participants about the actual goal of the data collection. This ensured that the participants were unbiased towards the experience of the target emotions. The data was recorded inside a test vehicle driving on real roads comparable to a quiet residential area. The collected data was validated by utilizing participants' self-reports and their peripheral physiological data. It was shown that the inducement of the positive and frustration scenario seemed to have been successful, regarding both self-reports and physiology measures. For the anxiety scenario only a milder state, more related to uncertainty or insecurity was experienced by the participants. The neutral driving scenario did not show significant differences in the participants' self-reports compared to the positive scenario. From the physiological data a clear difference was observed for the heart rate and finger temperature.

Overall it can be stated that the inducement of the four target emotions was successful. From emotion theory it is known that emotions are communicated by humans through changes in the facial expression, body movement and vocal expression. These feedback signals have not yet been considered. Therefore, in the scope of this Thesis, it further needs to be evaluated to which extend the experienced emotions are also reflected in the drivers speech. This annotation of the speech data will be presented in Section 5.1 and will be conducted in a much finer division, as emotions from speech are not continuously expressed by the speaker (cf. Section 2.1.4). The data with the final annotation will then further be utilized in Chapter 6 to evaluate the ability of detecting the driver's emotional state from real-world in-vehicle data. In the next chapter, I will draw the attention of the reader to speech quality and its impact on the emotion recognition task.

Chapter 4

Speech Quality Assessment and its Impact on Speech Emotion Recognition

Contents

FOR speech emotion recognition the quality of the speech signal is of great interest, as degraded speech samples will also lead to degraded speech features, due to an impaired quality of the speech signal. These features are further used to train the speech emotion recognizer. From literature it is well-known that the utilized feature set and recording setup can strongly influence the recognition performance (cf. Section 2.2.2). Therefore, it can be assumed that a degradation of the feature values will also influence the performance of a speech emotion recognizer. The big question is, how can we measure this signal degradation and does there exist a correlation between the level of degradation and the recognition performance? One way to measure the level of degradation of the signal is by utilizing a speech quality measure as presented in Section 2.4. To get an insight on the effect of degraded speech on speech quality measures and emotional speech, first investigations, realized in co-operation with colleagues of the Otto-von-Guericke University, were utilized on compressed speech data. In contrast to noisy speech data, where a data collection inside the considered noisy environment needs to be conducted to receive reliable data samples, compressed speech data is easy to obtain by applying common audio codecs on available emotional speech data sets. In this context, internal differences occurring during the transmission of the signal (e.g. jitter or package loss) are not further taken into consideration. To get a better insight on the utilized audio codecs, a short introduction on audio compression is given in Section 2.5.3 and the most relevant audio codecs are presented in Appendix B. Further, a brief overview on existing quality measures and a detailed description of the most relevant ones, in the scope of this Thesis, are given in Section 2.4.1. In the present Chapter, I will introduce an own measure based on the work presented in [Siegert; Lotz; Duong et al. 2016], which can be used to describe the difference between the power spectrum of two correlated speech signals (Compression Error Rate (CER)) (cf. Section 4.1.2). Afterwards, this novel measure as well as the MOS - Listening Quality Objective (MOS-LQO) and an adapted version of the Signal-to-Noise Ratio (SNR) (cf. Section 4.1.1) will be applied to compressed (cf. Section 4.2) and noisy speech (cf. Section 4.3), respectively, and it will further be investigated how the speech quality affects the ability to recognize the speaker's emotional state. The results presented in these two Sections are based on [Siegert; Lotz; Maruschke et al. 2016; Lotz et al. 2017] and [Lotz; Faller et al. 2018]. While for the investigations performed on compressed speech, a selected number of codecs is applied to the well-known EmoDB data samples, for the investigation performed on noisy speech, the re-recorded EmoDB under in-car recording conditions (EmoDB-Car) and re-recorded VAM under in-car recording conditions (VAM-Car) data sets presented in Section 3.1 are utilized. This Chapter will be concluded with a statement on how well the utilized speech quality measures can be used to assess the recognition performance of a speech emotion recognizer. As parts of this Chapter are based on work already published in [Siegert; Lotz; Duong et al. 2016; Siegert; Lotz; Maruschke et al. 2016; Lotz et al. 2017] and [Lotz; Faller et al. 2018], several phrasings are taken literally from these publications.

4.1 New Measures for Speech Quality Assessment

With regards to the commonly used quality measures MOS-LQO and SNR, as presented in Section 2.4.1, I will first present an adaptation of the well-known SNR measure, so that it can also be applied to data of varying recording conditions (cf. Section 4.1.1). As most quality measures have advantages and disadvantages when applied to compressed or noisy speech, respectively, I will furthermore present the self-developed CER, which I introduce to describe the direct difference between the power spectrum of the original and the compressed/ noisy speech signal (cf. Section 4.1.2). This measure can be utilized to get a general overview on how much the original speech signal was affected by a degradation through compression or noise. When the signal is compressed it can be assumed that the power of the compressed speech signal decreases, as signal parts which carry less information for the later application domain are being discarded. In case of noisy speech, the power of the noisy speech signal increases compared to the reference signal, as the noise is superimposed to the clean speech signal. Both, increase and decrease of the power spectrum can have a negative impact on the recognition performance. Furthermore, I investigate, whether an easy to calculate difference in spectral power can already give information on a later speech emotion recognition performance. One advantage of the newly introduced CER is that the signal is first segmented and then analyzed segment by segment in frequency domain. From literature there already exist several measures based on a segmentation of the signal in frequency domain. These are for example the segmented SNR, the spectral distance and the Bark distortion measure (cf. [Loizou 2011]). For the segmented SNR it can be assumed, that the calculation will have the same disadvantages as the SNR, as described in Section 4.1.1. The spectral distance is based on the cepstral coefficients of the clean and disturbed speech signal. As some of the relevant speech features used for speech emotion recognition are based on the cepstral coefficients (cf. Section 2.2.2), it is assumed that this difference will only describe the changes in the recognition performance affected by these features. The Bark distortion measure is closely related to the newly introduced CER, as it determines the mean difference between the loudness spectra of the clean and disturbed speech signal by utilizing the Bark frequency scale [Loizou 2007]. Other authors tried to use differences in paralinguistic features extracted from the speech signal itself like LPC coefficients, fundamental frequency, formants or spectral center of gravity (e.g. [Son 2005]), which are all based on the speech signal in frequency domain. These measures, however, are also designated features used in speech emotion recognition, such as the cepstral coefficients described previously.

4.1.1 Adapted SNR

In this Thesis, the SNR will be used to investigate the influence of acoustic conditions and in-vehicle noises on the clean speech signal and the ability to automatically detect the driver's emotional state. This will be done by utilizing the original EmoDB and VAM data sets and their corresponding re-recordings under silence and disturbed recording conditions (EmoDB-Car and VAM-Car). To do so, the speech samples of the original data sets will be compared to the corresponding re-recordings. However, when utilizing the SNR certain limitations need to be considered, which will be explained in detail now.

As defined in Equation (2.18) the power of the clean speech signal as well as the power of the noise signal need to be available under similar recording conditions to correctly calculate the SNR. This is not always the case, as, depending on the application domain, not all relevant power values can be determined. Whenever one of the required signal streams is not measurable using a designated microphone setup, an estimate of the SNR needs to be determined with regard to the unknown power value. For this investigation, as there exist only recordings of the clean speech signal and the noisy speech signal, an estimate of the SNR was utilized based on an estimate of the power of the noise signal (P_n). This was done by subtracting the power of the clean speech (P_s) from the power of the noisy speech (P_{ns}), under the assumption of superposition of multiple sound sources (cf. Section 2.5.1)

$$P_n = P_{ns} - P_s. \tag{4.1}$$

Applied to Equation (2.18) this led to

$$\widehat{SNR}_{dB} = 10 \cdot \log_{10}(\frac{P_s}{P_{ns} - P_s}), \tag{4.2}$$

as an estimate of the SNR in the logarithmic decibel scale. This assumption is valid for similar recording conditions of the clean and noisy speech signal. For the utilized data sets of the EmoDB-Car and VAM-Car data sets this corresponds to the re-recordings under silent and disturbed recording conditions, but not to the recordings of the original EmoDB and VAM data samples. In this Chapter I will, therefore, distinguish between two types of SNRs, which I refer to as *real SNR* and *relative SNR*. The *real SNR* denotes the estimate of the SNR ($\widehat{SNR}_{dB}$) as presented in Equation (4.2). The *relative SNR* denotes an estimate based on a clean speech signal obtained under "ideal" recording conditions (i.e. original EmoDB and VAM data samples unaffected by the in-car recording setup), which are not comparable to the recording conditions of the EmoDB-Car and VAM-Car re-recordings. These SNR values are only comparable among other SNRs based on this reference clean speech signal. To determine the *relative SNR*, the estimate presented in Equation (4.2) needed to be adapted. This led to two major difficulties, which will now be addressed and solved if possible:

Issue 1:

For the re-recordings the volume of the original speech samples was adjusted to match the loudness of a speaker inside a running vehicle (not considering the Lombard-effect occurring in natural human communication in noisy environments). Additionally, the acoustic conditions inside the simulator vehicle suppressed the speech signal replayed by the loudspeaker. This led to a general reduction of the signal's power.

Mitigation Strategy 1:

This issue could presumably be solved by normalizing the original and recorded speech samples to the same loudness. However, a speech signal is a highly dynamical signal and already small changes in the recording setup lead to changes in the waveform of the signal and subsequently to differences in the shaping of the frequencies. Therefore, none of the common normalization methods, as standardization or range normalization, can be utilized, as they would distort the characteristics of the speech signal. An alternative normalization method utilized for speech signals is the, so called, peak normalization. For this normalization approach, the speech signal gets normalized to a desired maximum amplitude of the waveform (dB). It is clear that this is only feasible if the in-vehicle acoustic and superimposed noise do not manipulate the waveform of the original speech signal, leading to a shift of the maximum amplitude. This, however, was especially the case for the re-recordings under disturbed recording conditions, as the amplitude of the noise signal exceeded the maximal amplitude of the speech part. This led to a normalization of the signal to the maximum amplitude of the noise and not the speech content of the signal. Furthermore, already small differences in the signal waveform accumulated to large differences in the signal's power, leading to deficient SNR values. This lies in the nature of the SNR, as it is calculated based on the total power of the speech signal. Consequently, it was concluded that Issue 1 was not satisfactorily solved by applying normalization.

Issue 2:

From a theoretical perspective it could be assumed that the SNR of the loudness normalized original speech sample and the recording under silence condition was considerably high, as only the in-vehicle acoustic and the microphone setup affected the recorded speech sample. However, as the power of the original speech signal was suppressed during the re-recording, this led to a lower signal power of the re-recorded speech sample compared to the original sample. When estimating the noise power by applying Equation (4.1), this led to a negative denominator in Equation (4.2). As the calculation formula of the SNR is only valid for positive ratios between speech and noise power, this led to an incorrect SNR value including imaginary parts.

Mitigation Strategy 2:

To cater for the suppression of the power of the original speech signal during the re-recording, the power of the clean speech signal P_s was weighted using a positive constant α, based on the approach presented in [Botinhao & Yamagishi 2017]

$$SNR_{dB,\alpha} = 10 \cdot \log_{10}(\frac{\alpha \cdot P_s}{P_{ns} - \alpha \cdot P_s}). \tag{4.3}$$

> Here, α suppressed the clean speech power and was chosen so that
>
> $$P_{ns} - \alpha \cdot P_s > 0 \tag{4.4}$$
>
> was true for all speech samples. To maintain the comparability of the different speech samples, α was calculated based on the speech samples of each utilized data set under silence recording condition. Then the minimum α value, that satisfied condition (4.4) was selected as global α and applied for the calculation of the $SNR_{dB,\alpha}$ under silence and disturbed recording conditions of the considered data set.

As the power values of the re-recorded speech samples were not normalized to the loudness of the original speech samples, α equated to a very small value ($\alpha_{\text{EmoDB-Car}} = 0.004, \alpha_{\text{VAM-Car}} = 0.0539$) indicating a strong suppression of the original clean speech power. Therefore, the obtained $SNR_{dB,\alpha}$ values are only comparable within the two recording setups and not to other SNR values stated in the literature. In the further context of this Thesis, this modified SNR value is referred to as *relative SNR*.

4.1.2 Compression Error Rate

In [Siegert; Lotz; Duong et al. 2016] Ingo Siegert and I investigated the impact of audio compression on the spectral quality of speech data. Here, I introduced a measure to assess the quality of speech by determining the differences occurring in the spectrum of the uncompressed high quality speech samples compared to the compressed version of said speech signal. This measure was further referred to as Compression Error Rate (CER). In the development process of the CER two different versions were introduced and utilized in [Siegert; Lotz; Duong et al. 2016; Lotz et al. 2017; Lotz; Faller et al. 2018]. Depending on the use case either a CER based on the average difference between the Power Spectral Density (PSD) in [dB] of the compressed and uncompressed speech signal (CER_{dB}) or the percentage difference between the PSD of the signals ($CER_{\%}$) was applied. For within data set evaluations the CER_{dB} gives a direct insight on the differences occurring in the signal energy. With regard to between data set evaluations, the differences in the signal energy are strongly dependent on the utilized data set. Therefore, the $CER_{\%}$ should be utilized.

Figure 4.1 gives an insight on how audio compression effects the PSD values of the speech signal. A clear deviation between the clean and compressed speech signal is observed. For the utilized speech codec (MPEG-1/MPEG-2 Audio Layer-3 (MP3) with 24 kbit/s) it was noticed that especially the higher frequency bins were strongly affected by the compression. This was in line with the statement made in [Eppinger & Herter 1993], where it is stated that especially frequency bins up to

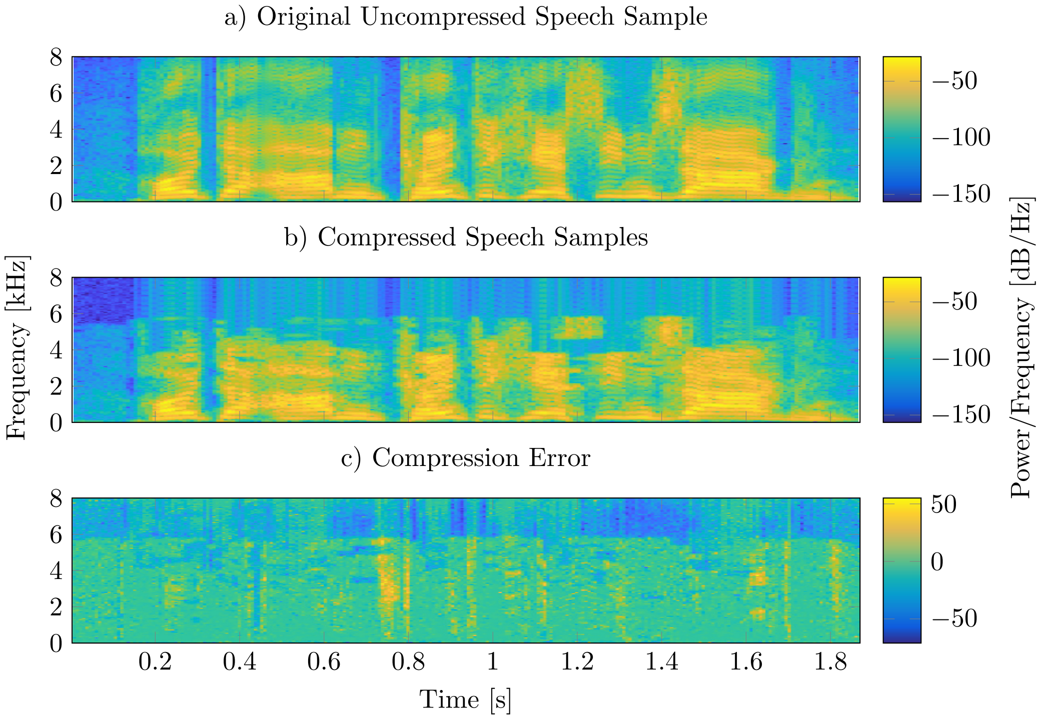


Figure 4.1: Power spectrum of an a) uncompressed and b) compressed speech sample when utilizing the MP3 encoder and decoder with 24 kbit/s. c) depicts the difference between the PSD-values of a) and b) in [dB].

4 kHz contribute to a good speech intelligibility, which is the desired goal of most audio codecs (cf. Section 2.5.3). From the results presented in [Siegert; Lotz; Duong et al. 2016] it was noticed that this phenomenon occurred for all evaluated audio codecs except for Speex (SPX). Furthermore, it was noticed that compression led to both in- and decreases of the PSD values. An increase of the PSD values was mostly observed in regions of low spectral power where no speech was present. Especially, audio codecs designed for music compression, as it is the case for MP3, showed this behavior. This was reasonable, as music does not contain large parts of silence. With an increase of the bit rate this phenomenon was attenuated.

The two approaches to calculate the CER will now be presented and discussed. It should further be noted that the development of the CER was, until recently, an ongoing process. Therefore, the result presented in this Chapter of the Thesis were solely based on the newest versions of the CER and may differ from the results presented in [Siegert; Lotz; Duong et al. 2016], [Lotz et al. 2017] and [Lotz; Faller et al. 2018].

Actual Change of Power Spectral Density in Compressed Speech (CER_{dB})

The first version of the CER, as presented in [Siegert; Lotz; Duong et al. 2016], was based on the averaged absolute difference between the spectral power of the compressed and uncompressed power spectrum in each speech segment and was referred to as CER_{dB}.

To calculate the CER_{dB}, first the power spectrum of both signals was calculated. This was done by applying the Wiener-Khintchine theorem, which defines the PSD as Discrete-Time Fourier Transform (DTFT) of the auto-correlated speech signal series ($s_{xx}[\kappa]$)

$$S_{XX}(k) = \sum_{\kappa=-\infty}^{\infty} s_{XX}[\kappa] \exp -i\frac{2\pi}{N}\kappa k, \tag{4.5}$$

where N denotes the sample size of the signal and k the frequency bin index ($k = 0, ..., N-1$). As this definition is only valid for stationary signals and speech is highly non-stationary, Equation (4.6) was applied to short consecutive signal segments of 12.5 ms length and an overlap of 5 ms, for which the signal is assumed to be short-time stationary [Wendemuth 2004]. Depending on the sampling rate F_s of the speech signal, the number of samples inside each segment was given as $N = F_s * 0.0125s$. The PSD was then calculated for each speech segment as

$$S_{XX}(t,k) = \sum_{\kappa=-\infty}^{\infty} s_{XX}[t,\kappa] \exp -i\frac{2\pi}{N}\kappa k, \tag{4.6}$$

with t as segment index ($t = 0, ..., T-1$, $T \hat{=}$total number of segments) and the auto-correlation series $s_{XX}[t,\kappa]$ determined as

$$s_{XX}[t,|\kappa|] = s_{XX}[t,-|\kappa|] = \frac{1}{N}\sum_{n=0}^{N-1-|\kappa|} x[t,n]x[t,n+|\kappa|], \tag{4.7}$$

where κ denotes the overlap of the speech segment with itself in terms of sample number.

Second, the difference between the PSD values of the compressed ($s_{XX,cs}[t,|\kappa|]$) and uncompressed/ clean speech ($s_{XX,s}[t,|\kappa|]$) was calculated for each speech segment in the logarithmic decibel scale as

$$\Delta S_{XX}(t,k) = 10 \cdot \log_{10}(S_{XX,cs}(t,k)) - 10 \cdot \log_{10}(S_{XX,s}(t,k)). \tag{4.8}$$

Negative $\Delta S_{XX}(t,k)$ values denote a decrease in the spectral power in dB of the compressed speech segment in frequency bin k compared to the uncompressed counterpart and positive values an increase, respectively. To get a general statement on the spectral power difference of the whole signal and not only on the considered segment and frequency bin, the Root Mean Square (RMS) of the power spectral difference for each signal segment was calculated by

$$RMS_{\Delta}(t) = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} \Delta S_{xx}(t,k)^2}. \tag{4.9}$$

Afterwards, the CER_{dB} was determined by averaging the $RMS_{\Delta}(t)$ over the total number of signal segments. This equated to

$$\begin{aligned} CER_{dB} &= \frac{1}{T}\sum_{t=0}^{T-1} RMS_{\Delta}(t) \\ &= \frac{1}{T}\sum_{t=0}^{T-1} \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} (10 \cdot \log_{10}(S_{XX,cs}(t,k)) - 10 \cdot \log_{10}(S_{XX,s}(t,k)))^2}. \end{aligned} \tag{4.10}$$

Negative CER_{dB} indicate an average decrease of the signal power in dB compared to the original clean speech signal and positive values an average increase. A value of 0 dB with an standard deviation of 0 dB indicates no changes of the signal power compared to the original clean speech signal. By utilizing the CER_{dB} on data originating from one data set it is possible to get a direct insight on the average changes occurring in the power spectrum of these signals.

Relative Change of Power Spectral Density in Compressed Speech ($CER_{\%}$)

The CER_{dB} only gives information on the actual change in spectral power and not on the relative change with regard to the original clean speech sample. Therefore, an adaption of the CER was realized so that the difference between the two power spectrums was given as a percentage value. By introducing this adaptation in the calculation of the CER, it was further possible to also compare samples originating from different data sets with each other, which was a limitation of the CER_{dB}. This adapted value was referred to as $CER_{\%}$. The $CER_{\%}$ was also based on the PSD values $S_{XX}(t,k)$ of each segment t and frequency bin k of the compressed and uncompressed speech signal. In contrast to the CER_{dB}, the $CER_{\%}$ was calculated by determining the percentage difference between $S_{XX,cs}(t,k)$ and $S_{XX,s}(t,k)$. To do so, the compression rate (CR) between the PSD values of the signals of each segment and frequency bin was calculated in the logarithmic decibel scale as

$$CR_{dB}(t,k) = \frac{10 \cdot \log_{10}(S_{XX,cs}(t,k))}{10 \cdot \log_{10}(S_{XX,s}(t,k))}. \tag{4.11}$$

Then the average ratio over all frequency bins in one signal segment was determined and averaged over all speech segments, and, afterwards, transformed into percentage values

$$CER_{\%} = 100 - 100 \cdot \frac{1}{T \cdot N} \sum_{t=0}^{T-1} \sum_{k=0}^{N-1} CR_{dB}(t,k). \tag{4.12}$$

In this process, outliers of the $CR_{dB}(t,k)$ values were excluded to overcome a bias of the average towards these values. In practice, the mean value excluding the outliers corresponded to the $Q_{0.5}$-Quantile and is identical to the median of the $CR_{dB}(t,k)$ distribution [Bleymüller & Weißbach 2015]. As for the the CER_{dB}, positive $CER_{\%}$ values indicate an increase of the signal power and negative values a decrease. A value of 0% with a standard deviation of 0% indicates no change compared to the original clean speech signal.

The CER as presented in Equations (4.10) and (4.12) is not limited to the application of compressed speech signals and can also be used to evaluate the distortion of a noisy speech signal compared to its clean speech counterpart. For this case $S_{XX,cs}(t,k)$ is replaced with $S_{XX,ns}(t,k)$, the PSD of the noisy speech signal. Even though the corresponding CER values are not anymore related to audio compression, we will further refer to this measure as the Compression Error Rate. As for the SNR, it should be noted that, when applied to the re-recordings of the EmoDB-Car and VAM-Car data set, there exists a general reduction of the signal's power compared to the original data samples. For the CER, however, no adaption of the measures' equation is necessary, as it is based on the PSD values, which describe the speech intensity per frequency bin and not the power of the signal over time.

Similarities and Differences to the Bark Distortion Measure

At the beginning of this Section I have mentioned the high similarity between the Bark Distortion Measure (BSD) (cf. [Loizou 2007]) and the newly developed CER. Therefore, I will now briefly introduce the BSD and its area of application. As for the CER the calculation of the BSD is based on the difference between the clean and the reference speech signal in their signal power:

$$BSD(k) = \sum_{b=1}^{N_b} [S_k(b) - \bar{S}_k(b)]^2, \tag{4.13}$$

with N_b being the number of critical frequency bands b of the bark frequency scale, and $S_k(b)$ and $\bar{S}_k(b)$ corresponding to the loudness spectra of the clean and enhanced signal, respectively. The bark scale is a psycho acoustic scale which divides the frequency scale into 24 critical frequency bands of human hearing. By utilizing this distribution of frequencies the measure takes into account psycho acoustic information of the human listening process, which is also indicated by its high correlation with the Mean Opinion Score (MOS) ($\rho = 0.9$) used to determine the listening quality of audio signals. Despite the fact that the BSD utilizes the enhanced speech signal as reference signal and the CER the compressed/ noisy speech signal, major difference lie in the utilized frequency scale and the applied statistics (i.e. mean square vs. root mean square). As the CER was not intended to describe the listening quality but rather the speech/ signal quality, and provide a general overview on the the signal's information loss, the application of the BSD would not meet these demands. Furthermore, by utilizing the RMS the physical quantity of the PSD is maintained.

4.2 Exploring Compressed Speech

In this Section I will investigate the effect of audio compression on speech emotion recognition based on the results published in [Lotz et al. 2017]. For that, I will first examine how well human labelers are able to perceive emotions from compressed speech. Afterwards, comparable automatic speech emotion recognition experiments are presented. The Section will be concluded by evaluating the correlation between speech quality and the results of the human labeling and speech emotion recognition, respectively.

To investigate the effect of audio compression three well-known lossy audio codecs were utilized (i.e. MP3, SPX and AMR-WB). In Appendix B an overview on the most relevant codecs and their application domain is given. To get a broad overview on the effect of audio compression, the codecs were chosen such that they would cover the three main application purposes: music (MP3), internet telephony/ Voice over Internet Protocol (VoIP) (SPX) and high quality mobile telephony/ Voice over LTE (VoLTE) (AMR-WB). Additionally, the lossless audio codec Free Lossless Audio Codec (FLAC) was utilized. As reference format the standard Waveform Audio File (WAV) format was used. The audio codecs were applied to the speech samples of the EmoDB data set by utilizing the encoders and decoders listed in Table 4.1. A convenient framework for audio and video conversion is provided by the software FFmpeg[1] (Version 2.8.2). It provides encoders and decoders for a broad range of audio codecs. Some codecs, however, are supported by external libraries. As some of the utilized audio codecs are open source, these designated reference encoders and decoders were used instead of FFmpeg. This was the case for FLAC encoding, MP3

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

Table 4.1: Overview on the utilized audio codecs and bit rates.

Name	WAV	FLAC	MP3
Compression	No	Yes	Yes
Lossless	–	Yes	No
Bit rate [kbit/s]	265	–	8, 16, 24, 32, 64, 96
Compression level	–	0-8	–
Encoder	–	flac	lame
Decoder	–	ffmpeg	lame
File size [%] of WAV	100	78.8 - 76.4	3.34, 6.47, 9.71, 12.94, 26.24, 39.36

Name	Speex	AMR-WB
Compression	Yes	Yes
Lossless	No	No
Bit rate [kbit/s]	6.6, 11.11, 22.06	6.6, 12.65, 23.85
Compression level	1, 3, 6	–
Encoder	speexenc	ffmpeg (libvo-amrwbenc)
Decoder	speexdec	ffmpeg
File size [%] of WAV	2.76, 4.34, 8.62	2.83, 5.18, 9.57

and SPX encoding and decoding, where flac[2], lame[3] as well as speexenc and speexdec [Valin 2007] were utilized, respectively. For AMR-WB the external encoding library libvo-amrwbenc of ffmpeg was utilized. Table 4.1 also gives a general overview on the utilized audio codecs and applied bit rates (for more details on bit rates and compression please refer to Appendix B). For FLAC it should be noted that the bit rates correspond to average rates over all compressed speech samples, as this codec is based on different compression levels with variable bit rates, to dynamically adapt to the complexity of the audio signal. Furthermore, the file size of the compressed speech signal is stated as average value over all speech samples of the EmoDB data set. Previous investigations, presented in [Siegert; Lotz; Duong et al. 2016], showed that with increasing bit rates for the MP3 codec, a saturation of the compression was reached and no further changes in the file size occured. This was also confirmed by applying the CER, which saturated from MP3 and 96 kbit/s upwards with a CER_{dB} of 2.76 dB.

4.2.1 Human Speech Emotion Perception

There only exists few research work on the ability of a human to perceive the emotional content of compressed speech (cf. [Labelle et al. 2016] and [Lahaie et al. 2017]). These studies, however, emphasize on bandwidth limitations and compression using the AMR-WB codec. To get a broader insight on the effect of audio compression on human emotion perception from speech, further listening experi-

[2] https://xiph.org/flac/index.html

[3] https://lame.sourceforge.io/index.php

ments were conducted on the original and compressed version of a sub set of the EmoDB data samples.

Listening Experiment

For the listening experiment seven native German speaking labelers (five females) were employed. None of the labelers had participated in this kind of listening experiment before. Before conducting the actual labeling task, each labeler underwent a training phase in which they had to listen to one original uncompressed sentence of the EmoDB data set spoken by one speaker in all seven emotion categories. This ensured that all labelers were able to distinguish between the different emotions uttered by the actors. The training samples were excluded from the samples set used for the listening experiment. During the experiment, the labelers listened to a subset of the EmoDB data samples under 12 encoding conditions (cf. Table 4.1) and the original WAV format. The FLAC format was excluded from the experiment, as no information loss was present and, therefore, no difference in the speech signal was audible compared to the original WAV format. The task of the listening experiment was to assign one out of seven emotion categories (i.e. anger, boredom, disgust, fear, joy, neutral and sadness) to the heard speech sample. This was done by utilizing the software tool *ikannotate* (cf. [Böck et al. 2011]). As listening experiments are highly time consuming and fatiguing, only a number of 26 different speech samples of the EmoDB data set was utilized. This subset comprised four different sentences spoken each by one speaker (two female and two male speakers) in all seven emotion categories, if possible. This resulted in four speech samples for each emotion category except for *joy* and *sadness*. For these categories only three speech samples were utilized. To sum up, each labeler listened to 338 speech samples (26 samples $\cdot$ (12 codecs + 1 reference) = 338). After each 26 speech samples the labeling was interrupted by a music file of 3 minutes length, this was done to reduce the probability of a labeler to memorize the speakers voice in combination with a certain emotion category. The samples were presented in pseudo-random order (i.e. no consecutive samples of the same speaker or emotion category) to overcome memorizing effects, but kept in this fixed order for each labeler to guarantee comparability of the experiment's results. On average the listening experiment took 100 minutes per labeler.

Experimental Results

To get an overall insight on the results of the listening experiment, the Unweighted Average Recall (UAR) over all emotions and labelers for each considered audio codec was determined (UAR_{h}). The results are presented in Figure 4.2. It can be seen that for all considered codecs and bit rates a UAR of over 86% was achieved. A repeated-measure Analysis of Variance (ANOVA) (cf. Section C.1) revealed that there exists no significant difference between the results of the listening experiment on com-

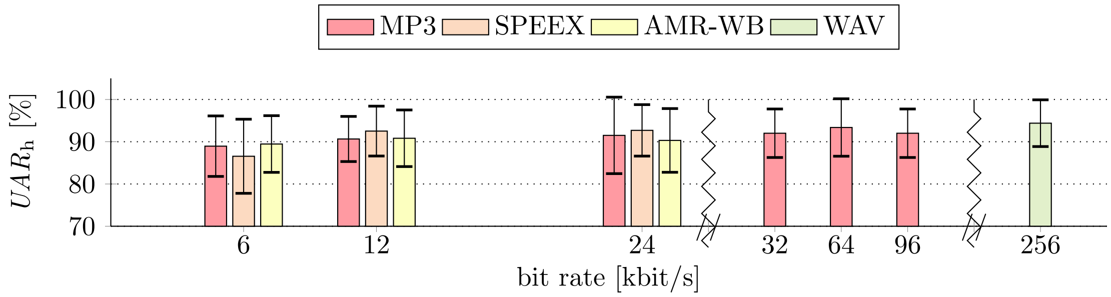


Figure 4.2: Mean and standard deviation of the UAR of the listening experiment (cf. [Lotz et al. 2017]).

pressed and uncompressed emotional speech (main effect codec: F(3.0,17.9) = 1.16, p > 0.05, Greenhouse-Geisser corrected). The lowest UAR was obtained for SPX with 6.6 kbit/s and the highest for the uncompressed WAV format. Furthermore, with increasing bit rate also an increase of the UAR was noticed for all evaluated audio codecs, except for MP3 with 96 kbit/s. For this audio codec configuration the UAR decreased. It was further noticed that not even for the uncompressed WAV samples an agreement of nearly 100% was achieved. This was reasonable, as the ground truth of the EmoDB data set relies on the emotional state given to the actors during the data collection and shows an emotion recognizability of over 80 % and naturalness of over 60% (cf. Section 2.1.5). This was in line with the emotion wise results presented in Table 4.2 where for each considered emotional state an UAR above 80 % was achieved during the listening experiment.

To evaluate how good the listeners were able to perceive the separate emotions, the UAR for each emotion averaged over all labelers ($UAR_{\mathrm{h,emotion}}$) was determined. The results obtained for each considered audio codec are stated in Table 4.2. Red and green entries indicate values outside of the standard deviation (above-green, below-red) and bold entries the highest UAR for each considered emotion. Overall, all emotions were perceived by the listeners with an UAR of at least 75%. Only for three cases an average recall of over 80% was not reached. This was the case for the

Table 4.2: $UAR_{\mathrm{h,emotion}}$ for each considered codec configuration. Recognition results outside of the standard deviation of each emotion are highlighted in red (below) and green (above). The best results are denoted in bold font.

codec	WAV	MP3	MP3	MP3	MP3	MP3	MP3	SPX	SPX	SPX	AMR	AMR	AMR
bit-rate	256	8	16	24	32	64	96	6.6	11.11	22.09	6.6	12.65	23.85
anger	**100**	**100**	**100**	**100**	96.43	**100**	96.43	**100**	**100**	96.43	96.43	96.43	96.43
boredom	92.86	85.71	85.71	**96.43**	92.86	92.86	92.86	82.14	**96.43**	89.29	78.57	85.71	85.71
disgust	**96.43**	82.14	85.71	85.71	82.14	92.86	82.14	82.14	89.29	85.71	82.14	89.29	89.29
fear	85.71	82.14	85.71	82.14	85.71	**92.86**	85.71	75.00	89.29	89.29	82.14	82.14	82.14
joy	**95.24**	90.48	**95.24**	**95.24**	**95.24**	90.48	**95.24**	90.48	90.48	**95.24**	90.48	90.48	**95.24**
neutral	**100**	96.43	96.43	85.71	96.43	89.29	96.43	**100**	96.43	92.86	96.43	96.43	92.86
sadness	90.48	85.71	85.71	95.24	95.24	95.24	95.24	76.19	85.71	**100**	**100**	95.24	90.48

emotions boredom, fear and sadness for the lowest bit rates of the codecs SPX and AMR-WB. These two audio codecs also showed the lowest UAR_{h}. Especially for SPX with 6.6 kbit/s, low UARs were achieved for a majority of the emotions. As each emotion was only represented with four or three samples in each audio codec and bit rate, already a small amount of misclassified samples led to a strong decrease of $UAR_{\mathrm{h,emotion}}$. To increase the validity of the results a more representative sample set needs to be considered. This, however, is only possible by increasing the number of speech samples in the listening experiment. As an increase of samples will lead to an increased duration of the listening experiment and therefore can lead to fatigue of the listeners, the listening experiment would need to undergo other limitations, for example, by considering certain emotional states only or by conducting several sessions of the experiment with the same subjects. This approach would on the one hand lead to a better validity of the experiment, but on the other hand also to much higher costs. This was not possible in the scope of this Thesis. The emotion wise results presented in Table 4.2 can, therefore, be seen as a first attempt towards investigating the ability of humans to recognize emotions from compressed speech.

4.2.2 Automatic Speech Emotion Recognition

Comparable state of the art speech emotion recognition experiments as presented in Section 4.2.1 were carried out to investigate to which extent audio compression influences the recognition performance and consequently the features used to automatically recognize the emotional state of a speaker. To accomplish speaker independency, the Leave-One-Subject-Out (LOSO) validation scheme was applied, resulting in 10 independent classification experiments (i.e. one per subject/ speaker) per considered audio codec configuration. Furthermore, the training and test sets were obtained by utilizing the same codec configuration (within codec classification experiments). As classifier a Support Vector Machine (SVM) with linear kernel ($C = 1$)) was applied using the software tool WEKA [Hall et al. 2009]. As feature set all features of the *emobase* set were utilized (cf. Section 2.2.2).

Similar to the results presented in Figure 4.2 and Table 4.2, Figure 4.3 and Table 4.3 present the corresponding results of the speech emotion recognizer. As for the listening experiment the UARs were calculated by averaging the recall over all carried out classification experiments for each codec (UAR_{a}). These results are stated in Figure 4.3. A repeated-measures ANOVA revealed that there exists a significant difference in the results obtained for the different codec configurations (main effect codec: F(3.5,31.7) = 3.87, $p < 0.05$, Greenhouse-Geisser corrected). Subsequently carried out post-hoc t-tests showed that significant differences occurred in the recognition performance between MP3 with 8 kbit/s and MP3 with 32, 64 and 96 kbit/s (all p's < 0.05, Bonferroni-corrected). Similar observations were made for MP3 with 16 kbit/s and MP3 with 32 kbit/s ($p < 0.05$, Bonferroni-corrected). For MP3 and AMR-WB an almost continuous increase of the UAR with an increase of

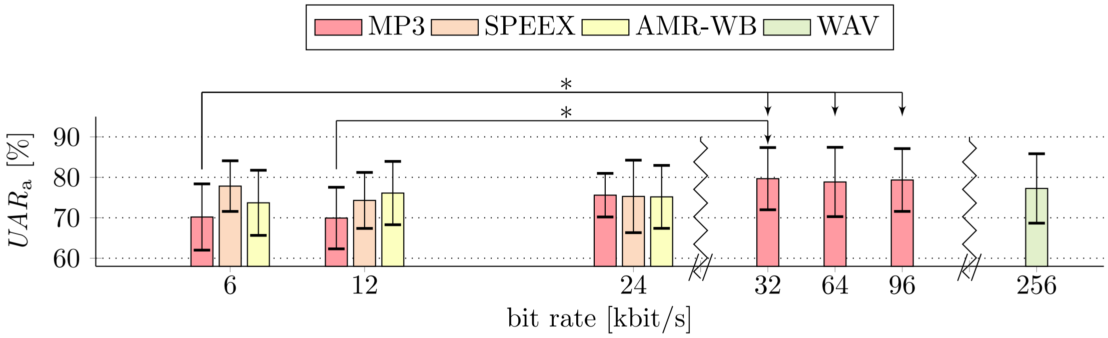


Figure 4.3: Mean and standard deviation of the UAR of the speech emotion recognition experiments. Stars denote level of significance ($\alpha = 0.05$) (cf. [Lotz et al. 2017]).

bit rate was noticed. This was not the case for SPX, where clearly the best recognition performance was present for a bit rate of 6.6 kbit/s. This codec configuration not only showed a comparatively high UAR but also the highest file size compression (2.76% of the original WAV). The worse results were obtained for MP3 with 8 and 16 kbit/s and the best for MP3 with 32 kbit/s.

Table 4.3 shows the results obtained for the individual emotions and codec configurations ($UAR_{\text{a,emotion}}$). Red and green entries indicate values outside of the standard deviation (above-green, below-red). The best results for each emotion are highlighted bold. A clear differentiation between good and bad performing codecs is visible. Especially for MP3 with low bit rates only low UARs were achieved. The codec configuration with 16 kbit/s even showed below average recognition results for six out of seven emotions. The best results were obtained when utilizing MP3 with higher bit rates (above 24 kbit/s). Above average results were obtained for three out of seven emotions for MP3 with 32 and 96 kbit/s. These results were not only above average but the best for four out of the seven evaluated emotion categories. The best results for fear and sadness were achieved by applying MP3 with 64 kbit/s. Only for the neutral state, the best result was obtained when utilizing SPX with a bit rate of 11.11 kbit/s. However, the results achieved for the

Table 4.3: $UAR_{\text{a,emotion}}$ for each considered codec configuration. Recognition results outside of the standard deviation of each emotion are highlighted in red (below) and green (above). The best results are denoted in bold font.

codec	WAV	MP3	MP3	MP3	MP3	MP3	MP3	SPX	SPX	SPX	AMR	AMR	AMR
bit-rate	256	8	16	24	32	64	96	6.6	11.11	22.09	6.6	12.65	23.85
anger	92.00	92.03	85.88	87.38	91.29	88.86	**92.77**	87.61	90.15	85.49	90.81	90.72	90.46
boredom	86.85	88.92	82.74	82.85	**90.81**	88.56	87.10	86.78	85.20	88.88	87.13	85.42	84.06
disgust	**53.57**	42.52	33.09	50.47	50.71	49.71	**53.57**	52.14	38.71	39.71	39.71	42.57	41.14
fear	80.92	57.49	80.33	80.58	84.58	**87.09**	86.53	84.63	73.60	83.18	84.87	86.52	85.74
joy	65.65	50.48	56.00	65.76	**70.21**	65.44	69.94	67.13	66.81	64.02	62.24	68.50	59.89
neutral	82.71	78.96	78.16	79.44	88.59	84.60	80.28	81.03	**88.77**	84.14	79.53	78.01	81.37
sadness	79.05	80.90	73.30	82.66	81.51	**87.66**	85.16	85.52	76.79	81.51	71.59	81.03	83.53

neutral state and MP3 with 32 kbit/s were only slightly below this value. It was further noticed that SPX with 6.6 kbit/s showed above average results for sadness and disgust. The results for all other states lay within the standard deviation. This was in line with the results presented in Figure 4.3, where this codec configuration achieved the 4th highest UAR_{a}. The AMR-WB codec showed no major benefit when applied for speech emotion recognition. For this codec a majority of the emotions and codec configurations showed an average recognition result. Surprisingly, almost none of the best results were achieved when utilizing the uncompressed WAV speech samples. For WAV only one emotion was detected with the highest recall, namely disgust. This emotion, however, showed the same recall for MP3 with 96 kbit/s. In contrast to the results of the human labeling, presented in the previous section, it was further possible to identify either increasing or decreasing influences (outside of the standard deviation) of the applied codec configurations on the UAR_{a}.

These results were explicable by the different application purposes and compression technologies of the codecs. The MP3 codec is specially designed for music compression and is based on perceptual coding. As one research field of music psychology focuses only on the relation between music and emotions [Juslin & Sloboda 2001], it can be assumed that the transmission of emotions for music is of high importance. From this assumption it could further be reasoned that not all information contained in the uncompressed WAV speech signal is necessary to automatically recognized the emotional state. Especially in the here presented approach, where only the prosodic information and no spoken content of the speech is evaluated. First investigations on this assumption were presented by Siegert and myself in [Siegert et al. 2018]. By applying the OPUS codec (cf. Appendix B), the successor of SPX, a hybrid coding based on both Analysis-by-Synthesis (AbS) and perceptual coding, to three benchmark speech emotion data sets, it was possible to achieve a remarkable performance increase of 1.66% to 4.47%. These results confirm that the differences in the speech signal caused by audio compression do not per se have a negative impact on the emotion recognition performance. To further evaluate this hypothesis, the results of the human labeling and the automatic speech emotion recognition, will now be compared to the speech/ audio quality to identify if there exist a correlation between these measures.

4.2.3 Correlation of Speech Quality and Speech Emotion Perception/ Recognition

To evaluate the correlation between speech quality and the ability of a human to perceive a certain emotion and automatic speech emotion recognition, respectively, two quality measures (MOS-LQO and CER, as presented in Section 2.4.1 and 4.1.2, respectively) were utilized. It is hypothesized that, related to the nature of the MOS and CER, representing subjective and objective speech quality measure, respectively, the subjective measure achieves a higher correlation with the results of the human

Table 4.4: $UAR_\mathrm{a}, UAR_\mathrm{h}$ and mean quality assessment measures (CER & MOS) over all evaluated subjects/ labelers. Brackets denote standard deviation.

codec	WAV	FLAC	MP3	MP3	MP3	MP3	MP3	MP3	SPX	SPX	SPX	AMR	AMR	AMR
bit-rate	256	–	8	16	24	32	64	96	6.6	11.11	22.09	6.6	12.65	23.85
UAR_h [%]	**94.39 (5.53)**	–	88.95 (7.17)	90.65 (5.34)	91.50 (9.05)	92.01 (5.74)	93.37 (6.80)	92.01 (5.74)	86.56 (8.77)	92.52 (5.90)	92.69 (6.09)	89.46 (6.71)	90.82 (7.00)	90.31 (7.54)
UAR_a [%]	77.25 (8.56)	77.25 (8.58)	70.19 (8.19)	69.93 (7.61)	75.59 (5.39)	**79.67 (7.69)**	78.85 (8.57)	79.34 (7.76)	77.83 (6.26)	74.29 (6.92)	75.25 (8.96)	73.70 (8.05)	76.11 (7.83)	75.17 (7.78)
CER_{dB} [dB]	0 (0)	0 (0)	11.91 (2.53)	8.85 (1.84)	7.32 (1.53)	4.78 (0.97)	3.25 (0.69)	2.76 (0.62)	6.03 (1.00)	5.49 (0.64)	4.72 (1.16)	6.11 (1.04)	5.23 (0.88)	4.86 (0.82)
$CER_{\%}$ [%]	0 (0)	0 (0)	-14.34 (3.23)	-8.09 (1.86)	-4.51 (1.24)	-1.57 (0.36)	-1.58 (0.30)	-1.69 (0.32)	-4.67 (1.38)	-2.53 (0.46)	-2.52 (0.44)	-5.06 (0.92)	-3.31 (0.93)	-2.08 (0.68)
MOS-LQO	4.59 (0.18)	4.59 (0.18)	1.68 (0.18)	2.82 (0.30)	3.51 (0.29)	4.04 (0.27)	4.55 (0.17)	**4.61 (0.17)**	2.07 (0.39)	2.96 (0.49)	3.91 (0.57)	2.71 (0.57)	3.51 (0.67)	3.79 (0.63)

listening test and the objective measure with the automatic emotion recognition results. Table 4.4 provides all relevant quality measures and recognition results of the listening experiment and the automatic speech emotion recognition task for all considered codec configurations. The CER_{dB} and $CER_{\%}$ were calculated as described in the previous Section, by applying Equations (4.10) and (4.12).

To determine the correlation between the different measures, Spearman's rank correlation coefficient (R_s) was calculated (cf. [Spearman 1904]). This coefficient can be utilized to determine the correlation between two ordinal and/ or metric scaled measures. Negative values indicate an inverse dependency, values from 0 to 0.2 a none to poor, 0.2 to 0.5 a weak to moderate, 0.5 to 0.8 a clear, and values above 0.8 a high to perfect correlation. To apply R_s, the ranks of the measures are needed, this was done by conducting a top-down ranking, ranking the best value with the highest rank "1". Depending on the evaluated measure, the best value corresponds to the highest or the lowest value (e.g. best value: $UAR_\mathrm{h} = 94.39$, $CER_{\%} = 0$).

First, a comparison of the quality measures presented in Table 4.4 was realized. This resulted in a high correlation of $R_{s(MOS\text{-}LQO/CER_\mathrm{dB})} = 0.9133$, $R_{s(MOS\text{-}LQO/CER_{\%})} = 0.9161$ and $R_{s(CER_\mathrm{dB}/CER_{\%})} = 0.9286$. Afterwards, the correlation of the results obtained from the listening experiment and the automatic speech emotion recognition task was determined ($R_{s(UAR_\mathrm{h}/UAR_\mathrm{a})} = 0.4154$), which corresponds to a moderate correlation. From these results it was concluded that there exist only small differences in the quality assessment but rather big differences in the performance assessment, respectively. It could be assumed that the differences in the performance assessment are related to different quality aspects, e.g. listening quality and spectral quality. From speech emotion recognition it is known that especially the spectral information of the speech signal has a high impact on the performance of a recognizer (cf. Section 2.2.2). Therefore, it was assumed

Table 4.5: Spearman's rank correlation coefficient determined to calculate the correlation between the performance measures (UAR_h and UAR_a) and quality measures (*MOS-LQO*, CER_dB and $CER_\%$).

	$R_{s(x/y)}$	x UAR_h	 UAR_a
	MOS-LQO	**0.7989**	0.6850
y	CER_dB	0.7510	0.7088
	$CER_\%$	0.7758	**0.7363**

that the CER is more relevant when it comes to speech emotion recognition as it is based on the spectral difference of the signals. To verify this assumption the correlations between the quality measures and the performance measures were determined, resulting in the R_s values presented in Table 4.5. These values indicated a clear correlation between the quality measures and the performance measures. Furthermore, the MOS-LQO showed a higher correlation with the results of the human listening experiment versus both CER measures and vice versa for the automatic speech emotion recognition results. This confirmed the previously made assumption, but it should also be noted that the differences between the R_s values were comparatively low, which was expectable because of the high correlation obtained in between the different quality measures.

Comparing the performance results of the human labeling with the automatic emotion recognition, it was noticed that the human labeling achieved an overall higher recall than the automatic emotion recognition. This was expectable, as the human auditory system is able to distinguish emotions in a much higher resolution and the listener has gained experience on how certain emotions are transmitted between humans, while the automatic emotion recognition is solely based on the data it was trained on. Especially for SPX with a bit rate of 6.6 kbit/s, a distinct difference in the performance of human and automatic emotion recognition was noticed. For this codec configuration the results of the human labeling showed the lowest UAR_h with the highest standard deviation. Surprisingly, in case of the automatic emotion recognition, this configuration showed a higher UAR_a than the one obtained for the uncompressed WAV codec with one of the lowest standard deviations. This reverse behavior was also noticed when comparing the results obtained for each emotion over all codecs of the two recognition tasks with each other (e.g. row *anger* in Table 4.2 vs. Table 4.3). By determining the correlation coefficient for each considered emotion, an average $\overline{R}_{s(\mathrm{emotion})}$ value of -0.05 (0.10) was reached. The lowest value was obtained for anger ($R_{s(\mathrm{anger})} = -0.25$), indicating an anti-correlation. For all other emotions no correlation was shown. Furthermore, it was determined if there existed a correlation between the ability to perceive/ recognize a certain emotion within a codec (e.g. column *MP3-8* in Table 4.2 vs. Table 4.3). Depending on the considered audio codec, the R_s value showed a different behavior: $R_{s(\mathrm{WAV})} = 0.25$, $\overline{R}_{s(\mathrm{MP3})} = 0.52$ (0.16), $\overline{R}_{s(\mathrm{SPX})} = 0.30$ (0.36) and $\overline{R}_{s(\mathrm{AMR\text{-}WB})} =$ -0.03 (0.03). It was noticed that especially for AMR-WB no correlation between the results of the human labeling and the automatic recognition was present. For MP3 a clear correlation of above 0.56 was reached for 8, 24, 32 and 64 kbit¡ For 16 and 96 kbit/s only a moderate correlation was achieved. This, on average, clear correlation of MP3 was explicable by the codec's compression algorithm, as mentioned earlier in this Section. For SPX the widest range of the R_s values was observed, ranging from 0.05 with 6.6 kbit/s to 0.71 with 11.11 kbit/s. This is in line with the previously observed reverse behavior for SPX with 6.6 kbit/s.

4.2.4 Findings and Recommendations on Compressed Speech

In this Section it was shown that the application of the audio compression led to significant differences in the results obtained by applying a speech emotion recognizer. For human speech emotion perception, the best UAR was achieved when utilizing the uncompressed WAV speech samples. For the speech emotion recognizer, this was not the case. Here, the best result was obtained when utilizing MP3 with higher bit rates (over 24 kbit/s). In case of a high file size compression, the SPX codec with 6.6 kbit/s is recommended. The results obtained for this codec configuration showed comparable high recognition results while achieving the highest file size reduction (2.76 % of the original WAV format). Surprisingly, the human ability to correctly identify the emotional state for this configuration was the lowest, compared to all the other codec configurations. From an emotion wise comparison of the audio codecs it was shown that there exists a clear correlation between the recall of human emotion perception and automatic emotion recognition for the MP3 codec configurations. This lies in the nature of the MP3 codec, which is specially designed for music compression. It is based on perceptual coding which is closely related to the human auditory system. From music psychology it is well-known that there exists a relation between music and emotions. Therefore, music compression not only keeps the audio quality high, but also maintains a good emotion perception. It was further investigated, if there exists a correlation between speech quality and speech emotion perception/ recognition. It was shown that there exists a high correlation between the speech quality measure MOS-LQO and the newly introduced CER measures. Furthermore, it was shown that there exists only a moderate correlation between the results of the human labeling and the speech emotion recognizer. From a theoretical perspective it was assessed, that the MOS-LQO would show a higher correlation with the ability of humans to perceive a certain emotion and the CER with the performance of a speech emotion recognizer. Because of the high correlation between the quality measures it was expected that the differences between these correlation coefficients would be comparatively low. This was confirmed by the results stated in Table 4.5, where a clear correlation for all measures was obtained.

It can be concluded that there exists a clear correlation between speech quality and speech emotion perception/ recognition. However, there exist exceptions for which this is not the case. For human emotion perception a clear deviation between the listening quality (*MOS-LQO*) and the human recall of emotions (UAR_{h}) was shown for MP3 with 96 kbit/s, SPX with 11.11 kbit/s and AMR-WB with 23.85 kbit/s. Even though the *MOS-LQO* indicated a good listening experience for the MP3 and AMR-WB configurations, the UAR_{h} only reached comparatively low values and vice versa for the SPX configuration. For the performance of the speech emotion recognizer a clear deviation between the speech signal quality (*CER* [%]) and the recall of the automatic emotion recognition system (UAR_{a}) was noticed for SPX with 6.6 kbit/s. This codec configuration showed a low signal quality but achieved a comparatively high recognition performance.

Nevertheless, it needs to be stated that the results presented in this Section are only based on the EmoDB data set, which contains highly expressive acted emotions. Especially for in the wild emotion recognition, where these kind of emotions occur less frequent and with low expressiveness, an application of audio compression may influence the recognition performance in a different manner. To get a more general insight on the effect of audio compression on speech emotion perception and recognition, the experiments need to be repeated on more natural speech emotion data sets. First investigations by Siegert and myself already showed, when utilizing MP3, AMR-WB and the successor of SPX, OPUS on the Danish Emotional Speech Database, that the recognition performances have similar tendencies as reported in this Chapter for the EmoDB data set [Siegert et al. 2017]. In [Siegert et al. 2018] we could even show that it is possible to improve the recognition performance by utilizing the OPUS codec in Constrained Energy Lapped Transform (CELT) mode (only based on perceptual coding). Furthermore, it would be interesting to evaluate the effect of audio compression on the feature space. This could be done by utilizing Wilcoxon signed-rank test (cf. Section C.2), similar to the approach presented in Section 5.2.2. This test can be used to determine if there exists a significant difference between a feature in the compressed and uncompressed state. It is not recommended to use the Pearson correlation coefficient, as it only identifies if there exists a linear relation between the feature values. This, however, is also the case when a feature shows a consistent increase/ decrease, which could still represent a significant alternation in the feature values. An evaluation of these differences was not done in the scope of this Thesis, but can be seen as potential future work.

4.3 Exploring Noisy Speech

In the previous Section I could show that there exists a clear correlation between the speech quality and the results of an automatic speech emotion recognizer. As not only compression leads to a degradation of the speech signal but also other disturbances like environmental noises, it was assumed that the evaluation approach presented for compressed speech can be adapted to be used for the analysis of disturbed speech, with a special focus on in-vehicle noises. This was done by utilizing two quality measures, the SNR and the CER. In the previous Section it was already shown that for emotion recognition the highest correlation between speech emotion recognition performance and speech quality was achieved by applying the CER [%]. Therefore, the CER [dB] and MOS-LQO were not considered for the investigations presented now. Furthermore, from literature it is known that there exists a high correlation between the SNR and MOS (cf. Section 2.4.2). Therefore, it can be assumed, that both quality measures can be used equally. As noisy speech data the EmoDB-Car and VAM-Car data samples, as presented in Section 3.1, were employed. By utilizing re-recordings of well-known emotional speech data sets, a reference clean version of the signals was available. Furthermore, the utilized data sets comprise different emotion naturalness and recording qualities. These are acted emotions recorded in an anechoic chamber (EmoDB) and scripted emotions recorded in a television studio setting (VAM), respectively. This made it possible to also evaluate the influence of the original recording quality on the speech quality and the emotion recognition performance of the re-recordings.

4.3.1 Quality Assessment of Emotional In-Car Speech Data

To evaluate the speech quality the SNR and the CER [%] measures, as introduced in Equation (4.3) and (4.12), were calculated. The original speech samples of the data sets were used as reference clean speech samples of the re-recorded noisy speech samples. For the re-recording two shotgun microphones were integrated into a driving simulator, mounted onto a strut profile at the left and right A-pillar of the chassis (cf. Figure 3.1 on page 86). Therefore, there also exist two recordings per recording condition. This resulted in eight experiments, which were carried out to evaluate the speech quality of in-car emotional speech data (cf. Table 4.6). As the clean speech samples were recorded under different recording conditions than the re-recordings (i.e. different room acoustics and microphone setups), the resulting SNR is not comparable to other SNRs presented in the literature and will be referred to as *relative SNR* (cf. Section 4.1.1). For the different recording conditions it was assumed that the recordings obtained under silence condition were only influenced by the in-vehicle acoustics and the recording setup, while for the recordings under disturbed condition an additional environmental noise was present. For the CER this would result in a negative CER_{dB} under silence condition, which would later increase under disturbed condition. A negative CER is possible for both, silence and disturbed conditions, as the reduction of the spectral power caused by the in-vehicle acoustics and recording setup may overrun the increases caused by the in-vehicle noises. Furthermore, it was assumed that the recordings of the left microphone,

Table 4.6: Carried out experiments to evaluate the quality of the noisy speech data of the EmoDB-Car and VAM-Car data sets under silence and disturbed recording conditions.

	Experiment	Recording Condition	Microphone
EmoDB-Car	(1)	silence	left
	(2)		right
	(3)	disturbed	left
	(4)		right
VAM-Car	(5)	silence	left
	(6)		right
	(7)	disturbed	left
	(8)		right

which was located closer to the loudspeaker, would receive better speech quality compared to the recordings of the right microphone.

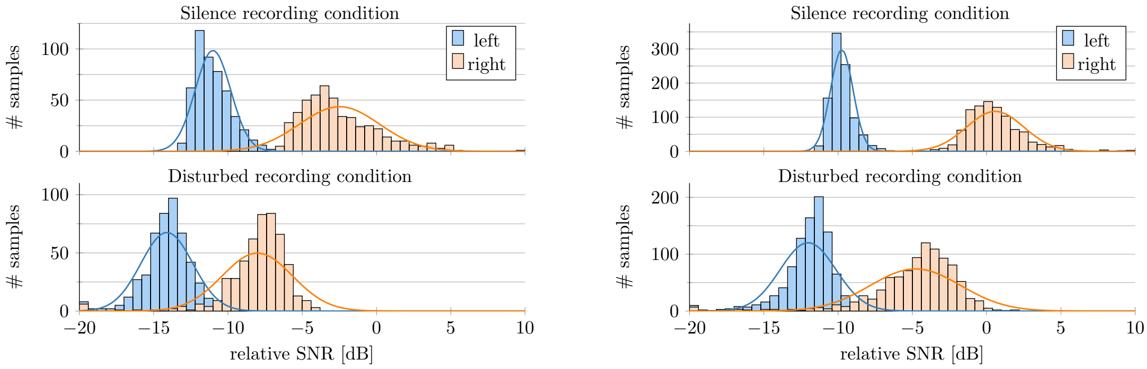
An overview on the results obtained for the quality assessment is presented in Table 4.7. The mean and standard deviation of the *relative SNR* and the $CER_{\%}$ are stated for each experiment. A detailed distribution of the quality measures is given in Figure 4.4 and 4.5 for both data sets, recording conditions and microphone setups. A repeated-measures ANOVA conducted for all data sets revealed that there exists a highly significant difference for both quality measures regarding the recording condition and microphone setup (all p's < 0.001, Greenhouse-Geisser-corrected). By conducting post-hoc t-tests it was shown, that the differences between the recording conditions and microphone setups were always highly significantly different (all p's $<$ 0.001, Bonferroni-corrected). Furthermore, it was noticed that the obtained quality measures of the right microphone setup showed a higher standard deviation than the left microphone setup. This was expectable, as the right microphone was located further away from the loudspeaker.

In Figure 4.4 it can be seen that for the *relative SNR* the left microphone always achieved lower values than the right microphone, contrarily to the assumption made

Table 4.7: Mean and standard deviation of the quality measures SNR and CER [%] for the eight carried out experiments averaged over all samples included in the corresponding data set.

	EmoDB-Car			
Experiment	(1)	(2)	(3)	(4)
relative SNR	-11.01 (1.20)	-2.49 (2.71)	-14.11 (1.75)	-8.02 (2.37)
$CER_{\%}$ [%]	-17.63 (1.54)	-25.75 (2.28)	-13.93 (2.44)	-21.15 (3.35)

	VAM-Car			
Experiment	(5)	(6)	(7)	(8)
relative SNR	-9.76 (0.76)	0.58 (1.93)	-12.02 (1.88)	-4.79 (3.05)
$CER_{\%}$ [%]	-6.84 (0.84)	-15.07 (1.48)	-4.62 (1.76)	-12.19 (2.70)

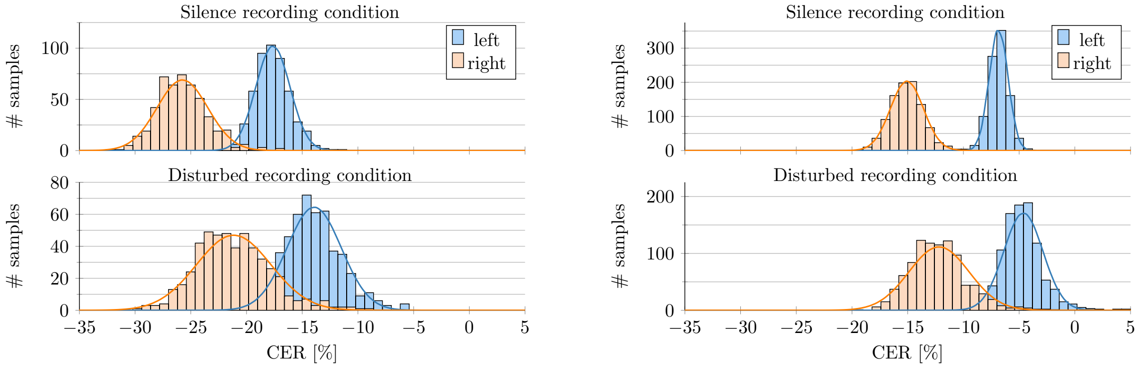


(a) relative SNR of the EmoDB-Car dataset

(b) relative SNR of the VAM-Car dataset

Figure 4.4: Histograms and fitted normal distributions of the *relative SNR* distribution of the EmoDB-Car and VAM-Car data samples under silence and disturbed recording conditions obtained for the left and right microphone location.

previously. As the left microphone was located closer to the loudspeaker, the signal was recorded with a higher loudness resulting in a higher power of the noisy speech signal P_{ns} compared to the recording of the right microphone. With regard to Equation (4.3) and applying one constant α for the left and right microphone setup, this resulted in a higher denominator for the left microphone. As the results were based on the same clean reference recording, an adaption of the α values for both microphones was not recommended. This would abolish the relation between the two recordings. It was further noticed, that the results obtained under disturbed recording conditions showed lower *relative SNR* values than the ones obtained under silence recording conditions. This was in line with the assumption that the disturbed recordings contain more noise than the silent recordings.



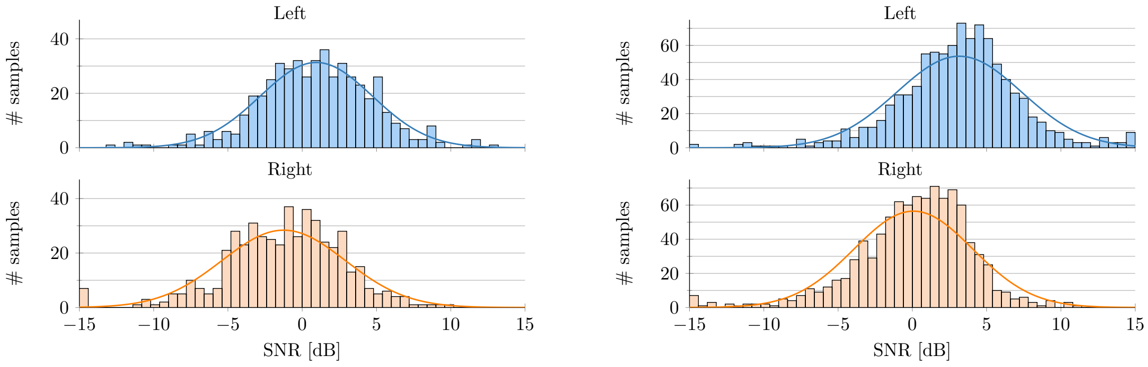
(a) CER of the EmoDB-Car dataset

(b) CER of the VAM-Car dataset

Figure 4.5: Histograms and fitted normal distributions of the $CER_{\%}$ distribution of the EmoDB-Car and VAM-Car data samples under silence and disturbed recording conditions obtained for the left and right microphone location.

The results of the CER can be taken from Figure 4.5 on page 128. Here, a contrary behavior compared to the results obtained for the relative SNR was observed. In case of the microphone setup, the left microphone always showed a higher CER than the right microphone. Considering the recording condition, the CER under disturbed condition always achieved higher values than under silence condition. The second observation, however, lies in the nature of the CER, as it describes the percentage difference of the spectral power. Whenever noise is added to a signal, the spectral power will increase, while for the recording under silence condition, the recording setup and in-vehicle acoustics suppress the spectral power. As the disturbed condition contains both, suppression by in-vehicle acoustics and additive noise, an increase compared to the results obtained under silence condition is possible. It was further noticed that for both quality measures, the results obtained for the VAM-Car data samples were higher than the results of the EmoDB-Car samples. This was explicable by the differences in the original recording quality of the two utilized data sets. The EmoDB samples were recorded in an anechoic chamber with no ambient noise being present, while the VAM samples were recorded in a television studio with audience and other interference factors. This made the EmoDB samples more prone to the in-vehicle acoustics and the environmental background noises of the simulator.

Because of the mentioned restrictions and difficulties occurring when applying the *relative SNR* (cf. Sections 4.1.1 and previously in this Section), it was assumed that this measure is rather unsuitable when it comes to the assessment of speech quality under non-ideal recording conditions. Nevertheless, a robust SNR, comparable to results presented in literature, could be obtained when the reference signal is recorded using a similar recording setup as the noisy counterpart. This was the case for the re-recordings under silence and disturbed recording conditions. Assuming that the re-recording correspond to the reference and noisy speech signal, respectively,



(a) SNR of the EmoDB-Car dataset

(b) SNR of the VAM-Car dataset

Figure 4.6: Histograms and fitted normal distributions of the SNR distribution of the EmoDB-Car and VAM-Car data samples.

the *real SNR* was calculated by applying Equation (4.2). This resulted in an average SNR of 0.95 (3.77) and -1.25 (4.16) for the left and right microphone of the EmoDB-Car samples and 1.68 (3.18) and -0.61 (3.60) for the VAM-Car samples, respectively. The detailed distribution of the SNR values for the left and right microphone of both data sets can be taken from Figure 4.6. As assumed, the left microphone setup outperforms the right microphone setup. Contrarily to the *relative SNR* and the CER, however, it is not possible anymore to give a statement on the speech quality compared to the original EmoDB and VAM data samples, as the original samples were not considered in the calculation of the *real SNR*.

4.3.2 Speech Emotion Recognition in In-Car Environments

Similarly to the experiments carried out to evaluate the speech quality, classification experiments were carried out to evaluate the performance of a speech emotion recognizer on noisy speech data. The classification experiments were realized by utilizing a SVM with a linear kernel ($C = 1$) using the software tool WEKA [Hall et al. 2009]. As feature set the *emobase* set was used. The classifiers were trained and tested by applying the LOSO validation scheme and the results were averaged over all speakers of the corresponding data set. Training and testing of the classifiers were performed on samples originating from the same recording setup only. Furthermore, additional baseline experiments were realized by applying the original samples of the EmoDB and VAM data sets.

The mean and standard deviations of the UAR for all performed classification experiments are stated in Figure 4.7. It was noticed, that the recognition results obtained for the EmoDB data samples achieved higher values than the ones obtained for the VAM samples. This was on the one hand explicable by the different emotion naturalness of the data, and on the other hand also by the different recording qualities. With an increasing level of naturalness, of both the uttered emotions and the recording setup, the ability to automatically detect the emotional state of a speaker decreases dramatically (cf. Section 2.1.1). This also explained the high standard deviations in the results obtained for the VAM data samples. It was further noticed that the differences in the results obtained under silence and disturbed recording conditions were comparatively higher for EmoDB than for VAM. This was attributed to the fact that the naturalistic VAM data samples were recorded under non-ideal recording conditions with perturbing acoustic conditions. Therefore, the in-car setting had little further disturbing effects on the recognition performance. By conducting repeated-measures ANOVAs for both data sets, it was shown that there exists no significant differences in the results obtained for the different recording conditions and baseline classification experiments (main effect EmoDB: F(0.17, 1.55) = 6.14, $p > 0.05$, Greenhouse-Geisser-corrected; main effect VAM: F(1.53, 70.41) = 1.98, $p > 0.05$, Greenhouse-Geisser-corrected). However, the results obtained under disturbed condition for the left and right microphone recordings of both evaluated

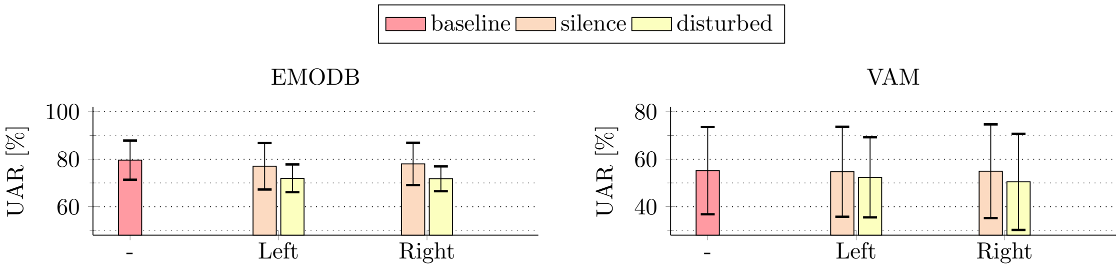


Figure 4.7: Mean and standard deviation of the UAR of the carried out LOSO classifiaction experiments of both utilized data sets (cf. [Lotz; Faller et al. 2018]).

data sets were noticeably lower compared to the baseline results. Furthermore, it was noticed that the baseline results of both data sets always outperform the results obtained under the different recording conditions for both microphone setups. This is reasonable, as the speech quality assessment, presented in Section 4.3.1, revealed a significant difference in the speech quality of the silent and disturbed recordings for all experiments, with the disturbed speech samples being of lower speech quality than the samples recorded under silent condition. It can therefore be assumed that speech quality affects the emotion recognition performance. However, the results obtained for the left and right microphones for each recording condition showed a high similarity, indicating that for speech emotion recognition the recording condition has a higher influence on the recognition performance and is, hence, of higher relevance than the microphone setup.

4.3.3 Findings on Noisy Speech

In this Section, I evaluated the effect of in-vehicle noises on the speech quality of emotional speech and the performance of a speech emotion recognizer. To evaluate the speech quality two quality measures were utilized ($CER_{\%}$ and *relative SNR*). It was shown that both measures have advantages and disadvantages when it comes to the assessment of the speech quality. For the *relative SNR*, especially the quality differences between the recordings of the left and right microphone were not assessed correctly. This was due to the restrictions and adaptations made in Section 4.1.1. To solve this issue, an adaptation of the constant α in Equation (4.3) for the speech samples originating from the left and right microphone would be needed. This, however, would abolish the relation to the original EmoDB and VAM data samples used as baseline for the calculation. Furthermore, the obtained results were not comparable to other SNRs presented in the literature. By utilizing the re-recording under silence and disturbed conditions, I was able to calculate the *real SNR* of the simulator recording setup. The obtained values lie in the region of comparable SNRs presented in Section 2.4.2. However, a statement on the speech quality compared to the original EmoDB and VAM data samples is now not possible anymore.

Overall, it can be concluded that the SNR, as introduced in Equation (4.3), is unsuitable for the assessment of speech quality when it comes to re-recordings under non-ideal recording conditions. For the CER it can be stated that, with regard to the assumptions made at the beginning of this Section (i.e. left microphone and silence recordings are of higher speech quality than the right microphone and disturbed recordings, respectively), the measure behaves as expected. However, it is of high importance to understand how the CER is designed, as it calculates the percentage difference compared to the reference signal. This could lead to a confusion in the interpretation of the measure, as the quality of the samples under disturbed recording conditions showed a higher percentage agreement compared to the recordings under silence condition. This is valid, as the additive vehicle and environmental noises included in the disturbed recordings lead to an increase of signal power compared to the samples recorded under silence condition.

By utilizing two different speech emotion data sets of different recording naturalness (i.e. acted vs. scripted emotions and anechoic chamber vs. television studio recording setup), I was able to show that the recording naturalness has a high impact on the utilized quality measures. For the VAM data samples, originally recorded under non-ideal recording conditions, the re-recording under silence and disturbed recording conditions were less influenced by the additional disturbances compared to the EmoDB samples. This was confirmed by the utilized quality measures which indicated an on average higher speech quality of the VAM-Car samples compared to the EmoDB-Car samples, with respect to the original VAM and EmoDB samples.

It was further investigated, if there exists a relation between the speech quality and the performance of a speech emotion recognizer. For both investigated data sets a clear decrease in the performance of the re-recordings was present. Especially the recordings under disturbed recording condition showed considerably lower UARs compared to the baseline results obtained by utilizing the original EmoDB and VAM data samples. This is in line with the hypothesis that the disturbed recordings are of lower speech quality and that there exists a relation between speech quality and recognition performance. For the microphone setup this, however, was not the case. Here, a significant difference in speech quality between the left and right microphone recordings was identified. The results of the corresponding emotion recognition experiments, however, did not reveal any significant difference. This indicates that, in case of speech emotion recognition in noisy environments, the recording/ noise conditions are of higher relevance than the microphone setup. Furthermore, the quality decrease from original to re-recorded speech was much higher for the (originally anechoic chamber-recorded, acted speech) EmoDB data set than for the (originally TV- studio-recorded, scripted speech) VAM data set. This also resulted in a higher performance decrease for the automatic emotion recognition task in case of the EmoDB, respectively.

4.4 Summary and Discussion

This Chapter highlights the effects of speech quality on the ability to perceive the emotional content by humans from speech and in an automatic application domain (i.e. speech emotion recognition), with a focus on compressed and disturbed emotional speech. As the number of available noisy speech data obtained in an in-vehicle environment is still limited and the realization of a reliable data set is highly time and resource consuming, first investigations were carried out by utilizing compressed speech samples of the well-known EmoDB data set. To get a broad insight on audio compression, a short overview on relevant compression techniques and audio codecs was given in Section 2.5.3 and Appendix B. Furthermore, measures to assess the quality of compressed and disturbed speech were introduced. As most of the common measures are highly limited in their application, a new measure, the CER, was introduced. This measure determines the average difference of two speech samples under different recording conditions as actual change ([dB]) or as relative change ([%]) compared to the original clean recording.

Initially, the effect of compressed speech was investigated. Three lossy audio codecs, intended to be used in different application domains and following different compression techniques, were utilized (i.e. MP3 - music - perceptual coding, SPX - VoIP - AbS, AMR-WB - VoLTE - AbS). As reference the standard WAV format was used. I was able to show that audio compression strongly influences both, the ability to perceive the emotional content by humans and to automatically recognize the emotional state of the speaker. For human emotion perception it was shown that the best emotion intelligibility was present for the reference high quality WAV format. This was not the case for the speech emotion recognizer, where the best results were obtained for MP3 with high bit rates (over 24 kbit/s). When it comes to a high file size reduction, the best results for the speech emotion recognizer were achieved by SPX with 6.6 kbit/s. Contrarily, this codec achieved the worst results when it came to human speech emotion perception. It was further noticed that the results obtained from the human labeling and automatic emotion recognition for each considered emotion were clearly rank correlated when applying the MP3 codec. Overall, the correlation between the human labeling and automatic emotion recognition was only moderate. This is explicable by the compression technique used for the MP3 codec, which is designed especially for music compression. From music psychology it is well-known that there exists a relation between music and emotions. Therefore, music compression not only keeps the audio quality high, but also maintains a good emotion perception. The introduced quality measures themselves showed a high correlation among each other. I could identify that the MOS-LQO measure achieved a higher correlation with the results obtained from the human labeling than the results of the automatic emotion recognition. A contrary correlation was shown for the $\mathrm{CER}_{\%}$, respectively. This is in line with the hypothesis made at the beginning of the corresponding Section that subjective measures better describe the ability of a human to perceive a certain emotion than objective measures and vice versa in case of speech emotion recognition.

With this prior knowledge, the $\text{CER}_{\%}$ was applied to investigate the performance of a speech emotion recognizer on disturbed/ noisy speech samples in an in-vehicle environment. As reference well-known quality measure for disturbed speech, an adapted version of the SNR was utilized. As data sets, two benchmark data sets re-recorded in an in-vehicle simulator environment were employed (EmoDB-Car and VAM-Car). It was shown that there exists a significant difference for both quality measures depending on the recording condition (silence and disturbed) and microphone setup (left and right). Also the recording setup of the original data set influenced the quality of the re-recordings. While the quality of the samples originating from the EmoDB data set (acted emotions in anechoic chamber) were strongly influenced by the recording conditions, the samples originating from the VAM data set (scripted emotions in television studio) showed a lower loss in quality, respectively. This finding corresponds to the observed differences in the results of the performed speech emotion recognition experiments. The results obtained for the EmoDB-Car data samples show a noticeably stronger decrease in the UAR compared to the baseline results than the results obtained for the VAM-Car data samples.

Overall it can be concluded that there exists a relation between speech quality and the ability to perceive and automatically recognize the emotional state of a speaker. This can be done by utilizing well-known quality measures but also by utilizing the newly introduced CER measures. Furthermore, it should be noted that the presented results are based on benchmark data sets, commonly used in the field of speech emotion recognition. By utilizing the VAM data set I was able to give a first insight on how in-vehicle noises influence emotional speech with a more realistic non-ideal recording setup. Nevertheless, especially in noisy environments other effects can occur that influence the way of speaking, which were left out of consideration in this Thesis. With a focus on in-vehicle communication the reader can refer to [Landgraf 2018]. However, it can be assumed that these effects do not influence the speech quality but rather the ability to perceive and automatically recognize the emotional state of the speaker. This aspect will be addressed in Chapter 6 where naturalistic real-world in-vehicle speech data is employed. The next Chapter will focus on the pre-processing of the raw audio material.

Chapter 5

Processing of In-Car Real-World Audio Data

Contents

UP to this point, the focus of this Thesis was drawn on the collection of simulated and real-world speech data in Chapter 3, and the influence of speech quality on speech emotion recognition in Chapter 4. It was already state in the introduction of this Thesis, that the performance of a machine learning based classifier is strongly affected by the quality of the data they are trained on (also see [Marsland 2015] and Section 2.2.6). Most of the factors affecting the quality of the utilized data set can be influenced in advance by an extensive experimental study design as highlighted in Section 3.2. Here, it was already shown that the collected real-world audio recordings are suitable for the present emotion recognition task. Other factors, such as inter-individual differences of the subjects (i.e. is the emotional state also reflected in the drivers speech) or restricting environmental conditions, need to be evaluated afterwards. The former is necessary to assess the trustworthiness of the data samples, while the later can contribute to the performance of the subsequent classification task or simplify the design of the classifier. Therefore, before designing a classifier to automatically recognize the driver's emotional state (cf. Chapter 6), a pre-processing of the raw audio material in necessary. This will on the one hand include the annotation of the real-world in-car audio recordings presented in Chapter 3 and on the other hand the evaluation of a speech enhancement method and its applicability in speech emotion recognition.

First, a detailed description of the annotation process and its results, which are based on [Lotz; Ihme et al. 2018] and [Requardt et al. 2018], will be given in Section 5.1. The pre-processing of the raw audio recordings of the real-world in-vehicle data collection, presented in Section 3.2, will be presented in Section 5.1.1. In Section 5.1.2, additionally to the results obtained when applying a standard manual labeling process, a machine-learning-assisted emotion labeling approach will be presented, which enables the full annotation of an emotion data set in noticeably less time compared to a conventional fully manual annotation. The presented approach is based on the semi-automatic annotation approach developed by Egorow and myself in [Egorow et al. 2017]. As the annotation of the data was performed on clean reference recordings of the driver's speech only, a post-annotation processing of the annotated data samples based on the noisy shotgun microphone recordings will be presented in Section 5.1.3.

Second, in Section 5.2, the suitability and necessity of a pre-processing of the raw audio material in terms of speech enhancement will be investigated. The main goal of speech enhancement is to improve the speech intelligibility of the disturbed speech signal by applying different digital signal processing algorithms in time and frequency domain [Benesty et al. 2009]. This, however, does not automatically lead to a better understanding of the emotional content, as they are primary optimized to increase the perceptual quality of degraded speech with disregard of the changes occurring in the speech signal. From literature it is known that, especially, for speech emotion recognition spectral-based speech features are of high relevance [El Ayadi et al. 2011; Eyben et al. 2016]. I will first give some basic theoretical background on how the selected speech enhancement algorithm operates (cf. Section 5.2.1). Afterwards, in Section 5.2.2, it will be analyzed how the application of a conventional speech enhancement algorithm modifies the original speech signal (in time and frequency domain and feature space). It will be further evaluated in Section 5.2.3 whether positive effects occur when executing emotion recognition tasks on the enhanced speech signal. The Chapter will be concluded in Section 5.2.4 by giving a recommendation on the application of speech enhancement for speech emotion recognition tasks.

As parts of this chapter are based on work already published in [Requardt; Egorow et al. 2020; Lotz; Ihme et al. 2018; Requardt et al. 2018] and [Egorow et al. 2017], several phrasings are taken literally from these publications.

5.1 Annotation

To obtain a ground truth of the real-world in-car recordings presented in Section 3.2 a manual annotation by trained labelers was performed. In total three independent, German speaking, female labelers in between the age of 20 to 35 were employed. They all underwent a pre-training where they were introduced to the annotation software *ikannotate2* (cf. [Siegert & Wendemuth 2017]) and had to conduct 20 test-annotations of speech samples originating from the Vera am Mittag (VAM) data set [Grimm et al. 2008]. This data set was chosen, as it contains spontaneous speech samples of non-professional German speakers under non-optimal recording conditions, which is comparable to the recording quality of the data collection presented. All samples in the VAM data set are provided with a corresponding emotion label in the dimensions of valence, arousal and dominance by the corpus developers, and a mapping onto the four quadrants of the two-dimensional valence-arousal space was conducted by [Schuller; Vlasenko; Eyben et al. 2009]. These mapped results served as ground truth and were used to verify the assessment of the test-annotations. If one of the labelers showed strong deviations from the ground truth they were given a more detailed instruction on how to assign the dimensions of valence and arousal and additional 20 test-annotations needed to be conducted by them. The test-samples were chosen such that they would cover all quadrants of the valence-arousal space equally in a randomized order.

For the actual annotation process, only the recordings obtained by the high quality headset microphone were used. Compared to the shotgun microphone recordings, the headset recordings contain considerably less noise, as the microphone inlet was directed towards the driver's mouth with a cardioid directional pattern suppressing sound coming from other directions. To increase the quality of the annotation results and to terminate the annotation effort by excluding unsuitable speech material, a pre-processing of the raw audio signal was necessary. The utilized *ikannotate* annotation software does not provide this kind of internal audio processing (i.e. voice activity detection and splitting of the speech material into suitable smaller sub-samples). Therefore, an additional pre-processing of the raw-audio material was performed before conducting the emotion labeling (cf. Section 5.1.1). All audio samples generated during the pre-processing were then annotated in a three step procedure:

1. Annotation of the dimensions of valence (from negative to positive) and arousal (from low/ calm to high/ aroused) using the 5-point Self Assessment Manikins (SAM) scale [Bradley & Lang 1994] (for more details see Section 2.1.4). This scale was also used as self-report measure in the detailed feedback form, which was filled out by the drivers after finishing all four driving scenarios (cf. Section 3.2).

2. Annotation of the four emotion categories: *positive*, *neutral*, *frustrated/ angry* and *anxious/ fearful*, which correspond to the induced emotional states in the four driving scenarios. Additionally, a free text input was available for the labelers to allow the annotation of a self-chosen emotion category in case one

of the predefined category did not match. This option, however, was never applied by the labelers.

3. Rating of the labelers' level of satisfaction of their own current labeling using a 5-point Likert-scale from 1 (very dissatisfied) to 5 (very satisfied). This step was included to get a direct subjective feedback on the quality of the annotation and conclude the reliance of their decision.

Another aspect which needs to be addressed is the reliability of the annotation results regarding the agreement of the three labelers. Therefore, the Inter-Rater-Reliability (IRR) utilizing Krippendorff's alpha was determined (cf. Section 2.1.4 and Section 5.1.2). The IRR is a measure to assess the agreement between the annotation results of two or more labelers [Hallgren 2012; Krippendorff 2004]. Depending on the data properties (i.e. nominal, ordinal or interval scaled data) the distance measure $\boldsymbol{\delta}_{c_a,c_b}$, defined as the distance between the assigned class c of labeler a and b, needs to be determined. When utilizing the SAM scale an ordinal distance measure is applied, while for the evaluation of categorial annotations the application of a nominal distance measure is required. Labelers lowering the IRR by 0.05 were excluded before assigning a certain label to the considered speech samples, as these small deviations in the IRR can already affect the annotation results in a noticeable way. This made the label assignment more conservative.

For the categorial labeling, the labels were assigned based on a majority voting of the labelers' annotations. When a labeler was excluded from the assignment process, the remaining labelers had to be fully conform in their annotation result. All samples for which no distinct assignment of a categorial label was possible, were excluded from the ground truth obtained for a categorial emotion evaluation of the data set. For the dimensional labeling, the average valence and arousal level over the reliable labelers was calculated. Here, no samples were excluded from the ground truth obtained for a dimensional emotion evaluation of the data set. The results obtained from the annotation of the categorial and dimensional labeling are presented in Section 5.1.2. Here, also a statement on the labeler's reliability by evaluating their level of satisfaction and the IRR is made. To confirm the consistency of both annotation approaches, a comparison of both approaches is conducted. Furthermore, to gather additional information on the quality of the annotation, a comparison of the results obtained by the drivers themselves (drivers' self-reports) and the annotation results is presented. As the annotation approach presented is very time consuming an alternative machine-learning-assisted annotation approach is introduced.

Finally, the labels obtained from annotating the high quality audio recordings of the headset microphones were mapped onto the corresponding speech samples of the noisy shotgun microphone recordings, which will later be used to classify the drivers' emotional state in a non-intrusive way. These noises do not only include environmental disturbances but also overlapping speech of the co-driver. As the inlet of the headset microphone was directed towards the driver's mouth, overlapping speech was not received by this microphone. Therefore, all samples containing overlaps in the shotgun microphone recordings were excluded from the data set. The results of this post-annotation processing are presented in Section 5.1.3.

5.1.1 Pre-Processing the Raw Audio Material

During the data collection the audio recordings of the three synchronized microphones, which were integrated onto the dashboard of the vehicle, were received for each of the four driving scenarios separately. This resulted in 12 audio recordings per driver (four audio recordings per driver for each of the three microphones). Unfortunately, due to the unstable energy supply inside the vehicle, which led to interruptions of some audio recordings, for some drivers more than one recording per driving scenario was received. By manually inspecting the recordings, they were sorted to the corresponding driving scenario afterwards. Furthermore, only those recordings containing speech data from the driver needed to be annotated by the labelers. These speech segments were extracted from the recordings by applying the *Sound Finder* tool provided by Audacity® [1] on the high quality headset recordings. To ensure a high accuracy of the speech segments, all segments for which sound was detected by the tool, were manually checked and, if needed, corrected or removed. Additionally, to ensure a reliable annotation of the extracted speech segments the recordings were divided into smaller sub-samples of about 2 seconds length, for which it can be assumed that no change in emotion occurs while an optimal length for a reliable decision making of the labelers is available [Pell & Kotz 2011]. If a segment needed to be divided, the remaining part should not be below a sufficient length of 0.5 seconds. Therefore, these short segments were added to the previous sample coming from the same speech segment, such that a sample could reach a maximum length of 2.5 seconds.

This segmentation process resulted in 16988 speech samples (6.96 hours of speech recordings) originating from 30 drivers (seven female drivers), which were annotated by the labelers as described. Table 5.1 gives an overview on the number of samples originating from the four different driving scenarios. On average female drivers uttered 594 speech samples during the data collection, while male drivers uttered 558 samples. A repeated-measures Analysis of Variance (ANOVA) (cf. Section C.1) revealed that the number of samples originating from the different driving scenarios are significantly different ($F(2.9, 84.4) = 35.4$, $p < 0.001$, Huynh-Feldt-corrected). By performing post-hoc t-tests it could be shown that the number of samples originating from the *mild anxiety* scenario is significantly higher than the number of samples recorded from the other emotion scenarios (all p's < 0.001, Bonferroni-corrected).

[1] Audacity® software is copyright ©1999-2019 Audacity Team. The name Audacity® is a registered trademark of Dominic Mazzoni.

Table 5.1: Number and length of time of samples originating from the four different driving scenarios. Brackets denote the share of male/ female samples.

Scenario	Samples [#]	Time [min]
Neutral	3968 (2987 / 981)	101.60 (77.17 / 24.43)
Positive	3529 (2681 / 848)	87.98 (67.19 / 20.79)
Frustration	4206 (3114 / 1092)	99.35 (73.79/ 25.57)
Mild Anxiety	5285 (4048 / 1237)	128.86 (97.48 / 31.39)
$\sum$	16988 (12830 / 4158)	417.80 (315.62 / 102.18)

5.1.2 Annotation Results

In total the three labelers needed to annotate 6.96 hours of speech material, comprising 16988 speech samples. It took on average 43.78 hours for each labeler to conduct the full annotation consisting of the dimensions valence and arousal, the emotion categories and the satisfactory level of the labeler's annotations. In sum 131.35 hours of annotation time was needed to obtain a ground truth of the real-world in-car data collection described in Section 3.2. From the labelers' satisfactory level a high usability of the annotation results can be concluded. For 4.44% of the annotations the labelers reported to be *very satisfied* in their decision process, for 80.71% *satisfied* and for 14.10% *neutral*. Only 0.60% of the annotations were assigned to *dissatisfied* and a very small share of 0.16% to *very dissatisfied*. For all three labelers an increase of satisfaction of their annotation with an increasing number of conducted annotations was noticed, indicating that a self-training of the labelers occurred while conducting the annotations.

Inter-Rater Reliability

By determining the IRR separately for the dimensional and categorial annotation, a labeler who annotated contrarily to the other labelers was excluded from the labeling process. Table 5.2 shows the average IRR over all evaluated drivers for all possible combinations of the three labelers. For the annotation of valence and the emotion categories the best results were obtained when leaving out labeler 2, while for the annotation of arousal leaving out labeler 1 resulted in the best IRR. The IRR when leaving out labeler 3 even dropped to negative values indicating a negative correlation between the labelers' results. This implies that labeler 1 and 2 have a strong disagreement within their annotations, especially for arousal.

For the dimensional approach, by leaving out labeler 2, an increase of the IRR of valence was achieved, while leaving out labeler 1 led to an increase of arousal. The difference between the IRR of the valence and arousal annotation when leaving out labeler 2 is noticeably higher compared to when leaving out labeler 1. To ensure that both dimensions are labeled sufficiently reliable, it could be assumed that it is

Table 5.2: Average and standard deviation of the IRR for all possible combinations of labelers for the dimensional and categorial annotation.

	Valence	Arousal	Categorial
All	0.37 (0.09)	0.16 (0.10)	0.24 (0.05)
w/o Labeler 1	0.35 (0.14)	**0.24** (0.15)	0.21 (0.07)
w/o Labeler 2	**0.49** (0.08)	0.18 (0.17)	**0.30** (0.07)
w/o Labeler 3	0.22 (0.12)	-0.05 (0.18)	0.20 (0.07)

advisable to rather leave out labeler 1, which results in a satisfactory IRR for both, valence and arousal, than leaving out labeler 2. However, as this assumption is based on the average values stated in Table 5.2 and leaving out labeler 2 for arousal showing a noticeably high standard deviation in relation to its average value, it is recommended to also perform a driver dependent evaluation of the IRR. In this driver dependent evaluation of the IRR, it was assumed that the threshold of leaving one labeler out of the labeling process is set to a delta of ±0.05. When doing so, it was noticed that using all labelers or leaving out labeler 3 would never result in an increase of IRR above this threshold, but leaving out either labeler 1 or labeler 2 always led to an increase above the threshold. As described above, in case of dimensional annotation, a compromise between a good annotation of valence and arousal needed to be made. This decision was made for each annotated driver individually. Out of all 30 drivers, labeler 1 was excluded from the labeling process 14 times and labeler 2 16 times. By considering these cases an average IRR of 0.44 for valence and 0.31 for arousal was achieved. From Section 2.1.4 these values indicate a fair and moderate agreement of the labelers (cf. Table 2.4 on page 23), respectively, and, in case of valence, even outperform reported IRRs of comparable data sets (e.g. 0.199 for valence and 0.485 for arousal in [Siegert et al. 2014]).

For the categorial annotation the best average IRR was achieved when leaving out labeler 2. From the driver dependent evaluation it was noticed, that the best results for each annotated driver were obtained when leaving out labeler 1 one time, labeler 2 16 times and labeler 3 one time. For the remaining 12 drivers the best IRR was achieved using the annotation results of all three labelers. Considering these cases an average IRR of 0.30 was reached. This indicates a fair agreement of the labelers (cf. Table 2.4 on page 23), comparable to other reported studies (i.e. 0.208 for a word list of size eight, and 0.195 for a word list of size 11 in [Siegert et al. 2014]).

Label Assignment

The labels of the dimensional approach were assigned to the desired audio sample by averaging the annotation results of the considered labelers. The averaged values of the valence/ arousal levels, obtained by utilizing the SAM scale, were then mapped onto the four quadrants and the origin of the valence-arousal space (q1 - q4, n) (cf.

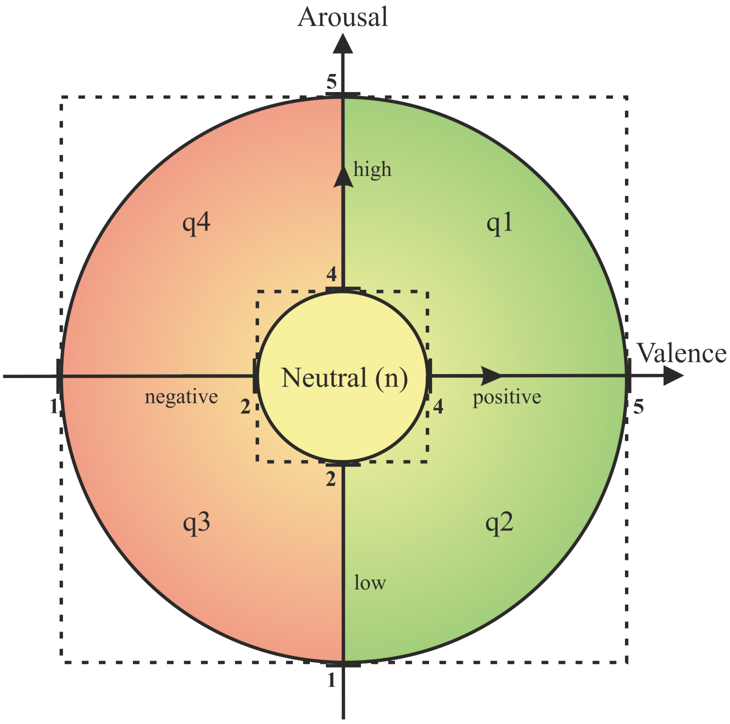


Figure 5.1: Mapping of the valence/ arousal values onto the four quadrants of the valence-arousal space using the 5-point SAM scale.

Figure 5.1). The detailed mapping of the averaged valence and arousal intervals onto dimensional categories is stated in Table 5.3. Own pre-studies indicate that these intervals show the best agreement with the categorial annotation approach (cf. Table 5.4 on page 143). Choosing the neutral interval in a more narrow region would result in a stronger confusion of *neutral* categorial samples lying in an area of distinct high/ low arousal and positive/ negative valence. Additionally to the four quadrants and the neutral area, samples lying outside of the neutral area and directly on either the valence or arousal axis were labeled as *high*, *low*, *positive* and *negative*, respectively. This mapping resulted in 14291 *n*, 280 *q1*, 499 *q2*, 968 *q3*, 102 *q4*, 1 *high*, 620 *low*, 156 *positive* and 70 *negative* speech samples.

Table 5.3: Mapping of the valence and arousal values, obtained by utilizing the 5-point SAM scale, onto the dimensional categories.

	Valence	Arousal
n	[2, 4]	[2, 4]
q1	(4, 5]	(4, 5]
q2	(4, 5]	[1, 2)
q3	[1, 2)	[1, 2)
q4	[1, 2)	(4, 5]

The large amount of samples mapped onto the neutral region of the valence-arousal space is striking, but reasonable for this kind of highly natural and low expressive recorded audio data.

For the categorial labeling a majority voting of the labelers' annotation results was conducted. In case of a labeler lowering the IRR below the threshold, as described before, these labelers were excluded from the majority voting and the remaining two labelers had to be fully concordant in their annotation result. This resulted in 11230 categorially labeled audio samples, which corresponds to 66% of the original speech samples. In detail 5139 *neutral*, 2150 *positive*, 2329 *frustrated* and 1612 *anxious* speech samples were obtained. The high number of neutral samples is explicable by the experimental design, as a neutral emotional state will naturally occur in all designed driving scenarios, without the need of being induced.

To verify the annotation results obtained by both annotation approaches independently, the confusion matrix of both approaches was determined. The results are presented in Table 5.4. Because of the low number of dimensionally annotated *high* samples, this label was left unconsidered. In Table 5.4, green entries denote a correlated assignment between both annotation approaches, while red entries denote an uncorrelated assignment. A high correlation of the annotation approaches is indicated by high values in green and low values in red entries. For the stated confusion matrix, a high consistency between the two annotation methods can be concluded. Already slight changes of valence and arousal indicate a change of the emotional state. Therefore, the assumption could be drawn that the true neutral region lies closer around the origin of the valence/ arousal dimensions than assumed. However, This would results in a much higher confusion of neutral categorial samples, as described earlier in this Section, which would reduce the high consistency of the other

Table 5.4: Confusion matrix of the speech samples assigned to the emotion and dimensional categories. Green entries with high values and red entries with low values indicate a high relation between the two assignment methods. Bold values indicate a high consistency between the assignment of a certain dimensional category and an emotion category.

	n	q1	q2	q3	q4	low	positive	negative
Neutral	4479	1	156	109	0	394	0	0
Positive	1587	**261**	165	0	0	2	**135**	0
Frustrated/ Angry	2060	1	3	99	**93**	17	0	**55**
Anxious/ Fearful	1126	0	6	423	2	51	0	4

annotation results. For the dimensional categories of q1 and q4 a clear assignment to the emotion categories of positive and frustrated, respectively, was possible. For q2 and q3 a large share of samples was assigned to positive and anxious, respectively. A majority of the remaining samples were assigned to the neutral emotion category. This is in line with the emotion models presented by [Holzapfel & Fuegen 2002] and [Almeida et al. 2016] where it is assumed that the neutral region is elongated towards low arousal in the valence-arousal space. Furthermore, this assumption was confirmed by the large number of labels assigned to the neutral emotion category but mapped to the low dimensional category.

Evaluating the Experimental Setup

The categorial annotation approach was designed in a way that the emotion categories would correspond directly to the induced emotional states during the driving scenarios of the experimental setup. By determining the share of categorially annotated samples originating from the four driving scenarios, a clear statement on the correctness of the experimental design was made. A confusion matrix of the results is given in Table 5.5. Each column of the Table contains the samples annotated as one of the emotion categories originating from each driving scenario. It can be noticed that the annoation results are in line with the experimental setup, as a relative majority of samples annotated to the emotion categories originated from the corresponding driving scenario where this emotional state was induced. The large number of samples labeled as neutral over all driving scenarios (first column) is reasonable, as neutral speech was uttered in all the designed scenarios. The same holds for the number of samples labeled as positive (second column) as most of the participants were very positive during the conversation with the interviewer. Also the low number of samples labeled as frustrated and anxious in the neutral and positive scenario is reasonable as the participants also talked about frustrating situations they experienced beforehand. As the mild anxiety scenario was conducted after the frustration scenario and they were both based on the evaluation of a technical system which did not work properly, some of the participants also experienced strong frustration during the anxiety task. This explains the high number of samples labeled as

Table 5.5: Confusion matrix of the categorial annotation in the four driving scenarios.

Scenario	Neutral	Positive	Frustrated/ Angry	Anxious/ Fearful
Neutral	1693	564	192	203
Positive	1156	916	189	104
Frustration	1035	349	1239	227
Mild Anxiety	1255	321	709	1078

Table 5.6: Mean and standard deviation of the valence and arousal annotation in the four driving scenarios. Arrows describe the tendencies of the mean values compared to the mean of the neutral scenario.

Scenario		Valence		Arousal
Neutral	-	3.14 (0.19)	-	2.32 (0.29)
Positive	↑	3.28 (0.17)	↑	2.44 (0.28)
Frustration	↓	2.83 (0.19)	↑	2.62 (0.29)
Mild Anxiety	↓	2.79 (0.14)	-	2.34 (0.25)

frustrated in the mild anxiety scenario. This tendency is in line with the drivers' self-report obtained by utilizing the Geneva Emotion Wheel (GEW) and presented in Table 3.1 on page 97. In this Tables, the items anger and fear were more likely selected during the corresponding driving scenarios, while positive affects were experienced in all driving scenarios. Only a small increase of experienced frustration during the mild anxiety scenario was noticed, which also corresponds to the annotation results obtained for this scenario.

Analogous to the categorial approach, the dimensional approach was used to validate the four driving scenarios. This was done by determining the average valence and arousal values over all speech samples originating from the four driving scenarios. The averages and standard deviations of these values are stated in Table 5.6. From a theoretical perspective and the results presented in Table 5.4, it was assumed that in comparison to the valence and arousal level during the neutral driving scenario the valence level would increase for the positive scenario and decrease for the frustration and mild anxiety scenarios. For the arousal level it was assumed that for the positive and frustration scenario an increase of arousal would occur, while for the mild anxiety scenario this value would decrease. Except for the average arousal value of the mild anxiety scenario, where a small increase of arousal was noticed, all theoretical assumptions were met. This also corresponds to the results obtained by the drivers' self-report presented in Table 3.2 on page 98. In this Table, the average valence values of mild anxiety behave as expected, while the average arousal values show a significant increase compared to the neutral and positive scenario.

Reducing Annotation Time: Semi-Automatic Labeling

From the previous Section it is known that the annotation of emotional states of the driver is highly time consuming (i.e. 43.78 hours of annotation for 6.96 hours of speech material in the present case). As many annotation processes are limited with their resources regarding time consumption and budget limitations, it is of large interest for the machine learning community to reduce the annotation time while maintaining a high quality of the annotation results. Therefore, a machine-learning-assisted annotation approach is presented in the following (as introduced in [Requardt; Egorow et al. 2020]), based on the semi-automatic labeling approach developed in [Egorow et al. 2017]. In this publication, Egorow and I present a semi-automatic approach for the detection/ annotation of filled pauses. Because of the low amount of available training material for this kind of rare speech events and the high effort in utilizing a conventional fully manual annotation [Böck et al. 2019], a semi-automatic annotation approach based on transfer learning was presented (cf. Section 2.1.4 on page 27). First, a filled pauses classifier trained on already existing annotated data (of low amount) was established. Afterwards, this classifier was tested on unknown data. By manually correcting the detected filled pauses of the unknown data the number of training material could be increased while the effort of a manual annotation was decreased dramatically. In the presented approach it was not aimed at achieving high detection rates but rather to increase the amount of exact training material of less-observed rare speech events. Therefore, the number of false negatives (not detected filled pauses) was left unconsidered and only the exactness of true positive detected speech segments (i.e. percentage overlap between the detected filled pause and the verified filled pause) was further evaluated while false positive detected segments were removed manually. By measuring the time it takes to adjust and verify the automatically detected filled pause, we could determine an average time of 20 s for the adjustment and verification of a true filled pause and 5 s to remove a wrongly detected filled pause. To estimate the time of the semi-automatic annotation process the following equation was introduced:

$$T_{semi\_auto} = TP \cdot 20 \text{ s} + FP \cdot 5 \text{ s}, \tag{5.1}$$

with TP being the number of true positives, for which the annotation of the filled pause was done correctly but an adjustment of the label was necessary and FP being the number of false positives with a filled pause being falsely detected and needed to be removed. Compared to the conventional way of a fully manual annotation, we could decrease the annotation effort by 85% (cf. [Egorow et al. 2017]).

To decrease the annotation time of the emotional in-vehicle data presented earlier in this Section, the developed semi-automatic approach of filled pause detection was adapted to be used for annotation of categorial emotional states. Therefore, a classifier was trained only considering the annotation results of the first subject. Afterwards, this classifier was applied to the unknown pre-processed speech samples of the next subject. Theoretically, the next step would be to let the labelers re-evaluate the outcome of the classifier. During this step, two cases need to be considered, which are True Positives (TPs) for which the emotion label is set correctly by the classifier and False Positives (FPs) for which the emotion label is set incorrectly. It was assumed that in case of expert labelers the time needed to verify a TP is much lower than the time needed to re-evaluate a FP, as the labeler not only needs to decide for a wrong classification, but also needs to assign a new label to the sample. As for the present in-vehicle data the labels of the full manual annotation were available, these labels and the statistics of this labeling process were used to present an estimate of this re-evaluation step. The results of the (estimated) re-evaluation were then used to train a new classifier using all annotated and verified data samples. The resulting classifier was then, again, applied to the unknown pre-processed speech samples of the next subject. This process was repeated until the data of all subjects was annotated. To evaluate, whether this approach would lead to a major decrease of annotation effort, the estimate of this semi-automatic approach is presented in the following.

From the annotation of the full speech data set (cf. previously stated in this section on page 140) it is known that the complete annotation of all 30 subjects took 131.35 hours (on average 43.78 hours for each labeler). In this time, the labelers annotated the dimensions of valence and arousal, the four emotion categories and rated the satisfaction level of their annotation. By excluding the time needed to rate their level of satisfaction, an average annotation time of 35.62 h ± 3.70 h for each labeler was obtained. For each speech sample the labelers took on average 6.66 s ± 1.02 s, which includes the time needed listening to the speech sample and excludes outliers. On average the labelers listened to each speech sample 1.46 ± 0.29 times before making their decision. As in total 16988 samples of approximately the same length were evaluated, these statistics represent the population in a high quality. A separation of the time needed to conduct the categorial and dimensional annotation was not possible, as these times were not tracked independently. In the annotation process this was not reasonable as the labelers were aware of the sequenced annotation process and it cannot be said, when they decided on which label to choose for the two different approaches. This, however, was not a problem, as both estimate and real value are based on the same averaged annotation time and the time needed to obtain the dimensional label can therefore be taken as systematic error. In the presented approach the time used to verify a TP ($\hat{T}_{TP}$) and re-evaluate a FP ($\hat{T}_{FP}$) was estimated using the above presented statistics of the full manual annotation results. This resulted in the following equation to determine $\hat{T}_{TP}$:

$$\hat{T}_{TP} = 1.46 \cdot 2 \text{ s} + 1 \text{ s} = 3.92 \text{ s} < 4 \text{ s}, \tag{5.2}$$

with the estimate of 1.46 times listening to the speech sample of an average length of 2 s (cf. Section 5.1.1) and the assumption of 1 s needed to verify the perceived emotional state. The estimated re-evaluation time of FP was calculated as

$$\hat{T}_{FP} = 6.66 \text{ s} < 7 \text{ s}, \tag{5.3}$$

which corresponds to a similar annotation time compared to a fully manual annotation of one speech sample. Based on $\hat{T}_{TP}$ and $\hat{T}_{FP}$ and the number of TP and FP for the considered subject, the estimated time needed to perform the semi-automatic annotation was determined as

$$\hat{T}_{semi\_auto} = TP \cdot \hat{T}_{TP} + FP \cdot \hat{T}_{FP} + T_{S_1} \tag{5.4}$$

$$< TP \cdot 4\text{ s} + FP \cdot 7\text{ s} + 4020\text{ s}, \tag{5.5}$$

with T_{S_1} denoting the time needed to manually annotate the first subject's speech samples which corresponds to 4020 s for the utilized data set.

To verify the stated averaged annotation times the total estimated time needed to conduct a fully manual annotation $\hat{T}_{manual}$ was calculated using

$$\hat{T}_{manual} = \#Samples \cdot 7s. \tag{5.6}$$

This led to an estimated manual annotation time of 33.03 h, which lies inside the standard deviation of the real value of 35.62 h± 3.70 h.

By conducting classification experiments as described, it was possible to identify 5313 samples correctly while for 11052 samples the label of the emotional state needed to be changed manually. As classifier a baseline Support Vector Machine (SVM) with a Radial Basis Function (RBF) was utilized. As hyper-parameter the default values of WEKA [Hall et al. 2009] were applied ($\gamma = 1/dim(FeatureSpace)$ and $C = 1$). The classifier was trained and tested on the *emobase* features of the high-quality headset microphone recordings. By applying Equation (5.5) an estimate of 28.51 hours for the semi-automatic annotation approach was obtained. Compared to the conventional fully manual annotation this corresponds to a time reduction of 19.96% for each labeler. The accumulated confusion matrix of the conducted experiments is stated in Table 5.7. As the annotations of the emotion categories were obtained by conducting a majority voting, for some samples no true label was obtained. This case was added as additional class to the classification problem and is referred to as *N/A*. However, it was not included in the training

Table 5.7: Confusion matrix of the performed classification experiments to determine the efficiency of the semi-automatic annotation approach. All TPs are highlighted in green. The last column of the table only contains "0" entries, as this class was not included in the classification process (adapted from [Requardt; Egorow et al. 2020].

		Predicted				
		anxious	frustrated	neutral	positive	*N/A*
True	anxious	565	140	655	149	*0*
	frustrated	324	667	853	370	*0*
	neutral	663	467	3199	580	*0*
	positive	308	246	654	882	*0*
	N/A	1297	703	2706	937	*0*

process of the classifier, as it would bias the classifier towards recognizing this widely spread class which could include samples belonging to any of the four emotional states. From conducted comparable classification experiments this statement was verified. As the classifier was forced to assign one of the four emotional states to each sample, even with the true label being *N/A*, it is assumed that the labelers re-evaluated these samples. By doing so, we achieve an overestimation of annotation time. Consequently, by applying the semi-automatic approach to a comparable categorial emotion labeling this would lead to an even higher percentaged decrease in annotation time compared to the estimated value presented. Nevertheless, an achieved annotation time reduction of 19.96% is of high importance especially for annotation processes which go along with high annotation effort.

An additional positive side effect which can be observed is the continuous increase of the classification performance with an increasing number of the manually labeled data samples inside the training set. Especially for applied research, this can be of great interest, for example if a certain performance of the classifier is needed before switching to a Semi-Supervised Learning (SSL) based labeling approach for incoming new data samples. The increase of performance in terms of F1-measure is stated in Figure 5.2. An increase of the classification performance with the amount of manually annotated training data can be noticed. The small drops in the classification performance are explicable by the recording setup. Even by utilizing high quality headset recordings, there exist quality differences between recordings of different drivers. Additionally from further evaluations presented in Chapter 6, it is known that not all drivers were able to express the relevant emotional state in a sufficiently expressive manner allowing it to be correctly detected by the classification algorithm, which was especially the case for subject no. 27.

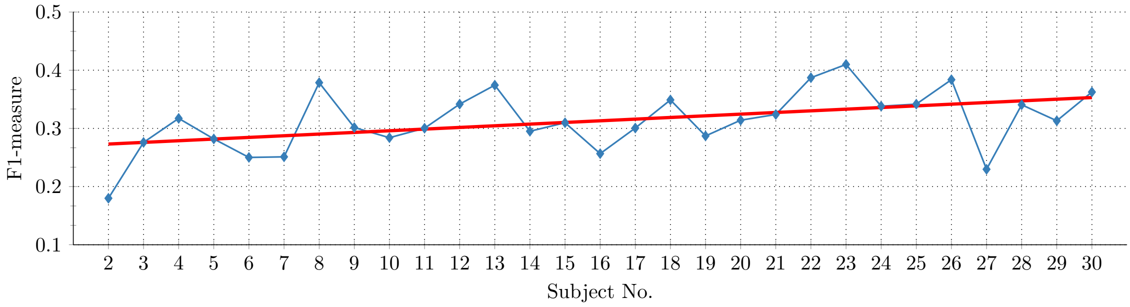


Figure 5.2: Evolution of the F1-measure with increase of manually annotated training data and index number of subjects. The red line represents the linear regression line through the data points, indicating a clear increase in F1 (adapted from [Requardt; Egorow et al. 2020]).

5.1.3 Post-Annotation Processing

All annotation results presented in the previous section were obtained by evaluating the high quality headset microphone recordings as well es the self-reports filled out by the subjects during the four different driving scenarios and served to gather the ground truth of the noisy speech samples recorded by the two non-intrusive shotgun microphones mounted on the dashboard of the vehicle. The speech samples extracted from the high quality headset recordings do not contain disturbances by other passengers inside the vehicle and considerably less noise, as the microphone's inlet was located closely to, and directed towards, the driver's mouth with a cardioid directional pattern suppressing sound coming from any other direction. Even though, the integrated shotgun microphones have a highly directional supercardioid/ lobar beam pattern, the location of the microphones on the dashboard of the vehicle and the distance between speaker/ mouth and microphone inlet make the recordings prone to external disturbances. These circumstances may lead to the occurrence of speech not associated with the driver or overlapping speech. Furthermore, additional disturbances of the speech signal, which make it impossible to perceive the speech of the driver may occur. Therefore, all corresponding speech samples of the shotgun microphone recordings were manually checked by the author of this Thesis. In this process each speech sample was labeled as *yes*, *no* or *overlap*. With *yes* assigned to all suitable samples, *no* assigned to all unsuitable samples (i.e. no perceptible speech content) and *overlap* assigned to samples containing overlapping speech. The explicit label of *overlap* was not further evaluated in the scope of this Thesis, but made the data suitable for the evaluation of overlapping speech as investigated by my colleague in [Egorow & Wendemuth 2019]. Only for those samples assigned with a *yes* a mapping of the emotion label for the noisy speech samples was carried out. As no clear distinction between the critical emotional states of frustration and anxiety was possible by evaluating the drivers' valence and arousal level (cf. Table 5.4), a focus was drawn on the categorial annotation, which allows to distinguish four emotion categories of neutral, positive, frustrated and anxious drivers.

The resulting sub set of annotated noisy speech samples is presented in Table 5.8. It contains in total 7562 samples originating from 28 participants (six females) and comprising 186.24 minutes of speech material. The data of two participants (one male, one female) were left unconsidered due to the high number of frame drop-outs in the audio recordings. It can be seen that all emotion categories are represented with a sufficient number of samples. A repeated-measures ANOVA revealed a significant effect of the emotion categories on the number of speech samples (main effect category: F(1.5,40.1) = 41.1, $p < 0.001$, Greenhouse-Geisser-corrected). A post-hoc t-test further revealed that the number of neutral samples is significantly higher than for all other emotion categories (all p's < 0.001, Bonferroni-corrected). Still, none of the emotions is strongly underrepresented, such that a bias of the developed classifiers towards over-represented categories is unlikely. Furthermore, classifiers

Table 5.8: Number of suitable categorially labeled samples contained in the emotional real-world in-vehicle recordings of the shotgun microphones. Brackets denote the share of male/ female samples.

Label	Samples [#]	Time [min]
Neutral	3689 (3005 / 684)	88.90 (72.45 / 16.44)
Positive	1168 (897 / 271)	30.15 (23.50 / 6.65)
Frustrated	1606 (1140 / 466)	40.34 (28.26 / 12.07)
Anxious	1099 (879 / 220)	26.48 (21.35 / 5.49)
$\sum$	7562 (5921 / 1641)	186.24 (145.57 / 40.67)

that are able to process the degree of unbalanced data distributions in the used data set were chosen.

The post-processed annotation results were used in Chapter 6 to evaluate the ability to detect the emotional state of a driver in a real-world driving scenario. By excluding samples with no perceptual speech, speech not associated with the driver and overlapping speech, it is possible to remove the influence of these factors on the emotion recognition performance and focus on the effects of in-vehicle noises on the speech signal. For future research it would be of great interest to also consider the difference of the individual recording setups on the recognition performance (e.g. headset vs. shotgun microphone vs. unprocessed recordings). This, however, was not investigated in the scope of this Thesis.

5.2 Speech Enhancement

In this Chapter I will investigate the effect of a selected conventional speech enhancement method on the recognition rates of speech emotion recognition and give a recommendation whether speech enhancement can lead to reliably significant improvement of the recognition performance. The experiments were performed using the re-recorded Berlin Emotional Speech Database (EmoDB) under in-car recording conditions (EmoDB-Car) described in Chapter 3.1. First, a brief overview on relevant speech enhancement approaches is given, followed by a more detailed description of the applied speech enhancement method (cf. Section 5.2.1). Afterwards, a focus is drawn on an evaluation of the enhanced speech signal compared to its counterparts under disturbed and silence recording conditions, respectively (cf. Section 5.2.2). This evaluation is based on a feature-value similarity analysis and occurring differences in the waveforms and power spectrums of the considered signals. Hereinafter, classification experiments based on the enhanced, disturbed and silence recordings are utilized (cf. Section 5.2.3). The Chapter is concluded in Section 5.2.4 by a recommendation on the application of speech enhancement methods for speech emotion recognition.

An overview on the most relevant speech enhancement methods applied to increase the speech intelligibility of disturbed speech is given in Chapter 2.5.2. Regarding these algorithms in combination with the simulated in-car recordings, one major challenge is the absence of a microphone dedicated to pick up the noise signal only, as it needs to be the case when applying adaptive noise cancellation approaches. Furthermore, most speech enhancement algorithms, without a reference microphone capturing the noise signal only, are based on the application of a microphone array. Even though two microphones were integrated into the simulator, this setting is not suitable for speech enhancement using a microphone array, as the array needs to satisfy special restrictions (e.g. number of microphones, distance between microphones, far-field source assumption). These restrictions can be met, if the application of an speech enhancement technique were the main scope of the research. In this Thesis, however, the focus is drawn on the emotion recognition of disturbed speech, for which high quality audio recordings are needed. To achieve this, highly directional shotgun microphones were chosen, which are unsuitable to be arranged in huge microphone arrays. Therefore, commonly used speech enhancement methods based on delay-and-sum beamforming cannot be utilized using the presented microphone setup. As no advantage of the two-microphone setup can be drawn for speech enhancement, a method suitable for the application on one single microphone receiving a noisy speech signal needs to be considered. In co-operation with a colleague from NUANCE, using his experience and expertise in speech processing, we opt for the Optimally-Modified Log-Spectral Amplitude (OM-LSA) speech estimation and Improved Minima Controlled Recursive Averaging (IMCRA) noise estimation approach as presented in [Cohen & Gannot 2008]. Contrarily to other speech enhancement methods, it is not based on a voice activity detection where the noise estimate is updated only in case of speech absence, but on a continuous update of the speech and noise estimate by utilizing the speech presence probability.

5.2.1 Theoretical Background on Speech Enhancement

This Section gives a brief overview on statistical model-based speech enhancement methods in frequency-domain as described in [Loizou 2007] and the applied OM-LSA-IMCRA method (cf. [Cohen & Gannot 2008]). The goal of speech enhancement in general is to improve the speech intelligibility of a disturbed speech signal. It is assumed that the signal is degraded by additive noise. The observed, sampled signal is given by

$$y(n) = x(n) + d(n), \tag{5.7}$$

with $x(n)$ representing the clean speech and $d(n)$ the additive noise signal. By applying the Short-time Fourier Transform (STFT) the signal is transformed into the frequency-domain. In comparison to the Discrete-Time Fourier Transform (DTFT), only short segments of (10-30 ms) of speech are analyzed by introducing a sliding analysis window $w(n)$ of the size N. This enables the possibility to process short-time stationary signals, as it is the case for speech. By applying STFT the observed discrete signal spectrum is given as

$$Y_{tk} = \sum_{n=0}^{N-1} y(n + tM)w(n) \exp(-i\frac{2\pi}{N}nk) = X_{tk} + D_{tk}, \tag{5.8}$$

where M denotes the number of samples separating two consecutive window frames ($N - M \hat{=}$ overlap of consecutive frames), t the time frame index ($t = 0, 1, ...$) and k the frequency bin index ($k = 0, 1, ..., N - 1$). X_{tk} and D_{tk} denote the corresponding STFT of the clean speech and noise signal. The goal of speech enhancement is to determine an estimate $\hat{X}_{tk}$ such that it minimizes the squared error distortion measure

$$d(X_{tk}, \hat{X}_{tk}) = \left| g(\hat{X}_{tk}) - \tilde{g}(X_{tk}) \right|^2, \tag{5.9}$$

with $g(X)$ representing specific functions as fidelity criteria. The estimate of the clean speech signal is then obtained by applying the inverse STFT to $\hat{X}_{tk}$,

$$\hat{x}(n) = \sum_t \sum_{k=0}^{N-1} \hat{X}_{tk}\tilde{w}(n - tM) \exp(-i\frac{2\pi}{N}k(n - tM)), \tag{5.10}$$

with $\tilde{w}(n)$ being a synthesis windows that is bi-orthogonal to $w(n)$. Considering two hypotheses of speech presence (H_1^{tk}) and speech absence (H_0^{tk}),

$$H_1^{tk} : Y_{tk} = X_{tk} + D_{tk}, \tag{5.11}$$

$$H_0^{tk} : Y_{tk} = D_{tk}, \tag{5.12}$$

an estimate of $\hat{X}_{tk}$ can be obtained by determining

$$\min_{\hat{X}_{tk}} E\{d(X_{tk}, \hat{X}_{tk}) \mid \hat{p}_{tk}, \hat{\lambda}_{tk}, \hat{\sigma^2}_{tk}, Y_{tk}\}, \tag{5.13}$$

with $\hat{p}_{tk}$ as speech presence probability, $\hat{\lambda}_{tk}$ as estimate of the variance of X_{tk} under H_1^{tk} and $\hat{\sigma^2}_{tk}$ as variance of D_{tk}.

OM-LSA-IMCRA

For the Optimally-Modified Log-Spectral Amplitude (OM-LSA) approach, as fidelity criteria the minimum mean squared error of the log-magnitude spectra is used, with $g(\hat{X}_{tk}) = \log\left|\hat{X}_{tk}\right|$, in Equation (5.9). Applying this function to Equation (5.13), this results in

$$\begin{aligned}\log\left|\hat{X}_{tk}\right| &= \hat{p}_{tk} \cdot E\{\log|X_{tk}| \mid H_1^{tk}, \hat{\lambda}_{tk}, \hat{\sigma^2}_{tk}, Y_{tk}\} \\ &\quad + (1-\hat{p}_{tk}) \cdot E\{\log(G_{min}|Y_{tk}|) \mid H_0^{tk}, Y_{tk}\}.\end{aligned} \tag{5.14}$$

Now, by solving this Equation we get

$$\hat{X}_{tk} = \left[G_{LSA}(\hat{\xi}_{tk}, \hat{\gamma}_{tk})\right]^{\hat{p}_{tk}} \cdot G_{min}^{(1-\hat{p}_{tk})} \cdot Y_{tk}, \tag{5.15}$$

with $G_{LSA}(\hat{\xi}_{tk}, \hat{\gamma}_{tk})$ representing the LSA gain function derived by [Ephraim & Malah 1985] and $G_{min} << 1$ as a constant attenuation factor needed to retain the naturalness of the noise during speech absence (cf. [Cohen & Berdugo 2001]). The terms $\hat{\xi}_{tk}$ and $\hat{\gamma}_{tk}$ denote the estimate of the a-priori and the a-posteriori Signal-to-Noise Ratio (SNR), respectively. To be able to solve Equation (5.15) we now need to determine the estimates of the speech presence probability ($\hat{p}_{tk}$), as well as $\hat{\xi}_{tk}$ and $\hat{\gamma}_{tk}$.

The speech presence probability is estimated by determining the a-priori speech presence probability estimate $\hat{p}_{tk|t-1}$ and applying the Bayes' rule. By applying local and global averaging windows in frequency domain on the recursive averaged a-priori SNR, a local and global estimate of speech presence in the k-th frequency bin of the t-th time-frame is determined (P_{tk}^{local} and P_{tk}^{global}). Additionally, P_t^{frame} is determined, which is based on the speech energy in neighboring frames and averaging the recursive averaged a-priori SNR over a certain frequency bin. Afterwards, $\hat{p}_{tk|t-1}$ is determined by calculating

$$\hat{p}_{tk|t-1} = P_{tk}^{local} P_{tk}^{global} P_t^{frame}. \tag{5.16}$$

The a-priori SNR can be recursively estimated under a Gaussian model as

$$\hat{\xi}_{tk} = \alpha_{tk}\hat{\xi}_{tk|t-1} + (1-\alpha_{tk})(\hat{\gamma}_{tk} - 1), \tag{5.17}$$

with α_{tk} $(0 < \alpha_{tk} < 1)$ as weighting factor that controls the trade-off between the noise reduction and transition distortion introduced into the signal (cf. [Cohen 2005]).

The a-posteriori SNR equates to

$$\hat{\gamma}_{tk} \hat{=} \frac{|Y_{tk}|^2}{\hat{\sigma^2}_{tk}}. \tag{5.18}$$

By applying the Improved Minima Controlled Recursive Averaging (IMCRA) method the estimate $\hat{\sigma^2}_{tk}$ of the noise spectral variance is obtained. This method is based on recursive averaging, by averaging over past spectral power values of the noisy measurement in periods of speech absence and holding the estimate during periods of speech presence:

$$\begin{aligned} \hat{\sigma^2}_{t+1,k} &= \tilde{p}_{tk}\hat{\sigma^2}_{tk} \\ &+ (1-\tilde{p}_{tk}) \cdot \left[\alpha_d\hat{\sigma^2}_{tk} + (1-\alpha_d)\,|Y_{tk}|^2\right], \end{aligned} \tag{5.19}$$

with α_d $(0 < \alpha_d < 1)$ as a smoothing parameter. The term $\tilde{p}_{tk}$ represents an estimate of the speech presence probability, distinct from $\hat{p}_{tk}$ used for estimating the clean speech $\hat{X}_{tk}$ in Equation (5.15), with $\hat{p}_{tk} \geq \tilde{p}_{tk}$. Hence, this approach is more prone towards detecting speech absence compared to $\hat{p}_{tk}$. The term $\tilde{p}_{tk|t-1}$ is determined by conducting two iterations of smoothing and minimum tracking. In the first iteration a rough voice activity detection for each frequency band is performed. In the second iteration only those components containing primarily noise, obtained in the first iteration, are analyzed. Finally, by applying the Bayes' rule $\tilde{p}_{tk}$ is determined and the estimates $\hat{\sigma^2}_{t+1,k}$ and $\hat{\gamma}_{tk}$ are calculated, respectively.

The OM-LSA-IMCRA speech enhancement can be applied to single-channel recording setups with no dedicated noise/ speech only recording being available. As described earlier, for the EmoDB-Car data samples, this is the case. With regard to the evaluation of the two individual microphone setups, a multi-channel speech enhancement approach was not necessarily needed, as this could lead to an interference of the microphone recordings, which would make an individual evaluation unfeasible. To generate the enhanced audio signals, a MATLAB-script provided by the inventor of the OM-LSA-IMCRA approach, Prof. Israel Cohen, was utilized [2].

5.2.2 Examining the Feature Space of Speech Enhanced Data for Speech Emotion Recognition

One important aspect which needs to be addressed is the effect of speech enhancement algorithms on the feature space. Speech enhancement is based on the modification of the disturbed speech signal to achieve better intelligibility of the audio parts where speech is present. The effect these enhancement method have on the

[2] Code available on `https://israelcohen.com/software/`

ability to recognize the emotional content of said speech parts is barely investigated. By modifying the speech signal not only the raw signal itself gets modified but also all the features characterizing the signal. As the emotion classification approaches presented in this Thesis are based on machine learning algorithms using the extracted features of the speech signal, it needs to be investigated to which extent the feature space is influenced by the applied speech enhancement. This was done by comparing the feature sets of the speech signal of the simulated in-car recordings EmoDB-Car under silence, disturbed and enhanced condition (cf. Section 3.1) with each other using the Wilcoxon signed-rank test of significance (cf. Appendix C). In the presented investigation, the samples comprise the recordings and their feature characteristics originating from different recording conditions (silence, disturbed and enhanced). The test was conducted over all ten speakers of the EmoDB-Car data set for each extracted feature of the *emobase* feature set. This resulted in 988 tests conducted on the same data set. From statistics it is well-known that this kind of multiple testing can lead to a so-called family-wise error also known as α-inflation (cf. Appendix C). One way to prevent α-inflation is by utilizing a Bonferroni-correction. With this approach, the level of significance gets reduced depending on the number of tests carried out. A different approach to prevent α-inflation is by dividing the data set into convenient sub sets and conducting a majority voting of the significant features over all sub sets. For this approach, no adjustment of α is needed as the majority voting aggravates the assumption of significance. One suitable way to split the EmoDB-Car data set is according to the speaker of an utterance, as there exist multiple utterances which originate from each speaker. One major advantage of doing this is the speaker dependency of the significantly different features, which would not be considered by utilizing a Bonferroni-correction. Therefore, I opted for this approach to prevent the α-inflation. The results of this approach obtained for three different level of significance with $\alpha = [0.05, 0.01, 0.005]$ are stated in Table 5.9. The Table gives an overview on the number of significantly altered features as percentage split of all evaluated features of the three performed experiments. With silence referring to the re-recorded data samples of EmoDB-Car with the simulator turned off (only in-car acoustics present), disturbed referring to the re-recorded data samples

Table 5.9: Percentage split of significantly altered features considering tree different level of significance ($\alpha = [0.05, 0.01, 0.005]$). The total number of altered features is denoted in brackets.

	Experiment		α	
		0.05	0.01	0.005
(1)	silence vs. disturbed	76% (751)	69% (680)	67% (661)
(2)	silence vs. enhanced	74% (732)	64% (636)	60% (588)
(3)	enhanced vs. disturbed	81% (796)	74% (736)	71% (700)

with the simulator turned on (additional environmental noises and engine sound) and enhanced referring the the speech enhanced disturbed data samples. By using re-recordings only and not considering the speech samples of the original EmoDB data set, we prevent differences in the volume-related features like loudness and intensity, as the recording's setup was kept unchanged throughout the whole data collection. Consequently, changes in the feature values originate from the in-vehicle noises and influences coming from the signal modification of the utilized speech enhancement algorithm only.

Figure 5.3 shows the waveforms of an original EmoDB sample and the corresponding re-recordings under silence, disturbed and enhanced condition of the EmoDB-Car data set, respectively. From a theoretical perspective it can be assumed that by utilizing speech enhancement the enhanced speech signal will align with the signal under silence condition, however, the spectra give more information and will be discussed later. Considering Figure 5.3, this is the case. A clear disturbance of the signal is present in Figure 5.3 c), which depicts the waveform of the noisy speech sample. Figure 5.3 d) corresponds to the speech enhanced version of c) and shows high resemblance with the original and silent waveform (a) & b)). It could be assumed that the features originating from waveform d) would also show a greater agreement to these two speech signals compared to the disturbed signal.

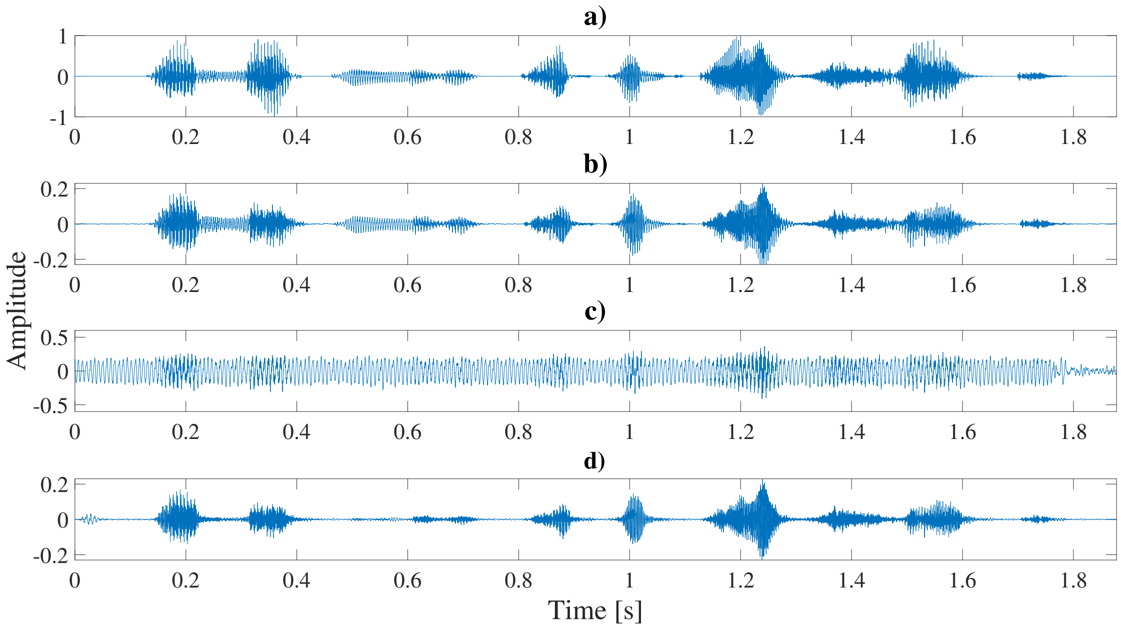


Figure 5.3: Waveforms of an exemplary a) original EmoDB sample, b) corresponding silent re-recording of EmoDB-Car, c) corresponding disturbed re-recording and d) speech enhanced disturbed re-recording.

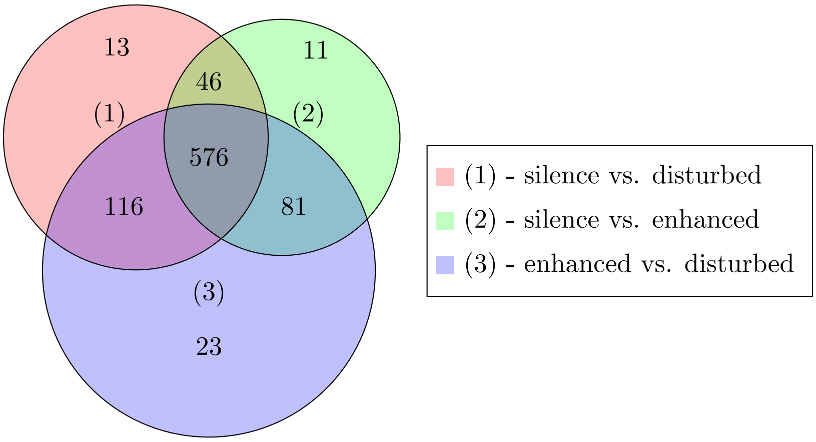

13
46
11
(1)
(2)
576
116
81
(3)
23
🟥 (1) - silence vs. disturbed
🟩 (2) - silence vs. enhanced
🟦 (3) - enhanced vs. disturbed

Figure 5.4: Venn diagram of significantly altered features ($\alpha = 0.05$) of the three experiments, showing the total number of dedicated and consistently altered features over several experiments.

This, however, is not reflected by the results stated in Table 5.9. When looking at the results of significantly altered features with $\alpha = 0.05$ in detail, the following was noticed: the number of significantly altered features of experiment number (2) shows a considerable high percentage split of 74%. Compared to experiment (1) this comprises a difference of 2% only, which corresponds to a total number of 19 features. From the assumption made previously, one would expect to get a much lower number of altered features for experiment (2) than for experiment (1), as the speech signal of the enhanced samples align considerably more to the original and silent samples than to the disturbed samples. Additionally, an even higher number of altered features was identified for experiment (3). The results imply that there exist features which are significantly altered over all given experiments, otherwise the number of altered features in (2) would need to be significantly lower. To confirm this statement a detailed evaluation of significantly altered features of each experiment is presented in Figure 5.4. The total number of altered features for each dedicated experiment and features altered over several experiments can be taken from this Venn diagram. As stated, the majority of features were altered over all three experiments (576 features) and therefore are significantly different in their speech signal even with a high alignment of the speech signal over time.

To understand this high number of altered features throughout the different recording conditions and to get a more detailed insight on the changes arising through speech enhancement we will now take a look at the power spectrum of the different signals, which was obtained by calculating the STFT (cf. Figure 5.5). From the experiments conducted in Section 4.3 it is already known that under non-ideal recording conditions the quality of the speech signal decreases considerably. For the silence condition a decrease of spectral power is noticed due to the absorbing characteristics of the in-vehicle recordings (non-linear distortion) while for the disturbed condition an increase of spectral power is noticed caused by added highway noises of the simulator environment. By applying speech enhancement to the disturbed re-recordings c), the enhanced speech signal d) is obtained. A clear decrease of spectral power, below the spectral power of the speech signal under silent condition b), is visible. This is also confirmed by the calculated Compression Error Rate (CER) (cf. Section 4.1.2 and Equation (4.12) on page 114) values for enhanced speech tested against the silence condition of both microphone settings ($CER_{\%,enh,l} = -3.94, CER_{\%,enh,r} = -7.70$). To recap, a negative CER indicates a decrease of spectral power while positive values indicate an increase. However, the CER only gives a general overview on the average signal power and does not distinguish between speech present and speech absent parts. This is an important issue, as the speech presence parts carry the most information considering speech emotion recognition. When going back to Figure 5.3 it can be noticed, that from second 0.48 to 0.74 there exists a voiced speech part in the original and silence signal, while this part is almost completely absent in the enhanced signal. This is also reflected in the power spectrums presented in Figure 5.5 where for this segment of the signal a noticeable lower spectral power is present, especially in the higher frequency bands above 4 kHz. In general, especially for the higher frequency bands it can be noticed that the signal power is suppressed by the speech enhancement method compared to the power spectrum presented in Figure 5.5 a) and b), respectively. Only for those

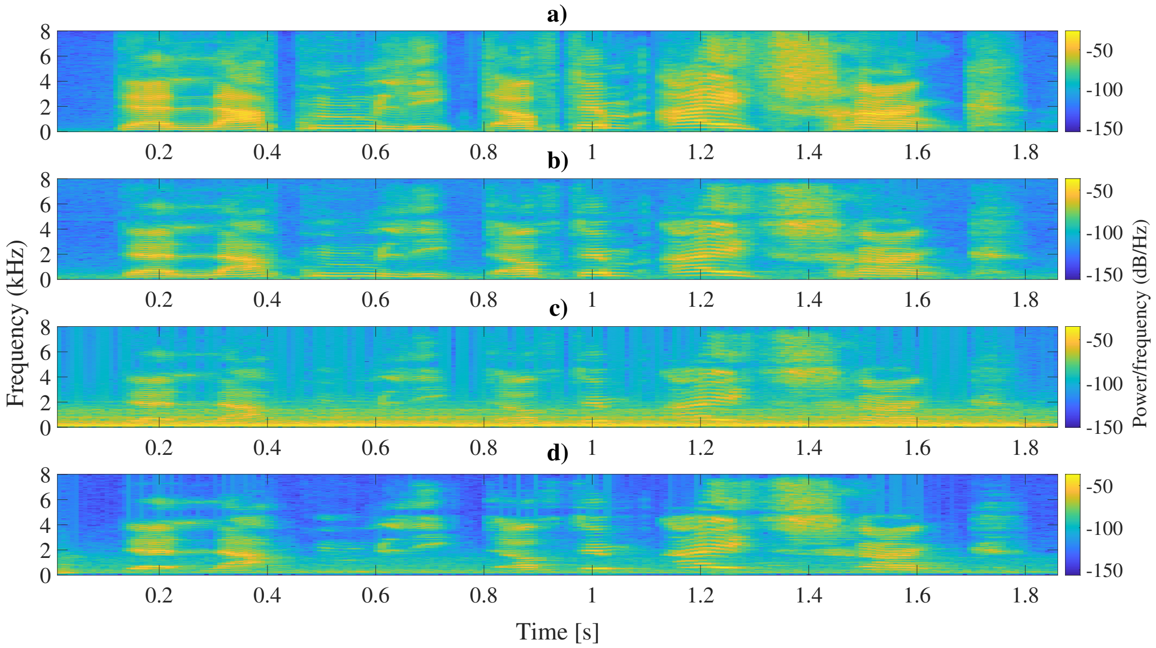


Figure 5.5: Power spectrum of an exemplary a) original EmoDB sample, b) corresponding silent re-recording of EmoDB-Car, c) corresponding disturbed re-recording and d) speech enhanced disturbed re-recording.

signal parts containing speech with a high spectral power a correct identification as speech present was obtained by the enhancement algorithm. Those parts containing speech with lower spectral power were strongly overlain by the in-vehicle noises and were not detected by the speech presence estimator of the OM-LSA-IMCRA speech enhancement algorithm. Consequently, these speech parts were assessed as noise by the enhancement algorithm and were therefore attenuated in the enhanced signal. For those signal parts of speech absence, it can be noticed that the enhanced signal shows a noticeable lower spectral power compared to the original and silence signal, respectively. Overall, it can be stated that by applying speech enhancement the signal obtained under silence condition shows distinct differences in the power spectrum compared to the enhanced signal. While the changes arising in the disturbed recordings can be explained by added environmental noises and engine sound, the changes obtained by applying speech enhancement are less reproducible.

5.2.3 Classifying Emotions from Enhanced Speech

To validate the statements made in the previous Section, several classification experiments, applying the Leave-One-Subject-Out (LOSO) validation scheme, were carried out using the baseline SVM-classifier with a linear kernel of the software tool WEKA [Hall et al. 2009]. As feature set all features of the *emobase* set were utilized. Additionally, two cross-recording evaluations were carried out, where the classifier was trained on disturbed and enhanced speech and tested on silent speech, respectively. The Unweighted Average Recalls (UARs) of the performed experiments are depicted in Figure 5.6. The baseline, silence and disturbed results correspond to the results presented in Section 4.3 and Figure 4.7 on page 131. Additionally, results obtained from experiments tested and trained on the enhanced recordings and the cross-recording experiments, which were tested on the silence recordings (cross-disturbed and cross-enhanced), are presented. A repeated-measures ANOVA revealed that there exist significant differences in the recognizer performance considering both microphone settings individually (main effect left: F(2.2,19.5) = 4.6, $p < 0.05$, Greenhouse-Geisser-corrected; main effect right: F(2.6,23.3) = 4.7, $p < 0.05$, Greenhouse-Geisser-corrected). The results obtained from the left and right microphone under similar recording conditions did not reach the level of significance. By conducting post-hoc t-tests a significant difference between the classification results obtained from the left microphone recordings under silence condition and obtained under disturbed and cross-disturbed condition was shown (all p's < 0.05, Bonferroni-corrected). For the right microphone recordings a significant difference between the results obtained under silence condition and under cross-enhanced and cross-disturbed condition was shown (all p's < 0.05, Bonferroni-corrected). For the remaining post-hoc tests the corrected level of significance was nearly reached. This is in line with the results obtained from the feature space analysis, where experiment (2) showed a lower number of significantly altered features as experiment (1).

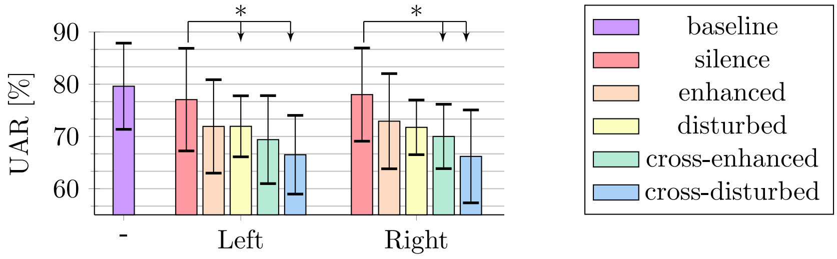


Figure 5.6: UAR and standard deviation of the performed LOSO cross-validation experiments on the original EmoDB, re-recorded EmoDB-Car and enhanced EmoDB-Car. Cross-experiments correspond to classifiers trained on enhanced/ disturbed speech and tested on silent speech. All other experiments were performed under matching training and testing condition. Stars denote the level of significance ($\alpha = 0.05$).

The high numbers of altered features in experiment (1) and (2) explain the strong decrease in the recognition performance between the matching results of the silent experiments and the cross-recording experiments, as it can be assumed that features of higher importance for detecting emotions from the silent speech samples were not chosen in the training process utilizing the enhanced and disturbed speech samples. This further supports the importance of choosing the correct data set, especially when it comes to non-optimal recording condition. In particular the results obtained for the cross-enhanced experiments are of high relevance for this statement, as it demonstrates that for emotion recognition speech enhancement will not automatically lead to an improvement of the recognition performance.

5.2.4 Findings and Recommendations on Speech Enhancement

It can be summarized that by utilizing the OM-LSA-IMCRA speech enhancement algorithm I was able to obtain an enhanced speech signal which aligns well to the speech signal under the silence condition. However, a majority of the extracted features from the enhanced signal is altered significantly compared to the features obtained under the silence and disturbed condition. These strong differences can be explained by the differences in the signals' power spectrum. Especially for higher frequency bands and speech absence signal parts, the enhancement algorithm lowers the spectral power noticeable compared to the silence recordings. For signal parts containing speech with a low spectral power, which were strongly disturbed by the in-vehicle noises, the speech presence estimator of the algorithm is unable to detect the speech reliably. Hence, these signal parts are incorrectly diminished. This is not the case for the disturbed signal, as the signal corresponds to the recording under silence condition (affected by the in-vehicle acoustics only) with added environmental noises and engine sound. The speech content of the signal itself remains unchanged, as it was recorded under the same room acoustics. Additional classification experiments revealed that there exists no significant difference in the recognition performance when using disturbed or enhanced speech signals, respectively. Therefore, it is advisable, for the investigations presented in this Thesis, not to use speech enhancement in case of speech emotion recognition but rather stick to the disturbed speech signal without speech parts being diminished by the enhancement algorithm.

5.3 Summary and Discussion

In this Chapter the pre- and post-processing of real-world in-car recordings was presented. This included the pre-processing of the raw-audio material obtained from the headset microphone worn by the driver. These recordings are assumed to be of higher recording quality, as the inlet of the microphone was directed towards the driver's mouth suppressing noises coming from other directions. The pre-processing included the partitioning of all voiced speech segments into smaller sub-samples in between 0.5 and 2.5 seconds of length. This resulted in 16988 speech samples which then were annotated by three independent, German speaking, expert labelers. The annotation process was conducted in three steps: first, an annotation of the emotion dimensions of valence and arousal was conducted. Second, one out of four emotion categories (neutral, positive, frustrated and anxious) was assessed to the speech sample. Finally, the satisfactory level of the current label assignment was assessed. From the annotation results I could show that there exists a high correlation between the two approaches of dimensional and categorial annotation. However, it is not possible to distinctively distinguish between the emotion categories of a frustrated and an anxious driver by only considering the emotion dimensions of valence and arousal (cf. Table 5.4). Therefore, for the further investigation on the classification of the driver's emotional state, only the results obtained by the categorial annotation are utilized. As the classification should be done by only considering the disturbed recordings of the shotgun microphones, a post-annotation processing of the disturbed speech samples was conducted. All speech samples containing noise and overlapping speech were manually removed from the data set. Afterwards, the labels obtained from the categorial annotation of the high quality headset recordings were mapped onto the disturbed speech samples. This resulted in 7562 speech samples distributed among the four emotion categories as presented in Table 5.8.

As a fully manual annotation of low-expressive naturalistic speech samples is strongly time consuming, an estimate of time needed to conduct a machine-learning-assisted manual annotation of categorial emotions was presented. This theoretical approach resulted in a decrease of annotation time of at least 19.96%. Because of the limitations accompanied by the presented manual annotation approach (e.g. sequenced dimensional and categorial annotation, majority voting to obtain categorial labels), it can be assumed that the actual decrease in annotation time would be even higher.

Furthermore, the effect of speech enhancement on the performance of a speech emotion recognition task was investigated. By applying speech enhancement I was able to align the waveform of a noisy speech signal well to its counterpart under silence recording condition. However, as speech enhancement is based on applying different filter techniques in time and frequency domain onto the original noisy speech signal, the denoised signal does not correspond to a similar speech signal recorded under silence condition. This led to a significantly high number of altered features between the denoised speech samples and their corresponding recordings under silence condition and to noticeable differences between their signals' power spectrum (cf. Figures 5.4 and 5.5). From conducted classification experiments it can be recommended to not apply speech enhancement in case of the speech emotion recognition task presented in this Thesis.

The Results obtained in this Chapter will be used in the next Chapter to evaluate the final research hypothesis of recognizing the drivers emotional state in an everyday driving environment.

Chapter 6

Towards Driver State Monitoring: Classifying Drivers' Emotions

Contents

UNTIL now I have presented the collection of real-world emotional speech data in an in-car driving environment (cf. Chapter 3) and its pre- and post-annotation (cf. Chapter 5). The classification results in Chapter 4 and Chapter 5, however, were solely based on processed benchmark data samples (i.e. re-recorded Berlin Emotional Speech Database (EmoDB) under in-car recording conditions (EmoDB-Car) and re-recorded Vera am Mittag (VAM) under in-car recording conditions (VAM-Car)). This Chapter will now focus on the detection of highly natural and low expressive emotional speech samples and is based on the work presented in [Requardt; Ihme et al. 2020]. The challenges associated with this kind of classification task were already addressed in Chapter 2 with a focus on different factors affecting this task, such as the experimental setup of the data collection, utilized speech pre- and post-processing steps or the design of the classification model. It was stated that for the present kind of low expressive and highly natural emotional speech data, under non-optimal recording conditions, the automatic emotion recognition task is highly challenging and will most certainly lead to low recognition performances of the classifier (cf. Section 2.3). Therefore, a feature selection and hyper-parameter optimization, to push the classifier to its recognition performance limits, is inevitable. Furthermore, with the present lack of available emotional speech data and strongly unbalanced data distributions, it is wise to utilize classification approaches, which can cope with these kinds of limitations. Suitable classifiers for this kind of data are Support Vector Machines (SVMs) and Random Forests (RFs), which were therefore applied for the present classification task (cf. Appendix A for detailed description of the algorithms).

To identify the best performing classifier, several classification experiments were conducted, which will be presented hereinafter. An overview on these experiments is given in Figure 6.1. It depicts all relevant steps from the feature extraction on the post-annotated data samples to the sum of cross-validated classification models, from which the best performing model was later chosen. As data the post-annotation processed data samples, as presented in Section 5.1.3 were utilized. This data comprises 186.24 minutes of speech material, which corresponds to 7562 speech samples originating from 28 participants (six females). The detailed samples' distribution can be taken from Table 5.8 on page 151. Using the feature extraction toolkit OpenSMILE, the *emobase* features (cf. Table 2.3 on page 39) were extracted from

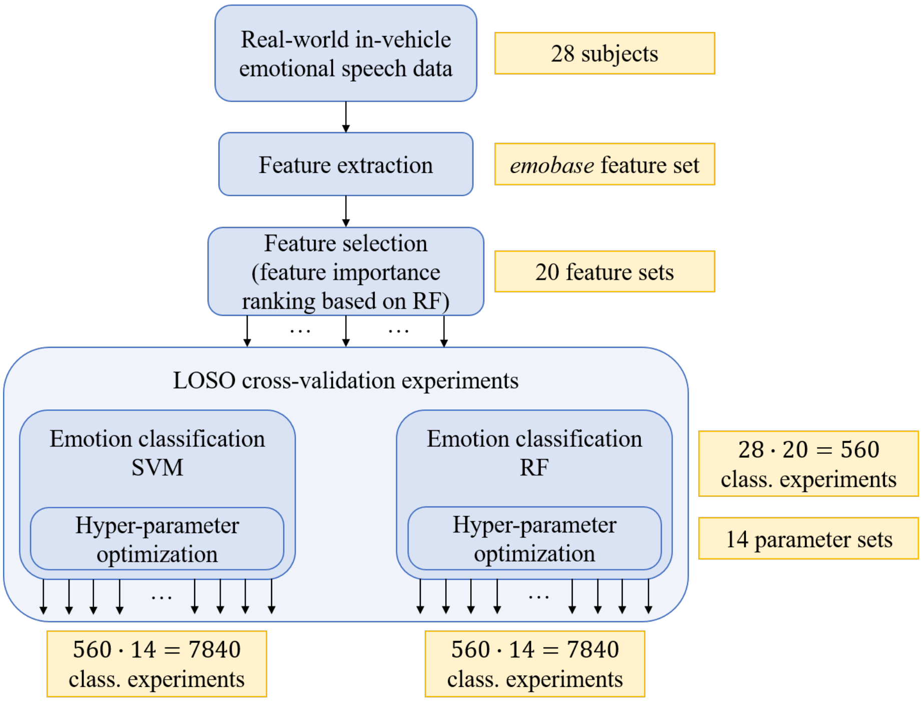

Real-world in-vehicle emotional speech data
28 subjects
Feature extraction
emobase feature set
Feature selection (feature importance ranking based on RF)
20 feature sets
LOSO cross-validation experiments
Emotion classification SVM
Hyper-parameter optimization
Emotion classification RF
Hyper-parameter optimization
28 · 20 = 560 class. experiments
14 parameter sets
560 · 14 = 7840 class. experiments
560 · 14 = 7840 class. experiments

Figure 6.1: Schematic diagram of the experimental setup (adapted from [Requardt; Ihme et al. 2020]).

the speech samples. Afterwards, a feature selection was performed by utilizing a feature importance ranking based on the recognition performance of a RF classifier (i.e. *wrapper* method, cf. Section 2.2.4). A detailed description of the utilized feature selection method and its results are presented in Section 6.1. This approach resulted in 20 feature sets, which were used to train and validate classification models of a SVM and RF, utilizing the Leave-One-Subject-Out (LOSO) cross-validation scheme (cf. Section 2.2.7). As for each cross-validation experiment each subject was used once to validate the classification model and all 20 feature sets needed to be evaluated, this resulted in 560 classification experiments for each classification approach. Additionally to the selection of the optimal feature set, a hyper-parameter optimization was performed. I here took into consideration 14 parameter combinations (optimal parameter candidates), which were obtained by performing a random search (cf. Section 6.2). The selection of the optimal hyper-parameter sets resulted in additional 14 classification experiments which had to be performed for each of the 560 cross-validation experiments. Considering this, in total 7840 classification models needed to be trained and tested to choose and validate the optimal classification model for each, SVM and RF, classification approach (cf. Section 6.3). All implementations presented in this Chapter were performed using MATLAB (Version R2018b).

As parts of this chapter are based on work already published in [Requardt; Ihme et al. 2020], several phrasings are taken literally from this publication.

6.1 Choosing the Optimal Feature Set

As already introduced in Section 2.2.3 and Section 2.2.4, the reduction of the utilized feature set plays a decisive role when it comes to validating the designed classification model. Especially in case of large feature sets, a low number of available data samples for the considered emotional class, and ambiguous data clusters, the probability of overfitting the classifier increases. Therefore, it is of high relevance to only include those features into the feature set contributing the most to the present classification task. By leaving out features not contributing to the classification task, not only overfitting is prevented, but also the recognition performance may be increased (cf. [Egorow et al. 2018]). Another side effect when utilizing feature selection is that in case of a later real-time application of the model, also the computational effort and, hence, latency of the system can be dramatically decreased, as only the relevant features need to be extracted (cf. Section 2.2.4). This, for instance, would not be the case when utilizing a feature extraction method (i.e. generating new features by combining correlated features).

To perform the feature reduction, I utilized a *wrapper* feature selection method based on the RF recognition performance and feature permutation. This feature selection was performed using the *emobase* benchmark feature set including 988 features originating from 19 functionals applied to 26 Low-Level Descriptors (LLDs) and their deltas (see Table 2.3 on page 39 for more details).

6.1.1 Feature Selection Using Random Forest and Feature Permutation

The approach utilized in the course of this Thesis is a RF based wrapper method (cf. Section 2.2.4). I opt for this approach, as a comparable approach shows promising results in [Egorow et al. 2018], where a feature importance ranking based on the RF recognition performance is utilized and the selected features are later applied on a SVM classifier. Independent of the utilized data set, the reduced feature set shows an increase of recognition performance compared to the original *emobase* feature set, when utilizing 40% to 60% of the original 988 features. This results in an increased UAR by approx. 2% to 3%, compared to the recognition performance when using all the features.

I will now further describe the used feature selection wrapper method, which is a greedy algorithm based on feature permutation and minimizing the Out-Of-Bag (OOB)-error (cf. [Breiman 2001]). The OOB-error corresponds to the error rate obtained on the OOB-observations of the RF. OOB-observations are those observations not included in the training process of the tree because of the bootstrapping procedure used to re-sample the training data (cf. Appendix A). For every individual decision tree ($t = 1, ..., numTrees$, with $numTrees$ being the total number of trees in the forest) of the RF, the OOB-error ($\epsilon(t)$) is computed. At each split of the decision trees, a pre-defined number of features ($numFeatures$) is chosen randomly from the feature set ($Subset(t) \subseteq F$, with F being the set of features). The chosen feature values of each feature used at each split of the individual tree are permuted for each feature successively among all OOB-observations and an updated error rate using the permuted feature vector is calculated ($\tilde{\epsilon}(t, f)$, with $f \in F$ for all $f \in Subset(t)$). By determining the difference between the OOB-error and the updated error rate ($d(t, f) = \tilde{\epsilon}(t, f) - \epsilon(t)$) the influence of this feature on the recognition performance of the considered tree is evaluated. If a change in the prediction error occurs, this indicates an influence of the feature permutation on the model and vice versa. As only the pre-defined $numFeatues$ are used at the split of each tree, in case of features which are not represented at the split of the tree, $d(t, f)$ is set to zero. Finally, the feature importance (I) of feature f is determined by

$$I(f) = \frac{\overline{d}(f)}{\sigma(f)}, \tag{6.1}$$

with $\overline{d}(f)$ being the average value of $d(t, f)$ for each feature f over all trees and $\sigma(f)$ representing the corresponding standard deviation. Using $I(f)$ we can now rank the features according to their importance and only consider those features having a high impact on the present classification task.

6.1.2 Evaluating the Feature Selection

As I wanted to determine the optimal feature set, which performs best also on unknown subjects, it was decided to not apply the presented feature selection approach on the data of all 28 evaluation subjects, but to perform LOSO cross-validation experiments. By doing so, one of the 28 subjects was always unknown in the decision process also resulting in 28 different sets of features ranked by their importance. In case of a high correlation ($r > 0.80$) between all the obtained feature rankings, a simple averaging of the feature importance for each feature would have been sufficient to select those features reaching the highest value and, hence, contributing the most to all LOSO experiments. However, by determining the Pearson correlation-coefficient r of the obtained feature rankings only a moderate positive correlation over all experiments was achieved ($r = 0.60$ (0.01), brackets denote standard deviation). I, therefore, introduced a new procedure of selecting the optimal reduced feature set in [Requardt; Ihme et al. 2020], which will be described in detail hereinafter.

Novel Feature Selection Approach

The novel feature selection approach was based on choosing the optimal feature set from the top 100 features of all LOSO experiments. It was first determined how many features were included in all 28 top 100 feature rankings. This resulted in 15 very highly correlated features being present among all experiments ($r = 0.90$ (0.02)). The low number of consistent features indicated a high diverseness of the important features between the different subjects and, hence, the feature characteristics. It was assumable that these features would most certainly not lead to the best recognition performance. To increase the probability of finding an optimal feature set, the feature selection was extended to generate multiple sets of optimal feature set candidates. These additional sets were generated by successively adding those features, which were consistent in $28-n$ LOSO experiments, with $n = 1, ..., 27$, while maintaining a high correlation of the newly generated reduced feature sets of $r \geq 0.80$. This correlation limit was reached in case of $n = 19$ and, hence, resulted in 19 additional optimal feature set candidates. An overview on all generated feature set candidates, including their number of features and their correlation among all experiments, is presented in Figure 6.2.

At this point it can already be anticipated that the classification results, when utilizing feature set 1 (15 features), did not per se differ significantly from the results obtained when utilizing all 988 features of the *emobase* feature set (cf. Figure 6.4 and Figure 6.5). This may be attributed to the fact that the utilized 15 features are

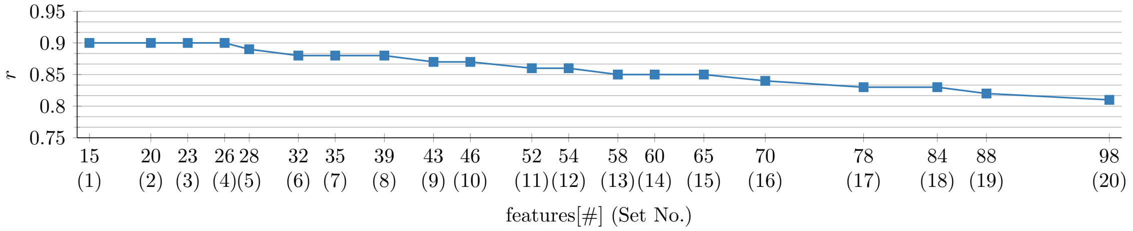


Figure 6.2: Correlation coefficient r of the optimal feature set candidates. Brackets denote the corresponding feature set numbers (adapted from [Requardt; Ihme et al. 2020]).

assumed to be those features contributing most to the present classification task. Therefore, it was assumed that this low number of utilized features was sufficient to obtain a comparable recognition performance, as when utilizing all features, while reducing the number of features to 1.52% of the original feature set. Further emotion recognition experiments, utilizing RF and SVM classification models, and statistical analysis (repeated-measures ANOVA) of the obtained recognition results for the different optimal feature set candidates, are presented in Section 6.3. This will also include results with regard to the influence of hyper-parameter optimization, as introduced in the next Section.

6.2 Improving the Recognition Performance using Hyper-Parameter Optimization

To further push the speech emotion recognition system to its performance limits it is inevitable to perform a hyper-parameter optimization. Especially in case of real-world data samples, which show a high variation in their feature characteristics and, hence, are more spaciously distributed among the search space (i.e. ambiguous class boundaries), a hyper-parameter optimization can lead to significant improvement of the recognition performance (cf. Section 2.2.8). Therefore, a random search was performed to optimize the relevant hyper-parameters of the designed RF and SVM classification models. This was done by randomly choosing ten parameter combinations from a pre-defined discrete search interval. To validate these parameter combinations, LOSO cross-validation experiments were performed and the parameter combination performing best among all LOSO-experiments was chosen as the optimal one. This was done individually for each optimal feature set candidate generated in the previous Section. In the following, I first present how the search interval of the random search was determined. Afterwards, the results of the LOSO cross-validation experiments are presented, first averaged over all generated feature sets, and later individually for each utilized feature set.

6.2.1 Determining the Search Interval

The search intervals of the RF and SVM hyper-parameters were determined separately and individually for each considered hyper-parameter. As hyper-parameters the most commonly used parameters, as introduced in Appendix A and Section 2.2.8, were utilized. The search intervals of both classification approaches were based on the heuristic choices presented in the literature (cf. [Hastie et al. 2009; Liaw & Wiener 2002] and [Hsu et al. 2016]).

In case of a RF classifier, relevant hyper-parameters are the number of trees included in the forest ($numTrees$) and the size of the feature set used at each split of the decision trees ($numFeatures$). A detailed description on how these two parameters influence the recognition performance and generalizability of the classifier is stated in Appendix A. Considering the recommendations presented there, the parameter interval of $numTrees$ was specified ranging from 10 - 1000 trees, which is comparable to the experimental setup presented in [Probst et al. 2019]. The interval of $numFeatures$ was chosen based on the recommendations made in [Hastie et al. 2009]. Depending on, whether the designed classifier is used to solve a classification or a regression problem, the authors recommend to use either $numFeatures = \sqrt(p)$, with p being the total number of features inside the feature set, or $numFeatures = p/3$, respectively. Considering the feature selection presented in the previous section, the utilized feature sets were comparatively low. With a maximum number of 98 features included in the largest set (cf. Figure 6.2 on page 170), choosing $numFeatures < \sqrt(p)$ would not be reasonable. Therefore, the lower limit of the search interval was chosen as $\sqrt(p)$. The upper limit of the search interval was chosen based on the assumption made in [Breiman 2001], where it is stated that for regression problems a larger number of $numFeatures$ is needed. Bearing in mind that the recommendations made in [Hastie et al. 2009] are only valid in case of ideal class separability, I extended the search space up to $numFeatures = p/2$. As the size of the utilized feature set varies from $p = 15$ up to $p = 98$ features, also the search interval changes. By generating a random seed, ten integer values lying in between the upper and lower limit of the search intervals of $numTree$ and $numFeatures$ were generated. This resulted in ten optimal parameter combination candidates as stated in Table 6.1. The numeration of the parameter sets was not chosen randomly, but in increasing order of the F1-Measure obtained during the later evaluation of the associated hyper-parameter combinations (cf. Table 6.2 and Table 6.3 on pages 175 and 178). In case of $numFeatures$, the percentage split of randomly chosen features out of all features in the set, averaged over all feature sets, is given. Additionally to the randomly generated parameter combinations, all combinations of the upper and lower limits of both search intervals were evaluated. With regard to the expected computational costs arising when performing a hyper-parameter optimization in combination with choosing an optimal feature set and LOSO cross-validation (cf. Figure 6.1 on page 166), evaluating a

Table 6.1: Optimal parameter set candidates of the RF and SVM classifier generated using random search on the pre-defined search intervals (adapted from [Requardt; Ihme et al. 2020]). In case of RF, the relevant parameter combinations correspond to *numTrees* and *numFeatures*, whilst for SVM the relevant parameter combinations correspond to C and γ.

	Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
RF	*numTrees* [#]	10	10	61	134	555	1000	516	215	711	446	898	298	1000	894
	numFeatures [%]	16	50	24	16	16	16	25	35	32	30	36	39	50	40
SVM	C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{-5}	2^{1}	2^{-5}	2^{7}	2^{3}	2^{15}	2^{7}	2^{1}
	γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-15}	2^{-3}	2^{-11}	2^{-13}	2^{-11}	2^{-15}	2^{-11}	2^{-5}

larger number of hyper-parameter combinations would go beyond the scope of this Thesis.

In case of a SVM, the classification algorithm is based on finding a hyperplane in the transformed feature space, which linearly separates two classes. As the utilized data samples contain real-world speech data recorded under disturbed environmental conditions, it was assumed that standard SVMs would not lead to satisfactory classification results. To also achieve reasonable results in case of non-linearly separable data samples, there exist ways to adapt a standard SVM. This can be done by introducing a soft margin using the cost-value C and by utilizing the so-called *kernel-trick* (cf. Appendix A). The cost-value C penalizes those samples lying inside and on the wrong side of the margin. Depending on the used kernel, different parameters are used for hyper-parameter optimization. The most sophisticated kernel is the Radial Basis Function (RBF)-kernel as presented in Equation (A.22) on page 259. For this kernel, the only relevant hyper-parameter is γ, which affects the width of the Gaussian.

For the recognition experiments performed in this Chapter, I opt for a SVM with a soft margin and a RBF-kernel. The parameters C and γ were therefore considered as relevant hyper-parameters for the present classification task. Unlike the parameter *numFeatures* of the RF, the parameters of the SVM were chosen independent from the utilized feature set. Therefore, the parameter values presented were identical for all considered feature sets. However, in [Steinwart & Christmann 2008] it is stated that the scaling of the kernel parameter γ has the same effect on the classifier as scaling the input space. With an additional change in the dimensionality of the considered feature space, this implies that with a change in the input space also the influence of γ changes. I, therefore, assumed that the optimal parameter combination is also dependent on the chosen feature set.

To identify the search space of C and γ, the findings made in [Hsu et al. 2016] were applied. In this work, the authors present an optimization based on grid-search. By conducting various cross-validation experiments they identify a discrete search interval of

$$C = [2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13}, 2^{15}]$$

and

$$\gamma = [2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-7}, 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}].$$

The optimal hyper-parameter candidates were randomly taken from these intervals. In this process repetitions of the parameter values were allowed. As for the RF, ten parameter combinations were chosen randomly and four combinations were generated using the upper and lower limits of both search intervals. The resulting 14 optimal parameter set candidates are stated in Table 6.1.

6.2.2 Evaluating Optimal Parameter Combinations

The optimal hyper-parameter candidates, as presented in Table 6.1 on page 172, were validated using the LOSO cross-validation scheme. This was done for each feature set separately and resulted in 7840 experiments performed for each considered classification approach (cf. Figure 6.1 on page 166). The computational costs arising during this process of training and testing the individual classification models are presented in Figure 6.3. The computational costs were summed up for each feature set (i.e. 28 LOSO-runs and 14 parameter sets). A clear relation between the size of the feature set and the computational cost of the classifier is noticed (see Figure 6.2 on page 170 for feature set sizes). In case of the RF classifier the computational cost grows with the increase of the feature set. This can be attributed to the increase of features from which the classifier can choose at each split of the decision trees. For the SVM classifier a strong decrease of computational cost with an increase of features is noticed. This may be due to the computational effort of finding a separating hyperplane, which is easier for high-dimensional spaces (but often leads to lower generalization). Already anticipating the recognition performances presented in Figure 6.4 and Figure 6.5, it can be stated that the high computational costs obtained with small feature sets were accompanied by lower UARs, indicating that the determined separating hyperplanes are less suitable to clearly divide the feature

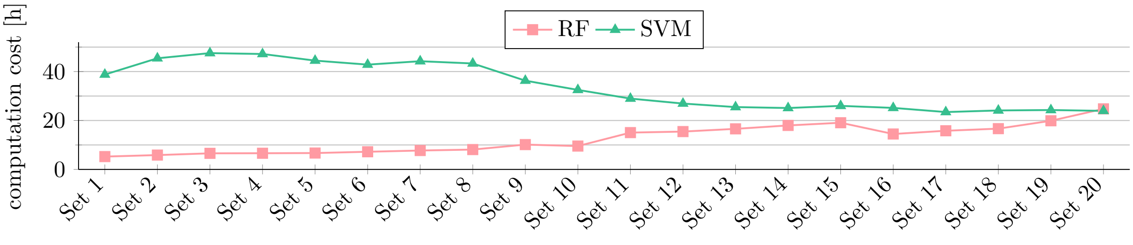


Figure 6.3: Computational costs of performing a hyper-parameter optimization on the SVM and RF classifier, utilizing a LOSO cross-validation for each identified feature set. With an increase of the feature set number also the size of the feature set increases.

space into clusters of the four considered emotion categories. Considering the total amount of computational cost, the optimization of the SVM took an overall training and testing time of over 675 hours, while the corresponding RF optimization took only 250 hours. However, with the last considered feature subset (Set 20) the courses of the computational costs of both classification approaches intersect. To confirm this trend, further feature subsets need to be generated and similar classification experiments need to be performed. This, however, was not done in the scope of this Thesis.

The obtained recognition performances were afterwards used to identify if there existed a distinct hyper-parameter combination which outperformed all others, regardless of the utilized feature set. As performance measures the UAR, UAP and corresponding F1-measure, as presented in Section 2.2.7 were utilized. As described earlier in this Section, the size of the feature set influences the search interval of the hyper-parameter optimization. Consequently, it is assumable that an optimal parameter combination needs to be chosen individually for each considered feature set. The results will be presented separately for the utilized classification approaches.

Parameter Optimization for the RF Classifiers

To receive an overall insight on how the different parameter candidates performed on the recognition task, the performance measures were averaged over all feature sets. In case of the F1-measure, the metric was determined for each feature set separately and later averaged over all feature sets (macro-average)[1]. The obtained values are stated in Table 6.2. The best results regarding the individual performance measures are marked in green color. The first column refers to the considered optimal parameter set candidates, as defined in Table 6.1 on page 172, and was chosen in increasing order of the obtained macro-averaged F1-measure, as this measure takes into account the trade-off between UAR and UAP. For all performance measures values above 25% were achieved, which indicates a recognition performance above chance level for the considered four-class emotion recognition problem. Brackets denote the standard deviation. It can be seen that for the UAR a very low standard deviation, ranging from 0.44 to 0.68 was achieved. This indicated, that the results obtained for the individual feature sets did not show a strong variation. In case of UAP and F1-measure a higher variation among the feature sets was noticed (indicated by the higher standard deviation). With regard to the results stated in Table 6.2, it was noticed that parameter sets 1 and 2 showed a considerably lower UAP and F1-measure compared to all other parameter sets. Therefore, it was assumed that these parameter sets do not serve as optimization candidates (marked in red color). The later presented detailed statistical analysis of the results will further reveal the unsuitability of parameter set 3.

[1] A different result would be obtained when determining the micro-average (e.g. calculating the F1 based on the already averaged UAR and UAP values)

Table 6.2: UAR, UAP and F1-measure of a 4-class driving condition problem averaged over all investigated feature sets for the different parameter combinations (Par. Set) of the RF classifier with ascending F1-measure (macro-averaged). Brackets denote the standard deviation. The best results regarding the individual performance measures are marked in green color. Red entries indicate parameter sets which were identifies as unsuitable during the statistical analysis.

Par. Set	UAR [%]	UAP [%]	F1[%]
1	31.86 (0.68)	33.97 (1.20)	32.87 (0.88)
2	32.01 (0.67)	33.86 (0.82)	32.91 (0.70)
3	32.10 (0.56)	41.88 (1.84)	36.34 (1.00)
4	31.73 (0.51)	43.01 (2.26)	36.50 (1.10)
5	31.48 (0.48)	43.70 (2.23)	36.59 (1.07)
6	31.41 (0.44)	43.94 (2.32)	36.61 (1.07)
7	31.67 (0.55)	44.00 (2.82)	36.81 (1.31)
8	31.99 (0.48)	43.45 (2.05)	36.83 (1.01)
9	31.83 (0.57)	44.13 (2.79)	36.96 (1.34)
10	31.86 (0.54)	44.08 (2.57)	36.96 (1.21)
11	31.80 (0.55)	44.22 (2.52)	36.98 (1.22)
12	31.92 (0.59)	43.98 (2.09)	36.98 (1.10)
13	31.92 (0.55)	44.39 (2.24)	37.12 (1.13)
14	31.89 (0.60)	44.59 (2.59)	37.16 (1.27)

The results obtained for each individual feature set (averaged over all subjects) are stated in Appendix D (see page 288 ff.). The Tables clearly indicate that, for most of the considered parameter sets, with an increased number of features also the performance increased up to a certain value. An exception was noticed in case of parameter sets 1 and 2, for which the UAR changed irregularly, also supported by its higher standard deviation in Table 6.2, and for which the UAP achieved noticeably lower values compared to all other parameter sets. Considering the associated parameter values of $numTrees$ and $numFeatures$, the number of trees of sets 1 and 2 both correspond to 10 trees. Therefore, it was assumed that this low number of trees is insufficient for a reliable classification, as they will not lead to a sufficient generalizability of the classifier (cf. Appendix A).

The assumptions made above were also confirmed by the results obtained from a repeated-measures Analysis of Variance (ANOVA) performed on the UAR, UAP and F1-measure over all considered parameter sets. This revealed a highly significant effect of the chosen parameter set (par. set) for all performance measures (main effect par. set UAR: $F(4.9,93.0) = 8.38$, $p < 0.01$, Greenhouse-Geisser-corrected; main effect par. set UAP: $F(5.2,98.2) = 148.79$, $p < 0.01$, Greenhouse-Geisser-corrected; main effect par. set F1: $F(5.2,98.3) = 115.25$, $p < 0.01$, Greenhouse-Geisser-corrected). By performing post-hoc paired t-tests on the F1-measure, I was able to identify three parameter sets which showed significant differences in their performance. These were parameter sets 1, 2 and 3. While the averaged performances of sets 1 and 2 were highly significantly lower than for all other sets (all p's < 0.01, Bonferroni-corrected), set 3 showed a significantly lower performance than six other sets (set 3 vs. sets [9, 10, 11, 12, 13 & 14], p's < 0.05, Bonferroni-corrected) and a higher performance than sets 1 and 2 (set 3 vs. sets [1, 2], p's < 0.05, Bonferroni-corrected). A similar significant effect was obtained when performing post-hoc paired t-tests on the UAP with sets 1 and 2 showing a highly significantly lower performance than all other sets (all p's < 0.01, Bonferroni-corrected) and set 3 showing a significantly lower performance than 10 other sets (set 3 vs. sets [5, 6, 7, 8, 9, 10, 11, 12, 13 & 14], p's < 0.05, Bonferroni-corrected). The results obtained on the F1-measure and UAP were, however, in conflict with the post-hoc results obtained on the UAR. Here, sets 1 and 2 showed no noteworthy significant differences and set 3 even outperformed seven other sets (set 3 vs. sets [4, 5, 6, 7, 9, 10 & 11], p's $<$ 0.05, Bonferroni-corrected). With regard to the present challenging recognition task with comparatively low recalls, it is of high importance to achieve a high precision, as this indicates a low number of false positives (cf. Table 2.5 on page 47 and Equation (2.14) on page 46) and, hence, a high trustworthiness of the identified emotional state. Considering the parameter values of set 3 it was further noticed that this set comprised a comparatively low value for *numTrees*. It was, therefore, assumed that the obtained RF classification model is not sufficiently generalizable, implying that the (on average) higher UAR could also occur by chance. This is also in line with the results presented in [Oshiro et al. 2012], where it is stated that for a good recognizer performance more than 100 decision trees are needed.

From the obtained results of the statistical analysis, three parameter sets (1, 2 and 3), which were identified as non-optimal parameter candidates, were discarded. However, it was not possible to identify one parameter set, which outperformed all other sets.

To verify this result, an additional repeated-measures ANOVA was performed on the remaining 11 parameter sets. For all performance measures, a significant effect of the parameter set was revealed (main effect par. set UAR: $F(4.4,83.3) = 14.25$, $p < 0.01$, Greenhouse-Geisser-corrected; main effect par. set UAP: $F(5.0,94.2) = 2.59$, $p < 0.05$, Greenhouse-Geisser-corrected; main effect par. set F1: $F(4.5,85.5) = 3.51$, $p < 0.01$, Greenhouse-Geisser-corrected). However, for the UAP and F1-measure this effect was noticeably lower compared to the results obtained when utilizing all parameter sets. Additional post-hoc paired t-tests revealed that there exists no parameter set, with an overall significant higher or lower performance in UAP or F1-measure, compared to all other parameter sets. Only considering the UAR, the post-hoc t-tests revealed that the sets 5 and 6 performed significant worse than a majority of the other sets (set 5 vs. sets [4, 8, 9, 10, 11, 12, 13 & 14], p's < 0.05, Bonferroni-corrected, and set 6 vs. sets [4, 7, 8, 9, 10, 11, 12, 13 & 14], p's

< 0.05, Bonferroni-corrected). However, especially when considering the trade-off between UAR and UAP, an overall significant effect in the F1-measure is reasonable when choosing the optimal parameter combination. Therefore, no further parameter sets were excluded from the list of potential candidates and the best performing parameter set was chosen independently for each feature set. At this point, it can already be disclosed that the performances of the different parameter sets for the individual feature sets did not differ significantly. The best performing parameter set for each feature set was chosen based on the micro-averaged F1-measure. The results of this feature set dependent LOSO cross-validation is presented in Section 6.3.

Parameter Optimization for the SVM Classifiers

Similar analyses were performed for the SVM classification models. Analogously as for the RF classification models, Table 6.3 gives an overview on the recognition performances (UAR, UAP and F1-measure), averaged over all feature sets. Green entries correspond to the highest value obtained for the individual performance measures. It is clearly noticeable that the first three parameter sets did not lead to a performance above chance level. This implies that the classifier was unable to differentiate between the individual emotional states. For these parameter sets almost all test samples were assigned to the same emotional state over all LOSO experiments, namely *neutral*. Therefore, these parameter combinations were excluded from the list of potential candidates. When considering the corresponding parameter values as stated in Table 6.1 on page 172, it can be seen that the value of γ was assigned to the highest possible value of the search interval ($\gamma = 2^3$), while the value of C was assigned as well to the highest, as to the lowest possible value of the search interval ($C = [2^{-5}, 2^{15}]$). This already gave a first evidence towards the choice of γ being of higher relevance than the choice of C. In case of the parameter sets 4 and 5, slightly higher performance measures were obtained, respectively. They were, however, outperformed by all other sets. Therefore, it was assumed that the parameter sets 1, 2, 3, 4 and 5 do not serve as optimal parameter candidates (marked in red color). Furthermore, it was noticed that, in contrast to the RF classifier, the range of the averaged performance measures was considerably wider. While the difference in between the highest and lowest values of the RF classifier equated to $\Delta UAR = 0.69\%$, $\Delta UAP = 10.73\%$ and $\Delta F1 = 2.29\%$, the difference in case of the SVM was considerably higher with $\Delta UAR = 6.21\%$, $\Delta UAP = 30.38\%$ and $\Delta F1 = 19.72\%$.

As for the RF approach, the results obtained for each individual feature set (averaged over all subjects) are stated in Appendix D (see page 291 ff.). For each parameter set a uniform increase and decrease of the individual performance measures was observed. While for most parameter sets the recognition performance increased with an increase of features included in the set, in case of parameter set 8 a decrease of performance was noticed. The utilized heat map in Appendix D visualizes the

Table 6.3: UAR, UAP and F1-measure of a 4-class driving condition problem averaged over all investigated feature sets for the different parameter combinations (Par. Set) of the SVM classifier with ascending F1-measure (macro-averaged). Brackets denote the standard deviation. The best results regarding the individual performance measures are marked in green color. Red entries indicate parameter sets which were identifies as unsuitable during the statistical analysis.

Par. Set	UAR [%]	UAP [%]	F1[%]
1	25.00 (0.01)	11.99 (0.68)	16.20 (0.58)
2	25.00 (0.01)	11.99 (0.68)	16.20 (0.58)
3	25.00 (0.01)	11.99 (0.68)	16.20 (0.58)
4	25.06 (0.16)	14.03 (4.10)	17.71 (2.92)
5	25.57 (0.41)	23.23 (6.73)	23.85 (4.13)
6	27.59 (0.66)	32.79 (1.94)	29.95 (1.12)
7	29.13 (1.23)	33.00 (1.95)	30.94 (1.49)
8	28.07 (0.78)	35.66 (2.66)	31.39 (1.46)
9	29.87 (0.80)	34.70 (3.17)	32.07 (1.78)
10	30.20 (0.91)	38.30 (5.37)	33.67 (2.64)
11	30.28 (0.99)	38.55 (5.48)	33.82 (2.74)
12	30.45 (0.81)	38.88 (4.40)	34.09 (2.20)
13	30.50 (0.87)	39.61 (4.66)	34.39 (2.34)
14	31.21 (0.72)	42.37 (2.56)	35.92 (1.34)

performance weakness of parameter sets 1, 2, 3, 4 and 5, which is in line with the averaged results presented in Table 6.3.

Even though the unsuitability of parameter sets 1, 2, 3, 4 and 5 seems evident. This assumption was confirmed by the results obtained from a repeated-measures ANOVA. This revealed a highly significant effect of the parameter sets on all performance measures (main effect par. set UAR: $F(1.2,13.6) = 303.69$, $p < 0.01$, Greenhouse-Geisser-corrected; main effect par. set UAP: $F(1.5,27.6) = 193.50$, $p < 0.01$, Greenhouse-Geisser-corrected; main effect par. set F1: $F(1.4,27.2) = 257.55$, $p < 0.01$, Greenhouse-Geisser-corrected). Post-hoc paired t-tests further revealed that sets 1, 2, 3 and 4 performed highly significant lower than the remaining ten sets for all three performance measures (all p's < 0.01, Bonferroni-corrected). In case of parameter set 5, a significant higher performance compared to sets 1, 2, 3 and 4 was obtained (all p's < 0.01, Bonferroni-corrected). Nevertheless, this set performed significant lower than the remaining nine sets (all p's < 0.01, Bonferroni-corrected). This is in line with the assumptions made previously, of sets 1, 2, 3, 4 and 5 not serving as optimal parameter candidates. It was further noticed that there exist three sets which outperformed a majority of the other sets for all performance measures, these were sets 12, 13 and 14. In case of set 14, it was even shown that this set performed significant higher than all other sets (all p's < 0.05, Bonferroni-corrected), and vice versa sets 12 and 13 performed significantly lower than set 14 (all p's $<$ 0.05, Bonferroni-corrected). Considering the post-hoc results obtained from the F1-measure, set 12 performed significant higher than nine other sets (set 12 vs. sets [1, 2, 3, 4, 5, 6, 7, 8 & 9], p's $<$ 0.05, Bonferroni-corrected) and set 13 significant higher than 11 other sets (set 13 vs. sets [1, 2, 3, 4, 5, 6, 7, 8, 9, 10 & 12], p's $<$ 0.05, Bonferroni-corrected).

From the obtained results of the statistical analysis, five parameter sets (sets 1, 2, 3, 4 and 5), which were identified as non-optimal parameter candidates, were discarded. Furthermore, it was possible to identify three parameter sets, which outperformed a majority of the other sets (sets 12, 13 and 14).

These results were verified by repeating the statistical analysis on the remaining nine parameter sets. As before, a highly significant effect of the parameter sets on the different performance measures was revealed (main effect par. set UAR: F(1.3,25.5) = 73.13, p $<$ 0.01, Greenhouse-Geisser-corrected; main effect par. set UAP: F(1.6,29.9) = 25.41, p $<$ 0.01, Greenhouse-Geisser-corrected; main effect par. set F1: F(1.6,29.8) = 36.62, p $<$ 0.01, Greenhouse-Geisser-corrected). In contrast to the RF approach, a noticeable higher effect on all performance measures, compared to the results obtained when utilizing all parameter sets, was obtained. This indicates that, with regard to the remaining parameter sets, the RF classifier is more robust against differences in the chosen parameter set than the SVM classifier. The post-hoc t-tests further revealed a consistent impact of the parameter sets on the different performance measures. For all considered measures, set 14 showed a significant higher performance compared to all other parameter sets (all p's $<$ 0.05, Bonferroni-corrected). Additionally, three other sets (sets 11, 12 and 13) performed significantly better than at least three other sets (sets 6, 7 and 9) for all considered performance measures (all p's $<$ 0.05, Bonferroni-corrected). Furthermore, for sets 6 and 7, a significant lower performance compared to a majority of the remaining sets was achieved (set 6 vs. sets [7, 9, 10, 11, 12, 13, & 14], p's $<$ 0.05, Bonferroni-corrected; set 7 vs. sets [9, 10, 11, 12, 13, & 14], p's $<$ 0.05, Bonferroni-corrected). It can be anticipated at this point that these two sets were also never chosen as optimal parameter set during the later feature set dependent evaluation in Section 6.3.

Considering the presented results, it was possible to draw conclusions on the influence of the cost parameter C and the kernel-parameter γ on the recognition performance of the SVM classifier. By evaluating the number of times a certain parameter set was chosen for the individual feature sets, it was noticed, that in cases of small feature sets ($\leq$ 20 features), a $\gamma \leq 2^{-3}$ was needed to obtain reasonable classification results. The only exception was noticed in case of $\gamma = 2^{-7}$ (set 6), which performed significant lower than seven other sets (see above). With an increased number of features also γ decreased. For feature set sizes in between 23 and 70 features a $\gamma \leq 2^{-5}$, and sizes from 78 features onwards a $\gamma \leq 2^{-7}$ was needed, respectively. However, these results only describe tendencies, which could not be confirmed by statistical analysis on the individual feature sets (see Section 6.3). Hence, as for the RF classifier, it is recommended to choose the parameter set individually for each feature set. During the later performance validation of the different feature sets, it was further possible to confirm the assumption made before, which state that cost-parameter C has a much lower influence on the performance of the classifier than the parameter γ.

6.2.3 Findings on Hyper-Parameter Optimization

To sum up, in case of the *RF classification approach*, no optimal parameter set which outperformed all other sets could be identified. A repeated-measures ANOVA revealed that the choice of the parameter set has a significant effect on the considered performance measures. However, this effect was mainly caused by three parameter sets performing significantly lower than a majority of the other sets. These were sets 1, 2 and 3, which were excluded for the further evaluation of the individual feature sets. For these sets, a comparatively low number of $numTrees$ was chosen, for which it can be assumed that the classifier is not generalizable. Furthermore, there exists a strong interdependency between the chosen hyper-parameters $numTrees$ and $numFeatures$, and a dependency between the parameter values and the size of the feature sets. This was also confirmed by the results obtained after excluding parameter sets 1, 2 and 3 from the statistical analysis. For this case, it was not possible to identify parameter sets, which outperformed a majority of the other sets. Therefore, I opt to choose the parameter sets individually for each considered feature set. The optimal parameter candidates which are used for this feature set dependent evaluation in Section 6.3 are parameter sets 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14.

However, it needs to be stated that not finding an optimal set which outperformed the other sets, may also be justified by the low number of evaluated parameter sets, which did not cover the whole search space. Furthermore, by considering the Tables presented in Appendix D, a much stronger influence of the chosen feature set, compared to the choice of the parameter set, was identified. The visualized heat map in Appendix D shows that with an increased size of the feature set also the classification performance increases. A further evidence of the strong influence of the feature set is the comparatively low range of the averaged performance measures presented in Table 6.2 on page 175. Especially after the exclusion of the parameter sets 1, 2 and 3, the difference of the mean values for each performance measure equated to 0.58% for UAR, 1.58% for UAP, and 0.66% for F1, accompanied with only small variations in the standard deviation.

In case of the *SVM classification approach*, by utilizing the identical evaluation procedure, it was possible to identify five parameter sets which were significantly outperformed by a strong majority of the remaining sets. These were sets 1, 2, 3, 4 and 5. With regard to the associated hyper-parameters, it was noticed that for these parameter sets the highest values of γ were utilized, while the cost-parameter C would range in between the highest and lowest possible value of the search interval ($C = [2^{-5}, 2^{15}]$). At this point it was already anticipated, that for an optimal performance of the classifier, an increase of the size of the feature set needed to be accompanied with a decrease of γ. The influence of the second optimization parameter C, however, was much lower, respectively. It was further possible to identify three parameter sets which would outperform a majority of the other parameter sets. These were sets 12, 13 and 14. This effect was also observed when excluding sets 1, 2, 3, 4 and 5 from the statistical analysis. This two stage approach was chosen to confirm the statistics when excluding "outliers". Additionally, this second statistical analysis further revealed two more parameter sets (sets 6 and 7) which were outperformed by a majority of the remaining sets leading to seven parameter sets which are used for the feature set dependent evaluation in Section 6.3 (sets 8, 9, 10, 11, 12, 13 & 14). In contrast to the RF classifier, the heat maps in Appendix D show that the recognition performance is strongly dependent on both, the chosen parameter set, and utilized feature set. This is also reflected by the comparatively high differences of the averaged UAR, UAP and F1-measures presented in Table 6.3 on page 178, which equate to 3.62%, 9.58%, and 5.97%, respectively (after exclusion of parameter sets 1, 2, 3, 4 and 5).

6.3 Classifying Drivers' Emotions

Until now, only the influence of the chosen hyper-parameter set on the recognition performance of the individual classifiers has been addressed. It was identified that the effect of the chosen parameter sets on the average recognition performance of the individual feature sets was, in most cases, not significant. Therefore, it was stated that the optimal parameter set needs to be chosen separately for each individual feature set. I will first evaluate if there exists a feature set which outperformed the other sets. Afterwards, I will validate the performance of the classifier by presenting the emotion-wise results of the best performing LOSO cross-validation experiments. This was done individually for the considered classification approaches.

6.3.1 The Influence of the Feature Set

The evaluation of the feature set was performed by utilizing the LOSO cross-validation results obtained when performing a hyper-parameter optimization for each feature set separately. The UAR and UAP, averaged over all LOSO experiments are stated in Figures 6.4 and 6.5, respectively. The parameter set was chosen based on the micro-averaged F1-measure. Additionally to the optimized results obtained for the individual feature sets, baseline results using all features of the *emobase* feature set are presented. The baseline results were obtained by utilizing a

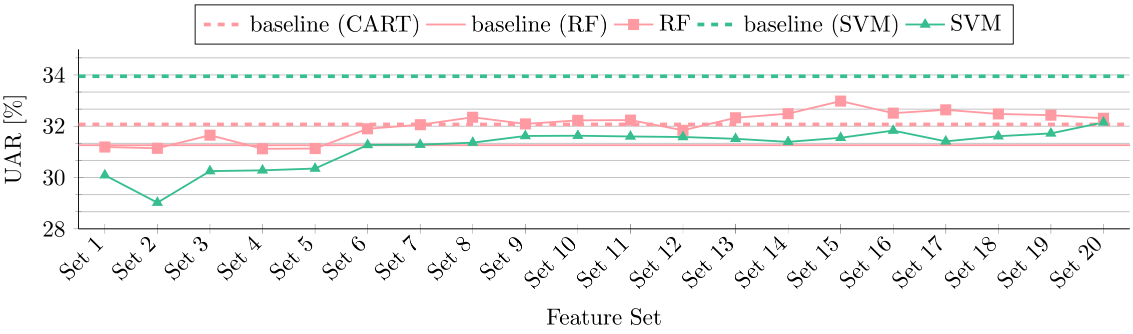


Figure 6.4: Mean UAR of the optimized RF/ SVM classifiers for each individual feature set.

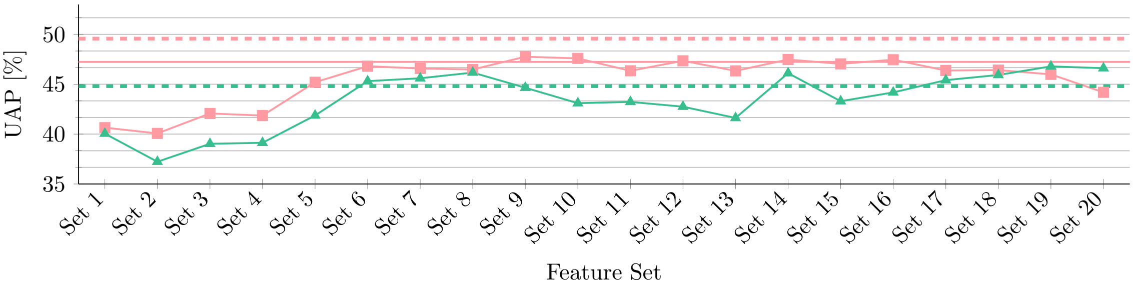


Figure 6.5: Mean UAP of the optimized RF/ SVM classifiers for each individual feature set.

standard non-optimized Classification and Regression Trees (CART) approach (cf. [Breiman et al. 1984]) with a default number of 200 decision tress in the forest (red dashed line), an optimized RF (red solid line) and an optimized SVM (green dashed line), using the remaining parameter sets of the previous Section 6.2.

Feature Set Analysis for the RF Classifiers

To identify if there exists a feature set having a significant effect on the recognition performance of the classifier, a repeated-measures ANOVA was performed using the results obtained from the optimized LOSO cross-validation experiments. In case of the UAR, the ANOVA revealed that the feature sets highly significantly affected the recall of the experiments (main effect feature set UAR: F(7.2,195.7) = 4.24, $p < 0.01$, Greenhouse-Geisser-corrected). By conducting a post-hoc paired t-test, one feature set was identified, which outperformed nine other sets and the baseline (RF) classifier, this was set 15 (set 15 vs. sets [baseline (RF), 2, 3, 4, 5, 6, 7, 9 & 12], p's < 0.05, Bonferroni-corrected). It was further noticed that especially small feature sets could not reach the performance of the baseline classifiers. This effect was even significant considering sets 1, 4, 5 and the baseline (RF) classifier in contrast to sets 14 to 18 (all p's < 0.05, Bonferroni-corrected). As for the UAR, a significant effect of the feature sets on the UAP was revealed (main effect feature set UAP: F(7.3,196.3) = 2.42, $p < 0.05$, Greenhouse-Geisser-corrected). However, from the post-hoc t-tests it was not possible to identify one feature set, which would outperform a majority of the other sets. Nevertheless, it was possible to identify two sets, which were outperformed by a majority of the considered other sets. These were sets 1 and 2 (set 1 vs. sets [baseline (CART), 7, 8, 9, 12, 14 & 16], p's < 0.05, Bonferroni-corrected; set 2 vs. sets [baseline (CART), 7, 9, 10, 11, 12, 13 & 15], p's < 0.05, Bonferroni-corrected).

Based on the results of the statistical analysis, I opted for feature set 15 as optimal feature set, as the recall obtained when utilizing this set was not only noticeable higher compared to all other sets but even significant. This was also in line with the statement made in the previous Section, of an increased performance of the classifier with an increased feature set size. Considering the Tables in Appendix D (see page 288 ff.), this effect was even independent of the chosen parameter set. This also substantiates the statement that, in case of a RF classifier, a feature selection is of higher importance when it comes to boosting the classifier's performance, compared to a parameter optimization for the individual feature sets.

Feature set 15 includes all features of the *emobase* feature set, which were ranked in the top 100 feature ranking of at least 14 LOSO experiments (cf. Section 6.1.2). From Figure 6.2 on page 170 it is known that set 15 contains 65 features of the original *emobase* set, which equates to 6.6% of the original features. It was noticed that the most frequently occurring LLDs (cf. Table 2.3 on page 39) were related to the Mel-Frequency Cepstral Coefficient (MFCC) No. 1 and 2 (7 and 12 features, respectively), and the Line Spectral Pair (LSP) of the Linear Predictive Coding (LPC)-coefficients No. 1, 3 and 7 (7 features, 6 features and 7 features, respectively). These features all belong to the feature category of *spectral features*, introduced in Section 2.2.2. Furthermore, all these LLDs, except for LSP frequency No. 7, describe harmonics close to the fundamental frequency F0 (lower frequency range). Only eight features included in feature set 15 did not belong to the *spectral features* category, but were related to loudness (1 feature), intensity (1 feature), Zero-Crossing-Rate (ZCR) (5 features) and probability of voicing (1 feature). It was further possible to identify those functionals, which were applied the most frequently to the LLDs. These were the arithmetic mean (7 occurrences), minimum and maximum value (6 occurrences each), quartiles (16 occurrences) and inter-quartile ranges (12 occurrences).

Feature Set Analysis for the SVM Classifiers

Similar evaluations were performed on the results obtained from the LOSO cross-validation experiments for the SVM classifier. The feature set dependent optimized results are presented in Figure 6.4 and 6.5. In case of the UAP, the baseline (SVM) classifier clearly outperforms all other feature sets. This was confirmed by the results obtained from the statistical analysis (main effect feature set UAR: F(5.5,147.4) = 8.69, $p < 0.01$, Greenhouse-Geisser-corrected). A post-hoc paired t-test revealed that the baseline (SVM) classifier would reach significantly higher UAR values compared to all other feature sets, except for feature set 16 and 20 (all p's $<$ 0.05, Bonferroni-corrected). As second best feature set, set 20 outperformed five other feature sets. These were sets 1 to 5 (all p's $<$ 0.05, Bonferroni-corrected). It was further possible to identify one feature set, which performed significantly worse than a majority of the other sets. This was set 2, which was outperformed by all other sets, except feature set 4, and the baseline (SVM) classifier (all p's $<$0.05, Bonferroni-corrected). In case of the UAP, again a significant effect of the feature sets was revealed (main effect feature set UAP: $F(5.8,155.6) = 2.93$, $p < 0.05$, Greenhouse-Geisser-corrected). The observed effect was, however, much lower compared to the effect on the UAR. Here, only one set, performing significantly worse than the others, was identified. This was feature set 2 (set 2 vs. sets [baseline (SVM), 6, 7, 8, 9, 10, 14, 16, 18, 19 & 20], all p's $<$ 0.05, Bonferroni-corrected). As it was not possible to identify one feature set which significantly outperformed the other sets, the best performing set was chosen. In both cases, UAR and UAP, this was feature set 20, which included 98 features (9.9%) of the original *emobase* set, which were ranked in the top 100 feature ranking of at least 9 LOSO experiments (cf. Section 6.1.2 and Figure 6.2 on page 170).

In comparison to feature set 15, chosen as optimal feature set for the RF classifier, feature set 20 includes 84 features belonging to the *spectral features* category (MFCCs and LSP of the LPC-coefficients) and 9 features related to the ZCR. The remaining features cannot be grouped into meaningful categories and were related to loudness (1 feature), intensity (2 features) and probability of voicing (1 feature). The most frequently applied functionals were the arithmetic mean (9 occurrences), minimum value (8 occurrences), maximum value (10 occurrences), quartiles (24 occurrences) and their inter-quartile ranges (17 occurrences).

Findings on Feature Selection

From the evaluation of the individual feature sets it was possible to identify one optimal feature set per utilized classifier. As expected, the optimal feature set of the SVM classifier (feature set 20) did not agree with the optimal feature set of the RF classifier (feature set 15). This was assumable, as the two investigated classification approaches differ strongly in their learning algorithm. Therefore, it is also possible that the different approaches are in need of individual feature sets to achieve the highest possible performance. Furthermore, the utilized feature selection method was a RF-based *wrapper* method, which chooses the relevant features based on the performance of the RF classifier. Nevertheless, this feature selection approach was chosen based on the results presented in [Egorow et al. 2018]. Here, the authors identified, that the number of features needed to outperform the baseline SVM classifier is strongly dependent on the utilized data set and can reach up to 40% of the original feature set. With the largest utilized feature set containing 98 features

(10% of the original *emobase* set), it may be considered as plausible that the optimal number of features needed to achieve above baseline recognition results may not yet be reached. Due to the low correlation of the feature importance over all features included in the *emobase* set ($r = 0.60(0.01)$) and the noticeable decrease of correlation with increasing size of the identified feature sets (from set 1 $r = 0.90$ to set 20 $r = 0.81$), it was not reasonable to also consider larger feature sets.

It was further noticed that the optimal feature sets obtained by evaluating the UAR and UAP were confirmed by the obtained macro-averaged F1-measure. This is not evident, as the F1-measure takes into consideration the trade-off between UAR and UAP, which may deviate strongly from the individual UAR and UAP results, especially, when the individual results drift apart. For both classifiers the best F1-measure was obtained when utilizing the corresponding optimal feature set (see Table 6.11 on page 194). The Table comprises the macro-averaged F1-measures obtained for the individual classifiers, utilized data set, and subset (see Section 6.3.3). In both cases one of the baseline classifiers outperformed these results (not significantly). Nevertheless, the main objective of the feature selection was to decrease the feature size drastically, to decrease the computational cost and latency of the system, and make the classifier applicable for a later real-time online application.

6.3.2 Emotion-Wise Evaluation of the Classifier's Performance

Up to now, I have identified the classifier's setup (optimal parameter and feature set combination), which on average showed the highest performance in the considered classification task. Nevertheless, I have not yet investigated the ability of the classifier to detect a certain emotional state. The results obtained for the considered emotional states will now be presented and evaluated. For this investigation, only the results obtained when utilizing the previously identified optimal feature set and corresponding best performing hyper-parameter set were taken into consideration.

Random Forest and Support Vector Machine

The performance measures of the optimized RF and SVM classifiers are stated in Tables 6.4 and 6.5. The measures were obtained by averaging the UAR, UAP and F1-measure over all LOSO cross-validation experiments per emotion. The best results obtained for each performance measure are denoted in bold. It can be seen that for both classifiers only *frustration* and *neutral* achieved values above chance level for the UAR (chance level $\widehat{=}$ 25%). Nevertheless, all emotional states achieved UAPs ranging from lowest 35.53% for the SVM-recognition of *positive* to highest 54.99% for the SVM-recognition of *frustration*. This indicated that even with a low percentage of truly *positive* or *anxious* speech samples being recognized, the ones recognized as said emotion truly belonged to this state above chance level.

From the Confusion matrices presented in Tables 6.6 and 6.7 further insight on the confusion in-between the recognition of the individual emotional states was drawn. The main diagonal corresponds to the percentage split of correctly classified samples for each state on basis of the true state of this sample. These values match the values obtained for the UAR in Tables 6.4 and 6.5. It was noticed that the low performance measures for the emotional states *anxiety*, *frustration* and *positive* were mainly caused by a confusion with the *neutral* state (red entries). The confusion between other emotional states was, with one exception, below 10% in both cases, FP and FN (see Table 2.5 on page 47 for explanation). This exception was observed in case of true *positive* samples being predicted as *frustration.* Especially for valence-related features, lying in the same half-space of arousal (cf. Figure 3.5), as it is the case for expressive positive emotions and frustration/ anger, there exists a strong ambiguity in the classification of these emotional states (cf. [Harimi et al. 2015] and [Wu et al. 2011]). This is due to the fact of *spectral features* mainly contributing to the differentiation between different arousal levels (cf. [Kim et al. 2009]).

When computing the macro-averaged F1-measure over all LOSO experiments and considered emotional states, it was further noticed that the RF approach slightly outperforms the SVM approach ($\text{F1}_{RF} = 38.77\%$ and $\text{F1}_{SVM} = 38.05\%$). From the emotion-wise evaluation it was shown that both classification approaches have individual strengths of recognizing certain emotional states. While the RF classifier outperformed the SVM in case of the *anxiety* and *frustration*, a contrary behaviour was observed in case of *positive.*

Findings on Emotion-Wise Performance Evaluation

From the presented results it can be concluded that both classifiers were unable to detect all emotional states with a sufficient recall. Especially for *anxiety* and *positive* the UARs did not reach the critical value of 25% (chance level). The highest mismatch of emotional states was related to the prediction of a neutral state with actually one of the other emotional states being truly present. This can be contributed to the fact that the utilized data set comprises highly natural and low expressive emotions, where the threshold between the induced mild emotion and the neutral state is too low to be fully distinguishable. This is also in line with the annotation results presented in Section 5.1.2 and the observations made in [Siegert et al. 2014]. In Section 5.1.2 it was shown that a majority of the samples labelled as a mild emotional category belonged to the neutral space of the dimensional annotation approach (cf. Table 5.4 on page 143). Nevertheless, for each emotional state a high precision of at least 35.53% was reached, which indicates an above chance level probability of the emotional state being correctly assigned to said state, for the considered categorial four-class classification task.

From the results obtained for the individual LOSO experiments, it was noticed that for some subjects the classifiers were unable to detect certain emotional states

Table 6.4: UAR, UAP and F1-measure in [%] of the optimized RF classifier (feature set 15 and parameter set 14), investigated separately for each emotional state. Brackets denote standard deviation. The best results regarding the individual performance measures are indicated in bold (adapted from [Requardt; Ihme et al. 2020]).

	anxious	frustrated	neutral	positive
UAR	9.52 (6.76)	28.11 (18.57)	**87.63** (10.83)	6.66 (7.45)
UAP	49.77 (31.73)	49.75 (24.82)	**51.96** (13.28)	36.67 (33.67)
F1	14.17 (8.71)	31.60 (18.15)	**64.06** (12.10)	10.48 (11.05)

Table 6.5: UAR, UAP and F1-measure in [%] of the optimized SVM classifier (feature set 20 and parameter set 11), investigated separately for each emotional state. Indications as in Table 6.4.

	anxious	frustrated	neutral	positive
UAR	6.36 (6.60)	25.05 (18.38)	**87.84** (10.57)	9.34 (10.26)
UAP	44.14 (31.36)	**54.99** (29.74)	51.68 (13.35)	35.53 (32.14)
F1	11.74 (11.22)	29.72 (18.00)	**63.77** (11.77)	11.64 (11.86)

Table 6.6: Confusion matrix of the optimized RF classifier (feature set 15 and parameter set 14) given as percentage split of the actual class. Brackets denote standard deviation. Grey entries on the main diagonal corresponds to the percentage split of correctly classified samples on basis of the actual class. Red entries denote the highest confusion for each emotional state (adapted from [Requardt; Ihme et al. 2020]).

		Predicted			
		anxious	frustrated	neutral	positive
Actual	anxious	9.52 (6.76)	6.63 (8.76)	82.59 (10.16)	1.25 (2.17)
	frustrated	1.91 (1.75)	28.11 (18.57)	65.97 (19.37)	4.02 (5.58)
	neutral	2.10 (2.45)	8.82 (10.15)	87.63 (10.83)	1.45 (2.96)
	positive	3.43 (6.58)	14.63 (15.63)	75.28 (20.21)	6.66 (7.45)

Table 6.7: Confusion matrix of the optimized SVM classifier (feature set 20 and parameter set 11) given as percentage split of the actual class. Indications as in Table 6.6.

		Predicted			
		anxious	frustrated	neutral	positive
Actual	anxious	6.36 (6.60)	5.70 (8.77)	85.18 (11.48)	2.76 (4.62)
	frustrated	2.05 (2.49)	25.05 (18.38)	66.53 (18.17)	6.37 (10.80)
	neutral	2.41 (3.59)	6.76 (9.16)	87.84 (10.57)	2.99 (5.32)
	positive	1.91 (3.35)	12.21 (15.27)	76.53 (19.46)	9.34 (10.26)

at all. This was especially the case when subjects showed a low expressiveness in all of their utterances. The dimensional annotation results of the relevant subjects revealed that a majority of their speech samples laid in the neutral space of the dimensional space (moderate to low arousal and neutral valence). To verify that the low recognition performance for these subjects was related to the low expressiveness of the speech samples, also subjects leading to high recognition performances were examined. For these subjects the dimensional annotation results also revealed a much higher expressiveness of the speech samples. This indicated that in case of only mild emotions with low expressiveness, the algorithms were unable to distinguish between these mild states and a neutral state and, hence, a minimum amount of expressiveness of the speech sample would be needed to receive reliable results from the classifiers. With regard to the desired application domain this limitation is acceptable and could only be discarded with a massive increase in the size of the data set.

6.3.3 Adjusting the Data Set

From the previous Section it was possible to identify subjects from the original data set, which would not contribute to the performed recognition task. These subjects were excluded from the cross-validation process, as they showed an inconsistency in their emotional behaviour obtained through the annotation process. Here, the results of the categorial and dimensional annotation approaches were heavily conflicting. Therefore, these subjects would contaminate the investigation. As a side result, which is by construction not a tempering with the data, the recognition performance of the classifiers for the remaining subjects increases, as confusions are reduced. Furthermore, the reliability of the classifiers increases.

Therefore, in this Section, I re-evaluated both classifiers using only the corresponding subset of subjects. This re-evaluation, additionally, involved an individual feature selection, hyper-parameter optimization, and performance validation of the adapted RF and SVM classification model. The process of feature selection and hyper-parameter optimization was performed similar to the one when utilizing the complete data set. The results will, however, be presented in a less extensive way, focusing on the most relevant findings only.

The Reduced Data Set

The reduced data set only contained those subject for which the classifiers were able to detect the considered emotional states. From previous findings in this Thesis (cf. Section 5.1 and Section 6.3.2), it was concluded that the left out subjects showed insufficient expressiveness to automatically and manually classify the emotional states correctly. By excluding these subjects from the data set, it is possible to increase the sensitivity of the classifiers towards more expressive emotions. This further implies that, if the speech based emotion recognition system is unable to detect any other state than *neutral*, a different modality needs to be considered to detect the driver's emotional state.

Table 6.8: Samples contained in the reduced subset of the emotional real-world in-car recordings. Brackets denote the share of male/ female samples (adapted from [Requardt; Ihme et al. 2020]).

Label	Samples [#]	Time [min]
Neutral	2137 (1605 / 532)	51.19 (38.04 / 13.15)
Positive	716 (527 / 189)	17.88 (13.41 / 4.48)
Frustrated	1087 (724 / 363)	27.02 (17.73 / 9.28)
Anxious	691 (533 / 158)	16.64 (12.49 / 4.15)
$\sum$	4631 (3389 / 1242)	112.73 (81.67 / 31.07)

In total 11 subjects were excluded from the original data set. The reduced data set contained 4631 speech samples originating from 17 speakers (four females). An overview on the corresponding subset is presented in Table 6.8.

Feature Selection

As the feature selection methodology presented in Section 6.1 is based on the performance of the RF classifier, the importance of the individual features also changed when utilizing only a subset of the original data samples. When repeating the feature selection on the reduced data set, it was noticed that the correlation among the feature importance of the 17 LOSO cross-validation experiments was much lower compared to the correlation obtained for the whole data set ($r = 0.49$ (0.02)). The decrease in correlation may be caused by several factors, for example, the highly inter-individual feature characteristics of the speakers themselves and the considered emotional states. This was also observed when considering those features, which were included in the top 100 feature importance ranking among all subjects. Surprisingly, this first feature set (Set 1*) also comprised 15 features (cf. Set 1 of original data set). These 15 features, however, were not identical. Nevertheless, there was a consistency of 73.33% in their feature composition. The corresponding correlation of the importance ranking of the new set, however, reached only $r = 0.79$ (0.04). In the previously presented feature selection process, only those feature sets achieving a high correlation of $r \geq 0.80$ were considered. To select the most relevant feature sets, for the reduced data set, this correlation limit needed to be reduced to 0.70. By doing so and proceeding the process presented in Section 6.1.2, 13 feature sets were identified. The resulting r-values and numbers of features included in the sets are presented in Figure 6.6. To indicate the difference of the feature sets, the sets obtained from the reduced data set are marked with a star.

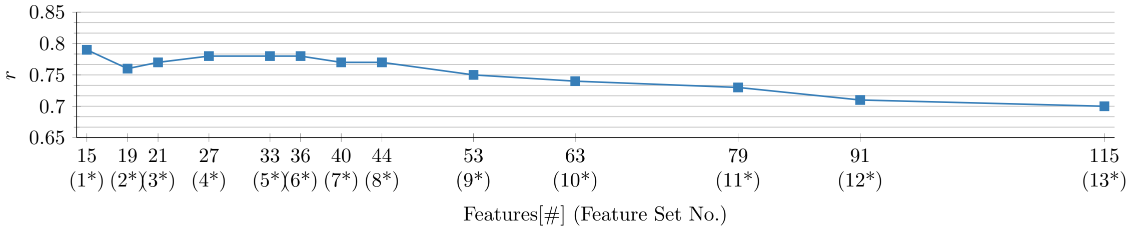


Figure 6.6: Correlation coefficient r of the reduced feature sets obtained when utilizing the reduced data set. Brackets denote the corresponding feature set numbers. Stars denote feature set numbers obtained from the reduced data set

Parameter Optimization

The hyper-parameter optimization was based on a random search. During this process I identified a search space from which the optimal parameter combinations were chosen randomly. This search space was identical to the search space utilized during the hyper-parameter optimization performed on the complete data set. Consequently the identical parameter combinations, as stated in Table 6.1 on page 172, were utilized. The numbering of the parameter sets was maintained, to prevent confusion. Nevertheless, the performance measures averaged over all feature sets

Table 6.9: UAR, UAP and F1-measure of a 4-class driving condition problem averaged over all investigated feature sets for the different parameter combinations (Par. Set) of the RF classifier with ascending F1-measure (macro-averaged) when utilizing the reduced data set. Brackets denote the standard deviation. The best results regarding the individual performance measures are marked in green color. Red entries indicate parameter sets which were identifies as unsuitable during the statistical analysis.

Par. Set	UAR [%]	UAP [%]	F1[%]
2	33.52 (0.88)	35.96 (0.90)	34.69 (0.86)
1	33.62 (0.64)	36.14 (0.84)	34.84 (0.73)
3	34.52 (0.62)	45.26 (1.95)	39.16 (1.08)
4	34.58 (0.52)	47.89 (2.44)	40.15 (1.18)
12	34.92 (0.55)	48.04 (2.76)	40.43 (1.35)
8	34.80 (0.59)	48.34 (2.77)	40.45 (1.35)
13	34.90 (0.65)	48.56 (2.71)	40.60 (1.38)
9	34.87 (0.72)	48.82 (2.84)	40.66 (1.47)
10	34.88 (0.57)	48.82 (2.73)	40.67 (1.33)
6	34.65 (0.52)	49.58 (2.90)	40.77 (1.34)
7	34.88 (0.48)	49.16 (2.71)	40.79 (1.25)
11	35.00 (0.58)	48.93 (2.83)	40.79 (1.37)
5	34.66 (0.49)	49.71 (2.47)	40.83 (1.15)
14	34.98 (0.64)	49.33 (3.08)	40.91 (1.50)

and ranked in increasing order of the macro-averaged F1-measure (cf. Tables 6.9 and 6.10), shows a deviation from the results obtained when utilizing all subjects included in the data set (cf. Table 6.2 on page 175 and Table6.3 on page 178).

Especially in case of the RF classification approach, the ranking differs considerably. By performing repeated-measures ANOVAs, it was possible to identify parameter sets, which were outperformed by a majority of the other sets (marked in red color). These parameter sets were identical to those identified when utilizing the complete data set for both classification approaches. The best results obtained for individual performance measures are marked in green color.

Again, the statistical analysis was repeated after exclusion of said sets. As before, in case of the RF, it was not possible to identify a parameter set outperforming all other sets. This was also indicated by the ranking of the remaining parameter sets (cf. Table 6.9), which shows a correlation to the original ranking in Table 6.2 on page 175 and again shows only small performance measure ranges. From the obtained results of the two statistical evaluation, three parameter sets (1, 2 and 3), which were identified as non-optimal parameter candidates, were discarded. However, it was not possible to identify one parameter set, which outperformed all other sets. Consequently, the 11 remaining parameter sets 4, 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14 were used later for the feature set dependent evaluation of the RF classifier.

Table 6.10: UAR, UAP and F1-measure of a 4-class driving condition problem averaged over all investigated feature sets for the different parameter combinations (Par. Set) of the SVM classifier with ascending F1-measure (macro-averaged) when utilizing the reduced data set. Indications as in Table 6.9.

Par. Set	UAR [%]	UAP [%]	F1[%]
1	25.00 (0.01)	11.36 (0.39)	15.62 (0.36)
2	25.00 (0.01)	11.36 (0.39)	15.62 (0.36)
3	25.00 (0.02)	11.36 (0.39)	15.62 (0.36)
4	25.03 (0.11)	13.92 (4.47)	17.53 (3.29)
5	25.51 (0.50)	20.58 (6.70)	22.22 (4.47)
6	29.62 (1.35)	37.26 (2.43)	32.99 (1.67)
8	29.40 (1.92)	37.94 (2.43)	33.08 (3.12)
7	32.37 (0.57)	40.05 (1.93)	35.79 (1.06)
9	32.50 (1.02)	42.99 (3.51)	36.99 (1.91)
12	33.57 (1.18)	45.93 (4.16)	38.75 (2.18)
10	33.33 (1.41)	46.95 (4.11)	38.95 (2.31)
14	34.00 (0.92)	45.69 (2.67)	38.97 (1.47)
13	33.83 (1.33)	46.56 (4.09)	39.15 (2.22)
11	33.44 (1.52)	47.67 (4.51)	39.26 (2.49)

In case of the SVM, parameter sets 10, 11, 12, 13 and 14 outperformed a majority of the remaining sets, and sets 6, 7 and 8 were outperformed by a majority. This is in line with the results obtained on the original data set, with a high correlation in the ranking of the parameter sets and again wider performance measure ranges (cf. Table 6.10 and Table 6.2 on page 175). With consideration of the first statistical analysis on all parameter sets, eight parameter sets, were excluded and the remaining six parameter sets 9, 10, 11, 12, 13 and 14 were used later for the feature set dependent evaluation of the SVM classifier.

Performance Validation

The UAR and UAP of the optimized RF and SVM classifiers, utilizing the remaining parameters sets identified in the previous section and averaged over all LOSO cross-validation experiments for the newly generated feature sets, are stated in Figures 6.7 and 6.8. To identify if there exists one feature set, which outperforms the other sets, repeated-measures ANOVAs were performed. They revealed a significant effect of the feature set on the performance of the classifiers. Nevertheless, it was not possible to identify one feature set outperforming a majority of the others, in case of the RF classifier. Here, the best feature set achieving the highest macro-averaged F1-value (see Figures on page 288 ff. in Appendix D) was chosen as best feature set (Set 10*). This set included 63 features with a 70% agreement in the features compared to set 15, chosen as optimal feature set for the complete data set. As for set 15, set 10* includes mainly features related to the MFCC No.1 and 2 (9 and 12 features, respectively) and the LSP of the LPC-coeffcient No. 3 and 7 (6 features, each). The main difference compared to feature set 15 lies in the share of *spectral features*. While set 15 included 88% of *spectral features*, set 10* only included 81% of these feature type. Especially the amount of features related to intensity (3 features) and probability of voicing (3 features) was increased. The most frequently applied functionals were the maximum value (8 occurrences), minimum value (7 occurrences), quartiles (13 occurrences) and inter-quartile ranges (8 occurrences). Furthermore it was possible to identify a decrease of features based on the arithmetic mean (5 occurrences) and an increase of features based on the standard deviation (5 occurrences).

In case of the SVM classifier, it was possible to identify one set outperforming a majority of the other sets in their UAR (set 13* vs. sets [1*, 2*, 3*, 4*, 5*, 6*, 7*, 9* & 10*], all p's < 0.05, Bonferroni-corrected). For the UAP no feature set could be identified. As set 13* also achieved the highest F1-measure, this set was chosen as the best performing feature set. As for the results obtained when applying the complete original data set, set 13* corresponds to the feature set containing the largest number of features (115 features). In comparison to set 20, set 13* only shows an agreement in 56% of the features included in both sets. The highest agreement was found in the category of *spectral features*. Again most of the features were

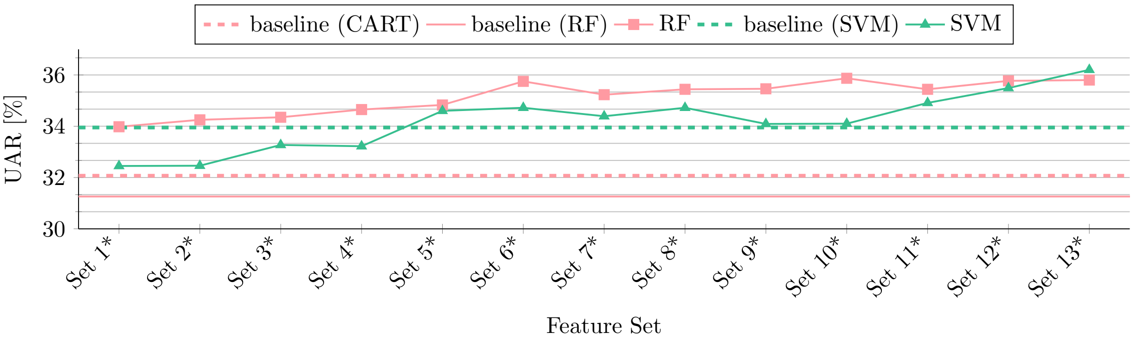


Figure 6.7: Mean UAR of the optimized RF/ SVM classifiers for each individual feature set. The classifiers were tested and trained on the reduced data set.

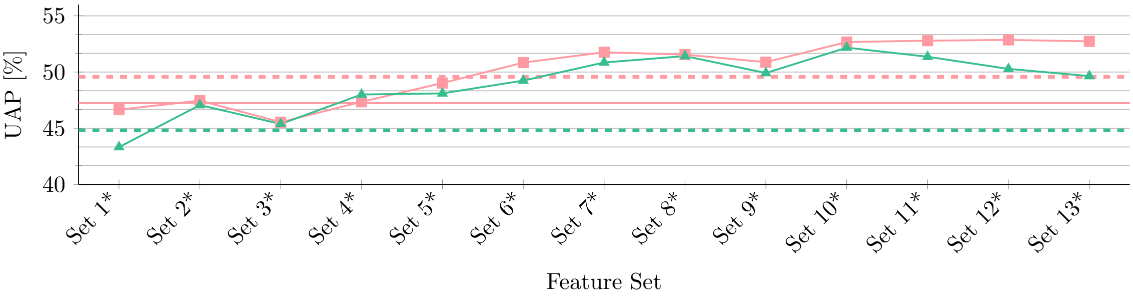


Figure 6.8: Mean UAP of the optimized RF/ SVM classifiers for each individual feature set. The classifiers were tested and trained on the reduced data set.

grouped in this category and related to the MFCC No.1 and 2 (16 and 14 features) and the LSP of the LPC-coefficient No. 1, 3 and 7 (7 features, 7 features and 8 features, respectively). Nevertheless, as for the RF classifier, the share of *spectral features* decreased from 87% (set 20) to 78% (set 13*) with an increased number of features related to the voice probability (6 features), intensity (6 features) and loudness (5 features). In case of the most frequently applied functionals, similar findings as for the RF classifier were made. Additionally to the minimum and maximum value, quartiles and inter-quartile ranges (cf. results obtained for RF classifier), it was noticed that for SVMs also the range (10 occurrences) played a decisive role. Furthermore, an increased number of features related to the standard deviation (7 occurrences) was identified.

By utilizing the reduced data set and applying the feature sets to the RF and SVM classifiers with optimized parameter set, it was possible to increase the overall recognition performance. The corresponding macro-averaged F1-measures obtained on the classifiers with optimized feature and parameter set, and trained and tested on the original data set and reduced subset, are stated in Table 6.11. Additionally, the results of the performed baseline classification experiments are stated. It can be seen that the overall recognition performance increased considerably for both classification approaches. In agreement with the results obtained on the complete data set, the RF classifier outperformed the SVM classifier.

Table 6.11: F1-measure in [%] of the most relevant RF and SVM classification experiments.

	baseline	set 15	set 20	subset(set 10*)	subset(set 13*)
CART	38.94				
RF	37.62	38.77		42.68	
SVM	38.63		38.05		41.87

The emotion-wise results of the LOSO cross-validation experiments, when applying the identified feature sets to the RF and SVM classifiers with optimized parameter sets, are stated in Tables 6.12 and 6.13. In both Tables, bold entries denote the best results obtained for each evaluated performance measure. The corresponding confusion matrices are stated in Table 6.14 and 6.15. For both classification methods a distinct increase of the recognition performance for the emotional states of *anxiety*, *frustration* and *positive* was noticed, while the recognition performance of the *neutral* state decreased. In case of the SVM classifier, a substantial increase for the recognition of the *anxiety* and *positive* state was noticed. In contrast to the results obtained on the complete data set, the F1-measure increased from 11.74% to 19.39% for *anxiety*, and from 11.64% to 19.40% for *positive*. While the UAR could still not reach the chance level of a random guess for a four class-classification problem (25%), the UAP noticeably increased. In case of the RF classifier, the achieved performance outperformed an UAP of 50% for each emotional state. In case of *frustration*, the RF classifier outperformed the SVM classifier in all performance measures. For this state, the UAR increased substantially from 28.11% to 38.60% accompanied with a decreased standard deviation. A nearly contrary finding was observed for the *anxiety* state, where the SVM classifier outperformed the RF classifier for UAR and F1, while the UAP was almost identical. Again, compared to the results obtained on the original data set, the standard deviation decreased for UAP and F1. Considering the confusion matrices stated in Tables 6.14 and 6.15, still a majority of the confusion was caused by a prediction as *neutral*. Nevertheless, this confusion was noticeably lower compared to the results obtained on the complete data set. Furthermore, an increase of confusion with *frustration* was noticed. This may be attributed to the higher expressiveness of the utilized speech samples. It can be assumed that a higher expressiveness is accompanied with a higher arousal. With all three considered emotional states lying in the same half-space of arousal (cf. Figure 3.5 on page 93) this, consequently, is also leading to a higher ambiguity in the classification of these emotional states (cf. results obtained on the complete data set).

Table 6.12: UAR, UAP and F1-measure in [%] of the optimized RF classifier (feature set 10* and parameter set 7), evaluated on the reduced data set, investigated separately for each emotional state. Brackets denote standard deviation. The best results regarding the individual performance measures are indicated in bold.

	anxious	frustrated	neutral	positive
UAR	9.93 (6.87)	38.60 (15.81)	**82.72** (10.29)	12.24 (7.57)
UAP	51.36 (32.19)	**54.76** (22.22)	51.28 (11.50)	53.30 (23.65)
F1	14.09 (7.05)	40.28 (12.55)	**62.16** (9.94)	18.63 (10.72)

Table 6.13: UAR, UAP and F1-measure in [%] of the optimized SVM classifier (feature set 13* and parameter set 11), evaluated on the reduced data set, investigated separately for each emotional state. Indications as in Table 6.12.

	anxious	frustrated	neutral	positive
UAR	13.82 (7.73)	35.69 (15.55)	**81.08** (10.10)	14.20 (10.59)
UAP	51.33 (25.06)	**52.78** (20.81)	51.64 (12.40)	42.81 (24.21)
F1	19.39 (8.20)	38.82 (11.71)	**61.91** (10.71)	19.40 (13.56)

Table 6.14: Confusion matrix of the optimized RF classifier (feature set 10* and parameter set 7), evaluated on the reduced data set, given as percentage split of the actual class. Brackets denote standard deviation. Grey entries on the main diagonal corresponds to the percentage split of correctly classified samples on basis of the actual class. Red entries denote the highest confusion for each emotional state.

		Predicted			
		anxious	frustrated	neutral	positive
Actual	anxious	9.93 (6.87)	9.91 (10.12)	78.19 (9.93)	1.97 (3.57)
	frustrated	1.80 (1.96)	38.60 (15.81)	54.67 (12.03)	4.92 (6.38)
	neutral	3.29 (4.89)	12.61 (12.07)	82.72 (10.29)	1.38 (1.41)
	positive	4.26 (7.11)	19.93 (17.12)	63.57 (17.50)	12.24 (7.57)

Table 6.15: Confusion matrix of the optimized SVM classifier (feature set 13* and parameter set 11), evaluated on the reduced data set, given as percentage split of the actual class. Indication of as in Table 6.14.

		Predicted			
		anxious	frustrated	neutral	positive
Actual	anxious	13.82 (7.73)	9.94 (11.79)	73.65 (10.33)	2.59 (4.40)
	frustrated	2.36 (2.29)	35.69 (15.55)	55.04 (13.58)	6.91 (9.24)
	neutral	3.96 (4.24)	11.80 (11.75)	81.08 (10.10)	3.15 (4.36)
	positive	3.42 (5.07)	19.26 (15.91)	63.12 (19.53)	14.20 (10.59)

Findings on the Reduced Data Set

From the results obtained on the reduced data set it can be concluded, that besides an optimization of the classification model, by utilizing feature selection and parameter optimization, also the utilized data samples play a decisive role. Even though the evaluation of the annotation presented in Section 5.1 showed a high reliability and consistency of the annotation results, it can be assumed that the utilized speech samples are not automatically suitable for speech emotion recognition. This can be mainly attributed to the human auditory system which is trained to distinguish emotions in very fine gradations and is also highly dependent on the individual's cultural, gender and family background (cf. Section 2.1.4). Especially for an automatic detection of emotions, low expressive emotions can only hardly be distinguished from a *neutral* state. Therefore, it was assumed that by excluding those subjects from the data set showing a low expressiveness in their utterances, the individual emotional state can be better distinguished from one another. This assumption was on the one hand confirmed by the changes of the features identified as most relevant for the present recognition task. Here, an increased importance of voice probability, intensity and loudness related LLDs was identified and the applied functionals showed an increased importance of range and standard deviation. These functionals are strongly related to the subject's individual speaking characteristics and indicate an increased importance of their inter-individual variability. On the other hand, also an increased overall and emotion-wise recognition performance of the classifiers was noticed. While the overall macro-averaged F1-measure could be increased from a maximum of 38.77% (RF) to 42.68% (RF), the emotion-wise increase in recognition performance was even more impressive and strongly classifier dependent. In case of *anxiety*, the F1-measure could be increased from 14.17% (RF) to 19.39% (SVM). For *frustration* the F1-measure was increased from 31.60%(RF) to 40.28% (RF) and for *positive* from 11.64% (RF) to 19.40% (SVM). In contrast to this large increase of at least 5.22% for the *anxiety*, *frustration* and *positive* state, the decrease in the F1-measure for the *neutral* state is comparatively low with a decrease from 64.06% (RF) to 62.16% (RF).

6.4 Summary and Discussion

In this Chapter, I investigated the ability to automatically recognize the driver's emotional state by utilizing real-world emotional speech data. To do so, two different machine learning approaches were utilized (RF and SVM) and validated using the data samples of the real-world in-car speech recordings presented in Section 3.2 and processed in Chapter 5. The validation was performed using the LOSO cross-validation scheme, which ensures speaker-independent recognition performances. To further boost the recognition performance of the classifiers, a feature selection and hyper-parameter optimization was conducted.

As feature selection a RF wrapper method was applied. It was possible to identify up to 20 feature sets for which the feature importance of the individual features showed a sufficient correlation. The parameter optimization was based on a random search performed on a predefined search interval of the most relevant hyper-parameters of the RF and SVM classification methods. By performing a statistical analysis on the overall recognition performances, it was possible to identify that, in case of the RF classifier, the feature selection is more relevant for a good recognition performance than choosing the optimal parameter combination. Furthermore, no relation between the parameter combination and the performance of the classifier could be drawn. This is contrary to the results obtained for the SVM classifier, where the optimal feature set was strongly dependent on the chosen parameter combination and a clear dependency of the hyper-parameter γ and the size of the feature set was noticed. When utilizing the identical feature set for both classification approaches, it was further noticed that the RF classifier would, in most cases, outperform the corresponding SVM classifier. This, however, may be also attributed to the utilized feature selection method which is a wrapper method based on the performance of a RF-classifier. It is assumable that the application of a SVM based wrapper or a filler feature selection leads to an increase in the recognition performance.

From the emotion-wise evaluation of the recognition performance, it was further noticed that there exist some subjects for which an automatic and manual differentiation of the emotional states was not possible. It was assumed that these subjects showed too low expressiveness in their speech utterances and, hence, their data was not contributing to the presented emotion recognition task. Furthermore, taking into account the annotation results from Section 5.1.2, an inconsistency in whether their speech samples contained emotional content or not was identified. Consequently, by leaving out these uncertain samples, not only the performance of the classifier can be increased, but also its trustworthiness. From these results, it was concluded that a certain amount of expressiveness is needed to be able to automatically distinguish between the individual emotional states. This was confirmed by repeating the LOSO cross-validation experiments on the reduced data set. This also led to differences in the feature selection and hence, led to different features inside the identified features sets. The overall statistical analysis, however, led to a similar behaviour of the considered classification approaches, with the RF classifier outperforming the SVM classifier.

Finally, by reducing the original data set, in combination with an independent feature selection and hyper-parameter optimization, it was possible to increase the overall F1-measure from 38.77% up to 42.68%. This corresponds to an increase from 32.98% up to 35.87% in the UAR and an increase from 47.04% up to 52.67% in the UAP. The utilized feature set comprised 63 features, which corresponds to an astonishing reduction down to 6.38% of the original *emobase* feature set. With regard to the computing time, especially for a later real-time in-car application, the number of features plays a decisive role. From the emotion-wise evaluation it was shown, that the investigated machine learning algorithms show individual strengths regarding the individual considered emotional state. While for both classification approaches the UAR of detecting *anxiety* and *positive* could not reach the chance level, the UAP of the RF classification approach achieved values above 50%. For the detection of *frustration*, UARs above chance level were achieved by both classifiers, with the RF outperforming the SVM.

Overall, it can be stated that for both classification approaches, the feature selection, hyper-parameter optimization and adjustment of the data set significantly affected the recognition performances, which reinforces the importance of such measures to be taken. By reducing the feature set to 6.38% of the original set, the performance of the speech emotion recognizer was not only maintained but even increased, which plays a decisive role for a later real-world application. It was further shown that a system needs a certain amount of expressiveness in the users speech to obtain applicable results. In case of too low expressiveness, a distinction from the neutral state is not possible.

With a focus on emotion recognition in-the-wild and in in-vehicle environments, these results pave the way to the detection of novel driver states in automotive industry. To my knowledge, these kind of highly natural and low expressive everyday driver's emotions have not been available to the research community. First attempts towards in-vehicle emotion recognition have already been made based on simulated and real-world speech data, and were presented in Section 2.6. Nevertheless, these results are not generalizable and an systematic extensive analysis, as presented in this Chapter of the Thesis, is rarely addressed. Furthermore, this Chapter also considering aspects for a later real-time in-vehicle application, by drastically decreasing the number of relevant features needed for a trustworthy emotion recognition and consequently also the computational cost.

In the next Chapter, I will conclude this Thesis and recapitulate the findings of the Chapters 3 to 6. The main results will be highlighted, and open research questions and future work will be discussed.

Chapter 7

Conclusion

Contents

At the beginning of this Thesis I have presented to the reader three hypotheses. In the progress of this Thesis we have now come to a point where these hypotheses have been examined.To recap the findings, I will now come back to these main hypotheses and draw a conclusion on the overall underlying research questions. Furthermore, I will present to the reader the open research questions, which have not yet been answered in the scope of this Thesis.

7.1 Conclusions on Main Hypotheses

1. **Hypothesis:** It is possible to induce naturalistic emotions in the driver, while driving in a real vehicle.

   To validate the first hypothesis, a data collection inside a real vehicle was conducted. This data collection included the induction of four target emotional states while driving in a conventional gasoline passenger car, namely, neutral, positive, frustration and anxiety. The emotions were induced by using emotion elicitation through supervised real-life studies (i.e. conducting secondary tasks while driving, designed to induce a certain emotion) and through retrospective (i.e. initiating a conversation on related topics and experiences of the driver). To prevent a bias of the driver towards a certain emotion, after each emotion scenario a short 5 minute recess was scheduled. Furthermore, the order of the emotion scenarios was chosen in a way to avoid a negativity bias for the neutral and positive emotional state.

   The data was collected in three modalities (audio, video and bio-physiological data) and afterwards validated by analyzing the bio-physiological signals of the driver, while driving under emotional influence. The validation results showed clear differences in their characteristics considering the different emotion scenarios of the experimental setup. These differences were, however, not significant for all emotional states. A further, more fine grained, annotation of the speech signal itself showed expressions of emotions via speech prosody (without considering the spoken content). By evaluating the individual driving scenarios it was shown, that the occurrence of emotional speech also match the induced emotional state of the different driving scenarios. Based on these results it can be concluded that it is possible to induce an emotional state in the driver while manually driving a vehicle.

2. **Hypothesis:** It is possible to compensate effects of speech distortion.

For the second Hypothesis, it was investigated if the effect of speech distortion can be compensated by applying speech enhancement. To do so, I first evaluated the quality of the speech signal of compressed and noisy speech. First investigations were done on compressed speech, as this kind of speech is "easy" to generate, as the signal is manipulated by digital signal processing steps and not by external noises. Hence, the original clean speech signal is available, which is necessary to determine the quality of the distorted signal (i.e. Signal-to-Noise Ratio (SNR) or newly developed Compression Error Rate (CER)). This is more challenging when it comes to noisy speech, as the clean speech signal is in most cases not available and highly challenging to obtain in real-life settings. For compressed speech, by applying different audio codecs and bit-rates, it was shown that there exists a clear correlation between the speech quality and the performance of speech emotion perception by humans and automatic recognition systems. Surprisingly, in case of automatic speech emotion recognition, the best results were not obtained on the original uncompressed signal, but when utilizing the MP3 codec with higher bit rates (over 24 kbit/s). This may be contributed to the fact that MP3 uses perceptual coding as coding technology, which reduces redundancies in the speech signal by disregarding those signal parts which are supposed to be beyond the resolution of the human auditory system.

For noisy speech, regarding speech quality, a highly significant effect of the recording condition and microphone setup was identified. In contrast, regarding the recognition performance, the effect of the recording conditions and the microphone setup was not significant, even though, a clear decrease of recognition performance with decrease of speech quality was identified. Considering these findings, the necessity of taking into account the speech quality in case of noisy speech is not evident. This was also confirmed by the results obtained when applying a speech enhancement algorithm (i.e. Optimally-Modified Log-Spectral Amplitude (OM-LSA)-Improved Minima Controlled Recursive Averaging (IMCRA)). Even though the statistical analysis of the utilized speech features showed a significant alteration, no significant effect on the performance of the speech emotion recognition system for enhanced and disturbed speech was identified. With regard to the high number of altered features in case of enhanced speech, it is recommended not to use speech enhancement in case of noisy speech.

For hypothesis two, it can be concluded that in case of noisy speech no compensation of its effects on the speech signal is needed. For compressed speech, however, its effect on the speech signal and the performance of the speech based emotion recognition system is strongly dependent on the utilized audio codec. In some cases it is even possible to increase the performance of the system by applying a specific audio codec.

3. **Hypothesis:** Under the assumption that hypotheses one and two apply, it is possible to automatically detect the emotional state of the driver by only considering the speech signal of the driver and its prosodic features.

   The final research hypothesis examines the ability to automatically recognize the driver's emotional state in a naturalistic everyday driving situation by applying suitable machine learning algorithms. Here, the evaluation results of hypotheses one and two play a decisive role, as they form the base of this research question. Only with highly natural and low expressive emotional speech data being available and the knowledge on how to cope with distorted in-vehicle speech, a meaningful validation of the classification results is possible. By utilizing LOSO cross-validation experiments and applying feature selection and parameter optimization strategies, and an adjustment of the original data set, two classifiers were identified. The best classification result of 42.68% F1-Measure was achieved when applying a RF classifier and a feature set of 63 features including 81% of spectral speech features. In comparison, the baseline RF classifier without feature selection, parameter optimization and data set adjustment only achieved 37.62% F1-Measure. Considering the emotion wise classification results of the investigation, it was possible to detect neutral and frustration emotions well above chance level. In case of anxiety and positive emotions, a rather low recall below chance level was achieved. Nevertheless, the precision for these states was almost equal as for neutral and frustration. A majority of the occurring mismatches in the emotional states were related to the confusion with the neutral state (ranging from 54.67% for frustration to 78.19% for anxiety). This implies that a mismatch with an emotional state other than neutral is comparatively rare.

   Therefore, in case of hypothesis three, it can be concluded that it is possible to detect the driver's emotional state from speech. However, the speech signal needs to contain a certain amount of emotion expressiveness, whereas the emotions occurring in everyday driving situations are in most cases of low expressiveness and only to some extent distinguishable from a neutral driver state. Nevertheless, a precision of detecting a certain emotion of above 50% was achieved for all emotions. With the highest confusion occurring with the neutral state, it is most certain that whenever a different emotional state is detected this state is also true.

7.2 Discussion, Open Issues and Future Research

As already stated in the previous Section, the results presented in this Thesis can be seen as first attempt towards detecting emotions from speech in a highly natural and low expressive in-the-wild environment. Even though the results are still capable of improvement, they are highly promising, especially in contrast to available comparable results presented in Section 2.6 at the beginning of the Thesis and with regard to the very limited data availability, leading to a comparatively high optimization demand. Furthermore, the results contributed to a successful final evaluation of the ADAS&ME research project and were acknowledged by the reviewer.

With a focus on a later real-world in-vehicle application, the investigations presented on compressed speech are of high interest, as cloud-based Advanced Driver Assistant Systems (ADAS) become more relevant in today's automotive industry [Volkswagen AG 2019]. By integrating the "Automotive Cloud" into their automotive technologies, the computational costs inside the vehicle are drastically decreased and upgrades to novel in-car applications or updates of already integrated systems can be remotely enabled. To implement these kind of systems, the raw sensor signals need to be communicated to the cloud server. In case of safety relevant systems, the latency of this data transmission plays a decisive role, as the raw sensor signals are often of large data volume. By utilizing signal compression this latency is reduced and a real-time application is feasible. Especially when utilizing multimodal data, this aspect is of increasing relevance. Considering the recommendation made on speech enhancement, these results can be of interest when it comes to the processing of in-vehicle speech, as omitting these signal processing steps will also reduce processing time and, hence, the latency of the signal transmission. However, these results can only be seen as a first tendency, as the investigations were not performed on real in-vehicle speech data but on benchmark emotional speech data sets re-recorded inside a fixed-based driving simulator.

It can be be summarized that the results pave the way to a later real-world application, but are not yet applicable inside a real vehicle. To achieve the goal of an in-vehicle application, a further evaluation and validation is needed. Only to mention some open issues, I will now concentrate on possible future work based on the contributions of this Thesis and how the work of this Thesis can contribute to other research questions out of its intended scope.

Considering the presented work, it is without question that there exists a large number of research questions which have not yet been (completely) investigated. As this Thesis focused on emotion recognition from speech, the data obtained in the real-world data collection (cf. Chapter 3) has not been analyzed to its full extent. Only considering this fact, there exists a comparatively large number of highly natural and low expressive emotional data of bio-physiological signals and facial expressions, which have not yet been evaluated. Regarding the evaluations of the real-world speech data, it is further possible to expand the presented work by making more use of the dimensional annotation results, which have only been considered in a small scope.

With a focus on in-vehicle noisy speech data and the experiments performed in the scope of this Thesis, it is further possible to identify lack in fundamental research in the field of speech emotion perception. While there exists some work on the emotion perception with additive white, brown or pink noises, environmental noises have rarely been investigated. With the existing simulated and real-world data presented in Chapter 3, first insight on emotion perception could be obtained by utilizing a listening experiment similar to the experiment performed in Section 4.2.1 on compressed speech. Another field of research, which has received less attention, is speech emotion recognition from enhanced speech. Until now, a focus of the research community was drawn on the potential to detect emotions from noisy or enhanced speech. However, in these studies the effect noises or digital signal processing steps can have on the raw speech signal is mostly neglected. Considering the feature based speech emotion recognition task presented here, the feature characteristics are of high relevance. First tentative insights on this research question are presented in Section 5.2. Here, it would be of great interest to break down the presented results to the emotion level, to also be able to investigate, if the utilized enhancement algorithm effects the recognition of certain emotions to a greater or lesser extent.

Another open issue already receiving attention for many years is the reliability of the ground truth, which the determined classification models are based on. In this Thesis the ground truth of the data was determined by performing annotations by expert labelers. Hence, the reliability of the annotated data conditions the quality and reliability of the recognition results. This issue is also addressed in Section 3.2.3 and Section 5.1.2 of this Thesis. Even though it was possible to validate the inducement of the target emotions, a significant effect compared to the neutral driving state was not present for all emotions. Furthermore, an averagely fair to moderate Inter-Rater-Reliability (IRR) was achieved for the performed annotation tasks. Compared to the results presented for other annotation tasks performed on naturalistic emotional speech data, these results are promising, especially with regard to the high inter-individual differences in emotion understanding (e.g. dependent on cultural, gender and family background). Nevertheless, compared to the annotation reliability of more objective measures these results are of rather little explanatory power. This also makes the process of defining the actual/ true ground truth highly challenging. It can be assumed that the annotation of the emotional state of the driver by (expert) labelers will never completely agree with the drivers' true intrinsic emotional states. Even though this problem has received increased attention, still, there exists no generally agreed state-of-the-art approach to solve this problem of emotion labeling.

Despite these open research questions, the results presented in this Thesis are highly representative when it comes to real-world in-vehicle speech emotion recognition. However, to make the results applicable in a wider scope, some further steps can be taken. This could, for example, be an increase of robustness towards variances in the recording conditions by evaluating the classifier in a cross-corpus setting. Furthermore, one major assumption made in the introduction of this Thesis is that a speech signal is available to the system. This, however, is a highly challenging endeavor, as speech is not naturally present in everyday driving situations. Nevertheless, it is assumable that in-car speech will become more natural and present, as it is increasingly being made operational to a multitude of service function inside the car. To, on the one hand, cope with this situation and, on the other hand, increase the reliability of the system, it is advisable to utilize a multimodal approach based on several modalities like speech, facial expression and bio-physiological parameters. Especially for the present emotion recognition task, speech and facial expressions supplement each other, as facial expressions are challenging to detect reliably whenever the driver is speaking and speech may not always be available while driving. By performing first investigations on the recorded speech and video data of the real-world data collection, it was already possible to show that the two approaches complement each other in case of signal outages. Furthermore, the confidence values of the individual recognition rates can be used to increase the reliability of the system, as these values can be used as weighting factor of the individual outputs.

Until now I have focused on open issues regarding the presented data and research results of this Thesis. To reach the major aim of developing a speech based emotion recognition system for in-vehicle application and reassure its performance in unknown real-world scenarios, some consecutive validation steps are needed. This can, for example, be done by performing large scaled field studies where the system is evaluated in real driving situations. A first small-scaled field study, only considering eight participants, was performed in the final phase of this Thesis and is shortly presented in Appendix E. In this study, frustration was induced to the participants in a similar way as for the data collection in Chapter 3. From these results a vague positive tendency on the performance of the audio-based classifier, with focus on detecting frustrated drivers, can be drawn. The results, however, lack of statistical evidence as the sample set does not represent the population in a sufficient way.

7.3 Paving New Ways

Even though the previous section has shown that there exist multiple research topics which have not been evaluated to their full extent, the results presented in this Thesis contribute strongly to the field of speech emotion recognition in natural non-ideal real-world surroundings. These contributions are, more evidently, paving the way towards empathic vehicles. The vision of this future technology is to recognize, understand and give a tailored response to the driver's internal state of interest. The main basis of this vision is to be able to recognize the driver's intrinsic state correctly. In combination with other fields of emotion research, i.e. emotion perception/ understanding, and Human-Computer Interaction (HCI), the results of this Thesis could contribute largely to a first prototypical implementation of such an empathic system. Not only do the investigations give evidence for natural speech emotion recognition, they also give first insights on the detection of complex driver states other than sleepiness, attention and distraction. These states have already been extensively investigated and are already seen as state-of-the-art in the automotive industry. With a focus on natural emotion recognition it is further possible to obtain first insights in other research areas, outside of the in-vehicle setting, where the detection of a natural emotional state is of high relevance. This is, for example, the case in natural everyday communication and, more essential, miscommunication. It is known from communication research that emotions play a decisive role when it comes to misunderstandings in interpersonal communication. This makes it reasonable that emotions also play a decisive role in natural HCI. A less apparent contribution of this Thesis can be seen in medical research. The collected and utilized data set comprised highly natural emotional speech of low-expressiveness. Despite these challenging circumstances, it was possible to identify all emotional states with a high precision, by only considering prosodic features and not the spoken content itself. This approach could be beneficial when communication with persons who are unable to communicate their emotional state verbally (i.e. in case of pervasive developmental disorders like autism).

To sum up, this Thesis gives evidence towards the ability of detecting emotions from speech in a natural real-world in-vehicle environment. It does not provide a complete solution on all the underlying research questions, however it paves the way to many subsequent investigations, for example, in the field of automotive research, communication research or medical research.

Glossary

Accuracy

Performance measure of a machine learning based classification approach representing the percentage of correctly predicted instances.

Annotation

Enriching data (e.g. video or audio) with additional information, for example, the emotional state of the speaker, by performing expert labelings.

Compression Error Rate (CER)

Novel measure to assess the quality of speech by determining the differences occurring in the spectrum of the uncompressed high quality speech samples compared to the compressed version of said speech signal. Can also be applied in case of noisy speech.

Cross-Validation

Method used to validate a classifier by separating the utilized data samples into independent data sets to train an test the algorithm. The independence of the data sets can be increased by performing a Leave-One-Subject-Out (LOSO) or Leave-One-Subject-Group-Out (LOSGO) cross-validation.

Deltas

First (delta) and second order (delta-delta) derivatives of the utilized features in the feature set.

EmoDB-Car

Re-recorded Berlin Emotional Speech Database inside a fixed-based driving simulator under silent and disturbed recording conditions (i.e. simulator turned on and simulator turned off).

Emotional expressiveness

Describing the intensity of showing an emotion. Can be related to the naturalness of the data. While acted data is mostly of high expressiveness, natural real-world data shows less expressiveness.

F1-Measure

Performance measure of a machine learning based classification approach considering the trade-off between recall and precision.

Inter-Rater-Reliability (IRR)

Performance measure to determine the reliability of a multi-rater labeling.

Labeling

Methodology of adding additional information to data.

Low-Level Descriptors (LLDs)

Global features/ characteristics of the speech signals which are most relevant for speech emotion recognition (e.g. pitch, formants, loudness, MFCCs, LPCs, ...).

Naturalness of Data

Describing the naturalness of the emotional content of a data set. It is distinguished between acted, scripted and natural emotions.

Precision

Performance measure of a machine learning based classification approach determining the percentage split of the correctly predicted positive samples out of all predicted positive samples, i.e. the probability of the true prediction to be correct.

Recall

Performance measure of a machine learning based classification approach determining the percentage split of correctly predicted positive samples out of all true condition positive samples. Also referred to as sensitivity.

Statistical Functionals

Statistics applied to the Low-Level Descriptors (LLDs) of a speech segment to extract the emotion speech features.

VAM-Car

Re-recorded Vera am Mittag Database inside a fixed-based driving simulator under silent and disturbed recording conditions (i.e. simulator turned on and simulator turned off).

References

3GPP/ETSI (2018a). *ETSI TS 126 071 V15.0.0.* Technical Specification 126071. Version 15.0.0. 3rd Generation Partnership Project (3GPP) and European Telecommunication Standards Institute (ETSI).

— (2018b). *ETSI TS 126 171 V15.0.0.* Technical Specification 126171. Version 15.0.0. 3rd Generation Partnership Project (3GPP) and European Telecommunication Standards Institute (ETSI).

— (2018c). *ETSI TS 126 290 V15.0.0.* Technical Specification 126290. Version 15.0.0. 3rd Generation Partnership Project (3GPP) and European Telecommunication Standards Institute (ETSI).

Abdelwahab, M. & Busso, C. (2019). 'Active Learning for Speech Emotion Recognition Using Deep Neural Network'. In: *Proc. of the 8th International Conference on Affective Computing and Intelligent Interaction.* ACII. Cambridge, United Kingdom: IEEE, pp. 1–7.

Abdić, I.; Fridman, L.; McDuff, D.; Marchi, E.; Reimer, B. & Schuller, B. (2016). 'Driver Frustration Detection from Audio and Video in the Wild'. In: *Proc. of the 25th International Joint Conference on Artificial Intelligence.* IJCAI. New York, NY, USA: ACM, pp. 1354–1360.

Akçay, M. B. & Oğuz, K. (2020). 'Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers'. *Speech Communication* 116, pp. 56–76.

Akerstedt, T. & Gillberg, M. (1990). 'Subjective and objective sleepiness in the active individual'. *International Journal of Neuroscience* 52.1–2, pp. 29–37.

Albahri, A.; Lech, M. & Cheng, E. (2016). 'Effect of Speech Compression on the Automatic Recognition of Emotions'. *International Journal of Signal Processing Systems* 4.11 (1), pp. 55–61.

Albahri, A. & Lech, M. (2016). 'Effects of Band Reduction and Coding on Speech Emotion Recognition'. In: *Proc. of the 10th International Conference on Signal Processing and Communication Systems.* ICSPCS. Gold Coast, Australia: IEEE, pp. 1–8.

Albanie, S.; Nagrani, A.; Vedaldi, A. & Zisserman, A. (2018). 'Emotion Recognition in Speech using Cross-Modal Transfer in the Wild'. In: *Proc. of the 26th ACM* *international conference on Multimedia*. MM. Seoul, Republic of Korea: ACM, pp. 292–301.

Aljanaki, A.; Yang, Y.-H. & Soleymani, M. (2017). 'Developing a benchmark for emotional analysis of music'. *Research Article* 12.e0173392 (3), pp. 1–22.

Almeida, P. R.; Ferreira-Santos, F.; Chaves, P. L.; Paiva, T. O.; Barbosa, F. & Marques-Teixeira (2016). 'Perceived arousal of facial expressions of emotion modulates the N170, regardless of emotional category: Time domain and time–frequency dynamics'. *International Journal of Psychophysiology* 99, pp. 48–56.

Altenburg, A. (29th Sept. 2017). *Wir sind die Freeses*. NDR2. URL: http://www.ndr.de/ndr2/wir%5C_sind%5C_die%5C_freeses/podcast425%20%5C%5C%200.html (visited on 20/08/2018).

Andersen, S.; Duric, A.; Astrom, H.; Hagen, R.; Kleijn, W. & Linden, J. (2004). *Internet Low Bit Rate Codec (iLBC)*. RFC - Experimental RFC-3951. Network Working Group.

Andersen, S.; Kleijn, W. B.; Hagen, R.; Linden, J.; Murthi, M. N. & Skoglund, J. (2002). 'iLBC - a linear predictive coder with robustness to packet losses'. In: *Proc. of the 2002 IEEE Speech Coding Workshop*. SCW. Ibaraki, Japan: IEEE, pp. 23–25.

Angkititrakul, P.; Petracca, M.; Sathyanarayana, A. & Hansen, J. H. (2007). 'UT-Drive: Driver Behavior and Speech Interactive Systems for In-Vehicle Environments'. In: *Proc. of the 2007 IEEE Intelligent Vehicles Symposium*. IV. Istanbul, Turkey: IEEE, pp. 566–569.

ANSI/ASA (1997). *Methods for Calculation of the Speech Intelligibility Index*. Standard ANSI/ASA S3.5-1997 (R2017). Washington, DC, USA: American National Standards.

Aristotle & McKeon, R. (1941). *The basic works of Aristotle*. New York: Random House.

Artstein, R. & Poesio, M. (2008). 'Inter-Coder Agreement for Computational Linguistics'. *Computational Linguistics* 34 (4), pp. 555–596.

ASME (2020). *National Transportation Safety Board Issues New Recommendation for Driver Monitoring in Autonomous Vehicles*. URL: https://www.asme.org/government-relations/capitol-update/national-transportation-

`safety-board-issues-new-recommendation-for-driver-monitoring-in-autonomous-vehicles`.

Avila, A. R.; Akhtar, Z.; Santos, J. F.; O'Shaughnessy, D. & Falk, T. H. (2021). 'Feature Pooling of Modulation Spectrum Features for Improved Speech Emotion Recognition in the Wild'. *IEEE Transactions on Affective Computing* 12 (1), pp. 177–188.

Avila, A. R.; Alam, M. J.; O'Shaughnessy, D. & Falk, T. (2018). 'Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition'. In: *Proc. of the 19th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Hyderabad, India: International Speech Communication Association (ISCA), pp. 3663–3667.

Avots, E.; Sapiński, T.; Bachmann, M. & Kamińska, D. (2019). 'Audiovisual emotion recognition in wild'. *Machine Vision and Applications* 30 (5), pp. 975–985.

Babel, M. & Munson, B. (2014). 'Producing Socially Meaningful Linguistic Variation'. In: *The Oxford Handbook of Language Production.* Ed. by Goldrick, M.; Ferreira, V. S. & Miozzo, M. Oxford University Press, pp. 308–325.

Banchhor, S.; Dodia, J. & Gowda, D. (2013). 'GUI Based Performance Analysis of Speech Enhancement Techniques'. *International Journal of Scientific and Research Publications* 3 (9), pp. 1–7.

Bänziger, T.; Pirker, H. & Scherer, K. R. (2006). 'GEMEP – GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions'. In: *Proc. of the 5th International Conference on Language Resources and Evaluation Workshops.* LREC. Genoa, Italy: European Language Resource Association (ELRA), pp. 15–19.

Bashirpour, M. & Geravanchizadeh, M. (2018). 'Robust emotional speech recognition based on binaural model and emotional auditory mask in noisy environments'. *EURASIP Journal on Audio, Speech, and Music Processing* 2018.9, pp. 1–13.

Bassano, C.; Ballestin, G.; Ceccaldi, E.; Larradet, F. I.; Mancini, M.; Volta, E. & Niewiadomski, R. (2019). 'A VR Game-based System for Multimodal Emotion Data Collection'. In: *Proc. of the 12th annual ACM SIGGRAPH conference on Motion, Interaction and Games.* MIG. Newcastle upon Tyne, United Kingdom: ACM, pp. 1–3.

Batliner, A.; Hacker, C.; Steidl, S.; Nöth, E.; D'Arcy, S.; Russell, M. & Wong, M. (2004). '"You stupid tin box" - children interacting with the AIBO robot:A cross-linguistic emotional speech corpus'. In: *Proc. of the 4th international conference on Language Resources and Evaluation.* LREC. Lisbon, Portugal: European Language Resources Association (ELRA), pp. 171–174.

Baveye, Y.; Dellandréa, E.; Chamaret, C. & Chen, L. (2015). 'Deep learning vs. kernel methods: Performance for emotion prediction in videos'. In: *Proc. of the 2015 International Conference on Affective Computing and Intelligent Interaction.* ACII. Xian, China: IEEE, pp. 77–83.

Becker-Asano, C. (2008). 'WASABI: Affect Simulation for Agents with Believable Interactivity'. PhD thesis. Bielefeld: Universität Bielefeld.

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour.* Princeton University Press.

Benesty, J.; Chen, J.; Huang, Y. & Cohen, I. (2009). *Noise Reduction in Speech Processing.* Ed. by Benesty, J. & Kellermann, W. Vol. 2. Springer Topics in Signal Processing. Springer-Verlag Berlin Heidelberg.

Benesty, J.; Chen, J. & Pan, C. (2016). *Fundamentals of Differential Beamforming.* SpringerBriefs in Electrical and Computer Engineering. Springer Singapore.

Benesty, J.; Sondhi, M. M. & Huang, Y. (2008). *Springer Handbook of Speech Processing.* Springer Handbooks. Springer-Verlag Berlin Heidelberg.

Bennett, E. M.; Alpert, R. & Goldstein, A. C. (1954). 'Communications Through Limited-Response Questioning'. *Public Opinion Quarterly* 18 (3), pp. 303–308.

Bergstra, J. & Bengio, Y. (2012). 'Random Search for Hyper-Parameter Optimization'. *Journal of Machine Learning Research* 13.10, pp. 281–305.

Bleymüller, J. & Weißbach, R. (2015). *Statistik für Wirtschaftswissenschaftler.* 17. Auflage. Munich, Germany: Franz Vahlen.

Blum, A. & Mitchell, T. (1998). 'Combining Labeled and Unlabeled Data with Co-Training'. In: *Proc. of the 11th Annual Conference on Computational Learning Theroy.* COLT. Madison, Wisconsin, USA: ACM, pp. 92–100.

Böck, R.; Egorow, O.; Siegert, I. & Wendemuth, A. (2017). 'Comparative Study on Normalisation in Emotion Recognition from Speech'. In: *Intelligent Human Computer Interaction.* Ed. by Horain, P.; Achard, C. & Mallem, M. Vol. 10688. Lecture Notes in Computer Science (LNCS). Springer, Cham, pp. 189–201.

Böck, R.; Siegert, I.; Haase, M.; Lange, J. & Wendemuth, A. (2011). 'ikannotate – A Tool for Labelling, Transcription, and Annotation of Emotionally Coloured

Speech'. In: *Affective Computing and Intelligent Interaction.* Ed. by D'Mello, S.; Graesser, A.; Schuller, B. & Martin, J.-C. Vol. 6974. Lecture Notes on Computer Science (LNCS). Berlin, Heidelberg, Germany: Springer, pp. 25–34.

Boll, S. F. (1979). 'Suppression of Acoustic Noise in Speech Using Spectral Subtraction'. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27 (2), pp. 113–120.

Bořil, H.; Sadjadi, S. O.; Kleinschmidt, T. & Hansen, J. H. L. (2010). 'Analysis and Detection of Cognitive Load and Frustration in Drivers' Speech'. In: *Proc. of the 11th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Makuhari, Chiba, Japan: International Speech Communication Association (ISCA), pp. 502–505.

Bortz, J. & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung - Leitfaden für die verteilungsfreie Analyse kleiner Stichproben.* 3rd edition. Springer Medizin Verlag Heidelberg.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler.* 7th edition. Springer-Verlag Berlin Heidelberg.

Botinhao, C. V. & Yamagishi, J. (2017). 'Speech Intelligibility in Cars: The Effect of Speaking Style, Noise and Listener Age'. In: *Proc. of the 18th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA), pp. 2944–2948.

Bradley, M. M.; Greenwald, M. K.; Petry, M. C. & Lang, P. J. (1992). 'Remembering Pictures: Pleasure and Arousal in Memory'. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18 (2), pp. 379–390.

Bradley, M. M. & Lang, P. J. (1994). 'Measuring emotion: The self-assessment manikin and the semantic differential'. *Journal of Behavior Therapy and Experimental Psychiatry* 25 (1), pp. 49–59.

Brandenburg, K. (1999). 'MP3 and AAC Explained'. In: *Proc. of the 17th AES International Conference: High-Quality Audio Coding.* Florence, Italy.

Brandenburg, K.; Eberlein, E.; Herre, J. & Edler, B. (1992). 'Comparison of Filterbanks for High Quality Audio Coding'. In: *Proc. of the 1992 IEEE International Symposium on Circuits and Systems.* ISCAS. San Diego, CA, USA: IEEE, pp. 1336–1339.

Braun, M.; Schubert, J.; Pfleging, B. & Alt, F. (2019). 'Improving Driver Emotions with Affective Strategies'. *Multimodal Technologies and Interaction* 3 (1), pp. 1–19.

Breiman, L. (1996). 'Bagging Predictors'. *Machine Learning* 24, pp. 123–140.

— (2001). 'Random Forests'. *Machine Learning* 45, pp. 5–32.

Breiman, L.; Friedman, J.; Stone, C. J. & Olshen, R. (1984). *Classification and Regression Trees.* 1st edition. Springer Series in Statistics. Chapman and Hall/CRC.

Breitenstein, C.; Van Lancker, D. & Daum, I. (2001). 'The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample'. *Cognition and Emotion* 15 (1), pp. 57–79.

Broekens, J. & Brinkman, W.-P. (2009). 'AffectButton: Towards a standard for dynamic affective user feedback'. In: *Proc. of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops.* ACII. Amsterdam, Netherlands: IEEE, pp. 1–8.

— (2013). 'AffectButton: A method for reliable and valid affective self-report'. *International Journal of Human-Computer Studies* 71 (6), pp. 641–667.

Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W. & Weiss, B. (2005). 'A Database of German Emotional Speech'. In: *Proc. of the 9th European Conference on Speech Communication and Technology.* INTERSPEECH. Lisbon, Portugal: International Speech Communication Association (ISCA), pp. 1517–1520.

Burns, K. L. & Beier, E. G. (1973). 'Significance of Vocal and Visual Channels in the Decoding of Emotional Meaning'. *Journal of Communication* 23 (1), pp. 118–130.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S. & Narayanan, S. S. (2008). 'IEMOCAP: Interactive emotional dyadic motion capture database'. *Language Resources and Evaluation* 42, pp. 335–359.

Busso, C.; Metallinou, A. & Narayanan, S. S. (2011). 'Iterative feature normalization for emotional speech detection'. In: *Proc. of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing.* ICASSP. Prague, Czech Republic: IEEE, pp. 5692–5695.

Bynion, T.-M. & Feldner, M. T. (2017). 'Self-Assessment Manikin'. In: *Encyclopedia of Personality and Individual Differences.* Ed. by Zeigler-Hill, V. & Shackelford, T. K. Springer, Cham. Chap. S, pp. 1–3.

Cairns, D. A. & Hansen, J. H. L. (1994). 'Nonlinear analysis and classification of speech under stressed conditions'. *Journal of the Acoustical Society of America* 96 (6), pp. 3392–3400.

Calem, R. (2019). *Health Sensing in Cars.* URL: `https://cta.tech/Resources/i3-Magazine/i3-Issues/2019/November-December/Health-Sensing-in-Cars`.

Cevher, D.; Zepf, S. & Klinge, R. (2019). 'Towards Multimodal Emotion Recognition in German Speech Events in Cars using Transfer Learning'. In: *Proc. of the 15th Conference on Natural Language Processing.* KONVENS. Nürnberg, Germany: German Society for Computational Linguistics & Language Technology, pp. 79–90.

Chandrasekar, P.; Chapaneri, S. & Jayaswal, D. D. (2014). 'Automatic Speech Emotion Recognition: A Survey'. In: *Proc. of the 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications.* CSCITA. Mumbai, India: IEEE, pp. 341–346.

Chen, J.-H. & Thyssen, J. (2008). 'Analysis-by-Synthesis Speech Coding'. In: *Springer Handbook of Speech Processing.* Ed. by Benesty, J.; Sondhi, M. M. & Huang, Y. Springer Handbooks. Springer-Verlag Berlin Heidelberg. Chap. 17, pp. 351–392.

Chen, J.-H. (2006). 'Novel Codec Structures for Noise Feedback Coding of Speech'. In: *Proc. of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings.* ICASSP. Toulouse, France: IEEE, pp. 1–4.

Chen, L.; Mao, X.; Xue, Y. & Cheng, L. L. (2012). 'Speech emotion recognition: Features and classification models'. *Digital Signal Processing* 22 (6), pp. 1154–1160.

Chen, M. & Hao, Y. (2020). 'Label-less Learning for Emotion Cognition'. *IEEE Transactions on Neural Networks and Learning Systems* 31 (7), pp. 2430–2440.

Chenchah, F. & Lachiri, Z. (2016). 'Speech emotion recognition in noisy environment'. In: *Proc. of the 2nd International Conference on Advanced Technologies for Signal and Image Processing.* ATSIP. Monastir, Tunisia: IEEE, pp. 788–792.

Cherkassky, V. & Ma, Y. (2004). 'Practical selection of SVM parameters and noise estimation for SVM regression'. *Neural Networks* 17 (1), pp. 113–126.

Chhikara, J. & Singh, J. (2012). 'Noise cancellation using adaptive algorithms'. *International Journal of Modern Engineering Research* 2.38 (3), pp. 792–795.

Chiaramello, E.; Moriconi, S. & Tognola, G. (2015). 'Objective Measures of Perceptual Quality for Predicting Speech Intelligibility in Sensorineural Hearing Loss'. In: *Proc. of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* EMBC. Milan, Italy: IEEE, pp. 5577–5580.

Chopade, C. R. (2015). 'Text Based Emotion Recognition: A Survey'. *International Journal of Science and Research* 4 (6), pp. 409–414.

Chuang, Z.-J. & Wu, C.-H. (2004). 'Multi-Modal Emotion Recognition from Speech and Text'. *Computational Linguistics and Chinese Language Processing* 9.2 (Special Issue on New Trends of Speech and Language Processing), pp. 45–62.

Cicchetti, D. V. (1994). 'Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology'. *Psychological Assessment* 6 (4), pp. 284–290.

Clarke, B.; Fokoue, E. & Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning.* Springer Series in Statistics. Springer, New York, NY.

Clynes, M. & Nettheim, N. (1982). 'The Living Quality of Music: Neurobiologic Patterns of Communicating Feeling'. In: *Music, Mind, and Brain: The Neuropsychology of Music.* Ed. by Clynes, M. Springer, Boston, MA. Chap. IV, pp. 47–82.

Coalson, J. & Xiph.Org Foundation (2020). *flac: free lossless audio codec.* Xiph.Org Foundation. URL: `https://xiph.org/flac/index.html` (visited on 08/02/2020).

Cohen, I. & Gannot, S. (2008). 'Springer Handbook of Speech Processing'. In: ed. by Benesty, J.; Sondhi, M. & Huang, Y. Springer Handbooks. Springer, Berlin, Heidelberg. Chap. Spectral Enhancement Methods, pp. 873–901.

Cohen, I. (2005). 'Relaxed Statistical Model for Speech Enhancement and a Priori SNR Estimation'. *IEEE Transactions on Speech and Audio Processing* 13.5, pp. 870–881.

Cohen, I. & Berdugo, B. (2001). 'Speech enhancement for non-stationary noise environments'. *Signal Processing* 81, pp. 2403–2418.

Cohen, J. (1960). 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement* 20 (1), pp. 37–46.

— (1968). 'Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit'. *Psychological Bulletin* 70 (4), pp. 213–220.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W. & Taylor, J. (2001). 'Emotion Recognition in Human-Computer Interaction'. *IEEE Signal Processing Magazine* 18 (1), pp. 32–80.

Cowie, R.; Douglas-Cowie, E.; Savvidou, S.; McMahon, E.; Sawey, M. & Schröder, M. (2000). "FEELTRACE': An instrument for recording perceived emotion in real time'. In: *Proc. of the Speech and Emotion, ISCA Tutorial and Research Workshop.* ITRW. Newcastle, Northern Ireland, UK: International Speech Communication Association (ISCA), pp. 19–24.

D'Agostino, R. B. (2006). 'Tests for Departures from Normality'. In: *Encyclopedia of Statistical Sciences.* Ed. by Samuel Kotz, S.; Read, C. B.; Balakrishnan, N. & Vidakovic, B. John Wiley & Sons Ltd.

Dahlgren, A.; Kecklund, G. & Akerstedt, T. (2005). 'Different levels of work-related stress and the effects on sleep, fatigue and cortisol'. *Scandinavian Journal of Work, Environment and Health* 31.4, pp. 277–285.

Dan-Glauser, E. S. & Scherer, K. R. (2011). 'The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance'. *Behavior Research Methods* 43 (2), pp. 468–477.

Daneshfar, F. & Kabudian, S. J. (2020). 'Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm'. *Multimedia Tools and Applications* 79, pp. 1261–1289.

Davies, M. & Fleiss, J. L. (1982). 'Measuring Agreement for Multinomial Data'. *Biometrics* 38 (4), pp. 1047–1051.

Davitz, J. R. & Beldoch, M. (1964). *The communication of emotional meaning.* McGraw-Hill series in psychology. McGraw-Hill.

Deng, L. & Li, X. (2013). 'Machine Learning Paradigms for Speech Recognition: An Overview'. *IEEE Transactions on Audio, Speech, and Language Processing* 21 (5), pp. 1060–1089.

Desmet, P. M. A.; Porcelijn, R. & Dijk, M. B. van (2007). 'Emotional Design; Application of a Research-Based Design Approach'. *Knowledge, Technology & Policy* 20 (3), pp. 141–155.

Desplanques, B. & Demuynck, K. (2018). 'Cross-lingual Speech Emotion Recognition through Factor Analysis'. In: *Proc. of the 19th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Hyderabad, India: International Speech Communication Association (ISCA), pp. 3648–3652.

Devillers, L. & Vasilescu, I. (2004). 'Reliability of Lexical and Prosodic Cues in two Real-life Spoken Dialog Corpora'. In: *Proc. of the 4th International Conference on Language Resources and Evaluation.* LREC. Lisbon, Portugal: European Language Resources Association (ELRA), pp. 1423–1426.

Devillers, L.; Vidrascu, L. & Lamel, L. (2005). 'Challenges in real-life emotion annotation and machine learning based detection'. *Neural Networks* 18 (4), pp. 407–422.

Dhall, A.; Goecke, R.; Joshi, J.; Sikka, K. & Gedeon, T. (2014). 'Emotion Recognition In The Wild Challenge 2014: Baseline, Data and Protocol'. In: *Proc. of the 16th International Conference on Multimodal Interaction.* ICMI. Istanbul, Turkey: ACM, pp. 461–466.

Dhall, A.; Goecke, R.; Lucey, S. & Gedeon, T. (2011). *Acted Facial Expressions In The Wild Database.* Technical Report TR-CS-11-02. Canberra, Australia: The Australian National University.

Dikecligil, G. N. & Mujica-Parodi, L. R. (2010). 'Ambulatory and Challenge-Associated Heart Rate Variability Measures Predict Cardiac Responses to Real-World Acute Emotional Stress'. *Biological Psychiatry* 67 (12), pp. 1185–1190.

DLR (9th Feb. 2011). *Das Auto der Zukunft kommt auf Knopfdruck – Fahrdemonstration mit dem FASCar II.* Deutsches Zentrum für Luft- und Raumfahrt. URL: `https://www.dlr.de/fs/en/desktopdefault.aspx/tabid-10714/18622_read-43366/` (visited on 04/05/2019).

— (2019). *FASCar - test vehicle for assistance and automation.* Deutsches Zentrum für Luft- und Raumfahrt. URL: `http://www.dlr.de/ts/en/`

`desktopdefault.aspx/tabid-11367/19950_read-46557/` (visited on 04/05/2019).

Dmitrieva, E. S. & Gelman, V. Y. (2012). 'The Relationship between the Perception of Emotional Intonation of Speech in Conditions of Interference and the Acoustic Parameters of Speech Signals in Adults of Different Gender and Age'. *Neuroscience and Behavioral Physiology* 42 (8), pp. 920–928.

Docherty, G. & Mendoza-Denton, N. (2011). 'Speaker-Related Variation–Sociophonetic Factors'. In: *The Oxford Handbook of Laboratory Phonology.* Ed. by Cohn, A. C.; Fougeron, C. & Huffman, M. K. Oxford University Press. Chap. 4, pp. 43–60.

Doclo, S. & Moonen, M. (2003). 'Design of far-field and near-field broadband beamformers using eigenfilters'. *Signal Processing* 83 (12), pp. 2641–2673.

Dong, H.-Y. & Lee, C.-M. (2018). 'Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering'. *EURASIP Journal on Audio, Speech, and Music Processing* 2018.3, pp. 1–13.

Drewitz, U.; Kaul, R.; Jipp, M. & Ihme, K. (2017). 'Verstehende und mitdenkende Assistenz für hochautomatisierte Fahrzeuge'. In: *Proc. of the VDI-Fachtagung: Der Fahrer im 21.Jahunder.* Braunschweig, Germany: VDI.

Dusenburg, D. & Knower, F. H. (1939). 'Experimental studies of the symbolism of action and voice—II: A study of the specificity of meaning in abstract tonal symbols'. *Quarterly Journal of Speech* 25 (1), pp. 67–75.

Eerola, T. & Vuoskoski, J. K. (2011). 'A comparison of the discrete and dimensional models of emotion in music'. *Psychology of Music* 39 (1), pp. 18–49.

Egorow, O.; Siegert, I. & Wendemuth, A. (2018). 'Improving Emotion Recognition Performance by Random-Forest-Based Feature Selection'. In: *Speech and Computer.* Ed. by Karpov, A.; Jokisch, O. & Potapova, R. Vol. 11096. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 134–144.

Egorow, O. & Wendemuth, A. (2019). 'On Emotions as Features for Speech Overlaps Classification'. *IEEE Transactions on affective computing*, s.p.

Ekman, P. (1972). 'Universals and cultural differences in facial expressions of emotion'. *Nebraska Symposium on Motivation* 19, pp. 207–283.

— (1999). 'Basic Emotions'. In: *Handbook of Cognition and Emotion.* Ed. by Dalgleish, T. & Power, M. John Wiley & Sons Ltd. Chap. 3, pp. 45–60.

Ekman, P. & Cordaro, D. (2011). 'What is Meant by Calling Emotions Basic'. *Emotion Review* 3 (4), pp. 364–370.

Ekman, P. & Friesen, W. V. (1975). *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions.* Prentice Hall.

El Ayadi, M.; Kamel, M. S. & Karray, F. (2011). 'Survey on speech emotion recognition: Features, classification schemes, and databases'. *Pattern Recognition* 44 (3), pp. 572–587.

Engberg, I. S.; Hansen, A. V.; Andersen, O. & Dalsgaard, P. (1997). 'Design, Recording and Verification of a Danish Emotional Speech Database'. In: *Proc. of the 5th European Conference on Speech Communication and Technology.* EUROSPEECH. Rhodes, Greece: International Speech Communication Association (ISCA), pp. 1695–1698.

Ephraim, Y. & Malah, D. (1985). 'Speech enhancement using a minimum mean-square error log-spectral amplitude estimator'. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33.2, pp. 443–445.

Eppinger, B. & Herter, E. (1993). *Sprachverarbeitung.* Munich, Germany: Carl-Hanser-Verlag.

Eyben, F.; Wöllmer, M. & Schuller, B. (2009). 'openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit'. In: *Proc. of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops.* ACII. Amsterdam, Netherlands: IEEE, pp. 1–6.

— (2010). 'openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor'. In: *Proc. of the 18th ACM international conference on Multimedia.* MM. Firenze, Italy: ACM, pp. 1459–1462.

Eyben, F. et al. (2016). 'The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing'. *IEEE Transactions on Affective Computing* 7 (2), pp. 190–202.

Fischer, M.; Richter, A.; Schindler, J.; Plättner, J.; Temme, G.; Kelsch, J.; Assmann, D. & Köster, F. (2014). 'Modular and Scalable Driving Simulator Hardware and Software for the Development of Future Driver Assistence and Automation Systems'. In: *New Developments in Driving Simulation Design and Experiments.* Driving Simulation Conference. Paris, France, pp. 223–229.

Fleiss, J. L. (1971). 'Measuring nominal scale agreement among many raters'. *Psychological Bulletin* 76 (5), pp. 378–382.

Fleiss, J. L.; Levin, B. & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*. 3rd edition. Wiley Series in Probability and Statistics. John Wiley & Sons.

Flemisch, F.; Meier, S.; Neuhöfer, J.; Baltzer, M.; Altendorf, E. & Özyurt, E. (2013). 'Kognitive und kooperative Systeme in der Fahrzeugführung: Selektiver Rückblick über die letzten Dekaden und Spekulation über die Zukunft'. *Kognitive Systeme* 2013 (1), pp. 1–10.

Fontaine, J. R.; Scherer, K. R.; Roesch, E. B.; Ellsworth & C., P. (2007). 'The World of Emotions Is Not Two-Dimensional'. *Psychological Science* 18.12, pp. 1050–1057.

Fontenla-Romero, Ó.; Martinez-Rego, D.; Guijarro-Berdiñas, B.; Pérez-Sánchez, B. & Peteiro-Barral, D. (2013). 'Online Machine Learning'. In: *Efficiency and Scalability Methods for Computational Intellect*. Ed. by Igelnik, B. & Zurada, J. M. IGI Global. Chap. 2, pp. 27–54.

Fox, E.; Cahill, S. & Zougkou, K. (2010). 'Preconscious Processing Biases Predict Emotional Reactivity to Stress'. *Biological Psychiatry* 67 (4), pp. 371–377.

Franke, T.; Attig, C. & Wessel, D. (2017). 'Affinity for technology interaction - a personal-resource perspective'. In: *Proc. of Human Factors and Ergonomics Society Europe Chapter 2017 Annual Conference*. Rome, Italy.

Freese, M. & Jipp, M. (2015). 'Zwischen Rational und Emotional – Ein Überblick über Entscheidungen und deren Einflussgrößen in kooperierenden Teams'. *Kognitive Systeme* 2015 (2), pp. 1–11.

French, N. R. & Steinberg, J. C. (1947). 'Factors Governing the Intelligibility of Speech Sounds'. *Journal of the Acoustical Society of America* 19 (1), pp. 90–119.

Frick, R. W. (1985). 'Communicating Emotion: The Role of Prosodic Features'. *Psychological Bulletin* 97 (3), pp. 412–429.

Gadsden, S. A. & Habibi, S. R. (2009). 'The Future of Automobiles: Cognitive Cars'. In: *Proc. of the 22nd Canadian Congress of Applied Mechanics*. CANCAM. Halifax, Nova Scotia, Canada: Dalhousie University, pp. 111–112.

Gagge, A.; Stolwijk, J. & Hardy, J. (1967). 'Comfort and thermal sensations and associated physiological responses at various ambient temperatures'. *Environmental Research* 1.1, pp. 1–20.

Gallardo, L. F.; Möller, S. & Beerends, J. (2017). 'Predicting Automatic Speech Recognition Performance Over Communication Channels from Instrumental Speech Quality and Intelligibility Scores'. In: *Proc. of the 18th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA), pp. 2939–2943.

Gao, H.; Yüce, A. & Thiran, J.-P. (2014). 'Detecting emotional stress from facial expressions for driving safety'. In: *Proc. of the 2014 IEEE International Conference on Image Processing.* ICIP. Paris, France: IEEE, pp. 5961–5965.

García, N.; Vásquez-Correa, J. C.; Arias-Londoño, J. D.; Várgas-Bonilla, J. F. & Orozco-Arroyave, J. R. (2015). 'Automatic Emotion Recognition in Compressed Speech Using Acoustic and Non-Linear Features'. In: *Proc. of the 20th Symposium on Signal Processing, Images and Computer Vision.* STSIVA. Bogotá, Columbia: IEEE, pp. 1–7.

Gelderblom, F. B.; Tronstad, T. V. & Viggen, E. M. (2019). 'Subjective Evaluation of a Noise-Reduced Training Target for Deep Neural Network-Based Speech Enhancement'. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27 (3), pp. 583–594.

Girard, J. M. (2014). 'CARMA: Software for Continuous Affect Rating and Media Annotation'. *Journal of Open Research Software* 2.e5 (1), pp. 1–6.

Gobl, C. & Chasaide, A. N. (2003). 'The role of voice quality in communicating emotion, mood and attitude'. *Speech Communication* 40 (1-2), pp. 189–212.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley.

Goli, P. & Karami-Mollaei, M. R. (2016). 'Speech Intelligibility Improvement in Noisy Environments for Near End Listening Enhancement'. *Journal of Information Systems and Telecommunication* 4 (1), pp. 27–33.

Gomez-Clapers, J. & Casanella, R. (2012). 'A Fast and Easy-to-Use ECG Acquisition and Heart Rate Monitoring System Using a Wireless Steering Wheel'. *IEEE Sensors Journal* 12 (3), pp. 610–616.

Grasberger, L. (2013). *Kampfmaschine Auto.* URL: `https://www.gdv.de/resource/blob/42812/b79becb7356f50aa253683e4985e574e/positionen-88-maerz-2013-pdf-data.pdf` (visited on 24/04/2021).

Grimm, M.; Kroschel, K.; Harris, H.; Nass, C.; Schuller, B.; Rigoll, G. & Moosmayr, T. (2007). 'On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving'. In: *Affective Computing and Intelligent Interaction*. Ed. by Paiva, A. C. R.; Prada, R. & Picard, R. W. Vol. 4738. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 126–138.

Grimm, M.; Kroschel, K. & Narayanan, S. (2008). 'The Vera am Mittag German audio-visual emotional speech database'. In: *Proc. of the 2008 IEEE International Conference on Multimedia and Expo*. ICME. Hannover, Germany: IEEE, pp. 865–868.

Grimm, M.; Kroschel, K.; Schuller, B.; Rigoll, G. & Moosmayr, T. (2007). 'Acoustic Emotion Recognition in Car Environment Using a 3D EmotionSpace Approach'. In: *Fortschritte der Akustik - DAGA 2007*. Ed. by Mehra, S.-R. & Leistner, P. Vol. 33. Dautesche Jahrestagung für Akustik. Deutsche Gesellschaft für Akustik e.V. (DEGA), pp. 313–314.

Grimme, C. & Bossek, J. (2018). *Einführung in die Optimierung - Konzepte, Methoden und Anwendungen*. Springer Vieweg.

Haider, F.; Pollak, S.; Albert, P. & Luz, S. (2021). 'Emotion recognition in low-resource settings: An evaluation of automatic feature selection methods'. *Computer Speech & Language* 65, pp. 1–10.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I. H. (2009). 'The WEKA Data Mining Software: An Update'. *ACM-SIGKDD Exploitations Newsletter* 11.1, pp. 10–18.

Hallgren, K. A. (2012). 'Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial'. *Tutorials in Quantitative Methods for Psychology* 8 (1), pp. 23–34.

Han, W.; Li, H.; Ruan, H.; Ma, L.; Sun, J. & Schuller, B. (2013). 'Active Learning for Dimensional Speech Emotion Recognition'. In: *Proc. of the 14th Annual Conference of the International Speech Communication Association*. INTERSPEECH. Lyon, France: International Speech Communication Association (ISCA), pp. 2841–2845.

Hansen, J. H. L. & Bou-Ghazale, S. E. (1997). 'Getting Started With SUSAS: A Speech Under Simulated And Actual Stress Database'. In: *Proc. of the 5th European Conference on Speech Communication and Technology*. EUROSPEECH. Rhodes, Greece: International Speech Communication Association (ISCA), pp. 1743–1746.

Haq, S. & Jackson, P. J. B. (2010). 'Multimodal Emotion Recognition'. In: *Machine Audition: Principles, Algorithms and Systems.* Ed. by Wang, W. IGI Global. Chap. 17, pp. 398–423.

Harimi, A.; AhmadyFard, A.; Shahzadi, A. & Yaghmaie, K. (2015). 'Anger or Joy? Emotion Recognition Using Nonlinear Dynamics of Speech'. *Applied Artificial Intelligence* 29 (7), pp. 675–696.

Harimi, A. & Esmaileyan, Z. (2014). 'A Database for Automatic Persian Speech Emotion Recognition: Collection, Processing and Evaluation'. *International Journal of Engineering* 27 (1), pp. 79–90.

Harmon-Jones, E.; Amodio, D. M. & Zinner, L. R. (2007). 'Social psychological methods of emotion elicitation'. In: *Handbook of Emotion Elicitation and Assessment.* Ed. by Coan, J. A. & Allen, J. J. B. Series in Affective Science. Oxford University Press. Chap. 6, pp. 91–105.

Harmon-Jones, E. & Sigelman, J. (2001). 'State anger and prefrontal brain activity: Evidence that insult-related relative left-prefrontal activation is associated with experienced anger and aggression'. *Journal of Personality and Social Psychology* 80 (5), pp. 797–803.

Harrigan, J.; Rosenthal, R. & Scherer, K. R. (2005). *The New Handbook of Methods in Nonverbal Behavior Research.* Series in Affective Science. Oxford University Press.

Hastie, T.; Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction.* 2nd edition. Springer Series in Statistics. Springer Science+Business Media.

Healey, J. A.; Nachman, L.; Subramanian, S.; Shahabdeen, J. & Morris, M. (2010). 'Out of the Lab and into the Fray: Towards Modeling Emotion in Everyday Life'. In: *Pervasive Computing.* Ed. by Floréen, P.; Krüger, A. & Spasojevic, M. Vol. 6030. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 156–173.

Healey, J. A. & Picard, R. W. (2005). 'Detecting Stress During Real-World Driving Tasks Using Physiological Sensors'. *IEEE Transactions on Intelligent Transportation Systems* 6 (2), pp. 156–166.

Heide, A. & Henning, K. (2006). 'The "cognitive car": A roadmap for research issues in the automotive sector'. *Annual Reviews in Control* 30 (2), pp. 197–203.

Hernandez, J.; McDuff, D.; Benavides, X.; Amores, J.; Maes, P. & Picard, R. W. (2014). 'AutoEmotive: Bringing Empathy to the Driving Experience to Manage Stress'. In: *Proc. of the Designing Interactive Systems Conference 2014.* DIS. Vancouver, BC, Canada: ACM, pp. 53–56.

Herre, J. & Lutzky, M. (2008). 'Perceptual Audio Coding of Speech Signals'. In: *Springer Handbook of Speech Processing.* Ed. by Benesty, J.; Sondhi, M. M. & Huang, Y. Springer Handbooks. Springer-Verlag Berlin Heidelberg. Chap. 18, pp. 393–410.

Hodgson, M. & Nosal, E.-M. (2002). 'Effect of noise and occupancy on optimal reverberation times for speech intelligibility in classrooms'. *Journal of the Acoustical Society of America* 111 (2), pp. 931–939.

Holzapfel, H. & Fuegen, C. (2002). 'Integrating Emotional Cues into a Framework for Dialogue Management'. In: *Proc. of the 4th IEEE International Conference on Multimodal Interfaces (ICMI'02).*

Hoque, M. E.; McDuff, D. J. & Picard, R. W. (2012). 'Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles'. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING* 3 (3), pp. 323–334.

Hornsby, B. W. Y. (2004). 'The Speech Intelligibility Index: What is it and what's it good for?' *The Hearing Journal* 57 (10), pp. 10–17.

Hornstein, M. G. (1967). 'Accuracy of emotional communication and interpersonal compatibility'. *Journal of Personality* 35 (1), pp. 20–30.

Houtgast, T. & Steeneken, H. J. M. (1973). 'The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility'. *Acta Acustica united with Acustica* 28.1, pp. 66–73.

Hsu, C.-w.; Chang, C.-c. & Lin, C.-j. (2016). *A practical guide to support vector classification.*

Hübner, D.; Vlasenko, B.; Grosser, T. & Wendemuth, A. (2010). 'Determining Optimal Features for Emotion Recognition from Speech by Applying an Evolutionary Algorithm'. In: *Proc. of the 11th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Makuhari, Chiba, Japan: International Speech Communication Association (ISCA), pp. 2358–2361.

IBM Corporation & Microsoft Corporation (1991). *Multimedia Programming Interface and Data Specifications 1.0.* Tech. rep.

Ihme, K.; Dömeland, C.; Freese, M. & Jipp, M. (2018). 'Frustration in the face of the driver: A simulator study on facial muscle activity during frustrated driving'. *Interaction Studies* 19.3, pp. 487–498.

Ihme, K.; Unni, A.; Zhang, M.; Rieger, J. W. & Jipp, M. (2018). 'Recognizing Frustration of Drivers From Face Video Recordings and Brain Activation Measurements With Functional Near-Infrared Spectroscopy'. *Frontiers in Human Neuroscience* 12.327, pp. 1–15.

ISO/IEC (1993). *Information technology - Coding of moving pictures and associated audio for digital storage media up to about 1,5 Mbit/s - Part 3:Audio.* ISO/IEC 11172-3:1993. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

— (1998). *Information technology - Generic coding of moving pictures and associated audio information - Part 3:Audio.* ISO/IEC 13818-3:1998. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

— (2006). *Information technology - Generic coding of moving pictures and associated audio information - Part 7:Advanced Audio Coding.* ISO/IEC 13818-7:2006. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

ITU-T (1996). *Methods for subjective determination of transmission quality.* Recommendation P.800 (08/96). Geneva, Swiss: International Telecommunication Union (Telecommunication Standardization Sector).

— (2003a). *One-way transmission time.* Recommendation G.114 (05/03). Geneva, Swiss: International Telecommunication Union (Telecommunication Standardization Sector).

— (2003b). *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm.* Recommendation P.835 (11/2003). Geneva, Swiss: International Telecommunication Union (Telecommunication Standardization Sector).

— (2003c). *Wideband Coding of Speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB).* REC G.722.2. International Telecommunication Union (Telecommunication Standardization Sector).

— (2009). *G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729.* Recommendation G.729.1

(05/06). Geneva, Swiss: International Telecommunication Union (Telecommunication Standardization Sector).

ITU-T (2014). *Perceptual objective listening quality assessment.* Recommendation P.863 (09/2014). Geneva, Swiss: International Telecommunication Union (Telecommunication Standardization Sector).

— (2018). *Perceptual objective listening quality assessment.* Recommendation P.863 (03/2018). Geneva, Swiss: International Telecommunication Union (Telecommunication Standardization Sector).

Ivanov, A. & Riccardi, G. (2012). 'Kolmogorov-Smirnov test for feature selection in emotion recognition from speech'. In: *Proc. of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing.* ICASSP. Kyoto, Japan: IEEE, pp. 5125–5128.

Izard, C. E. (1977). *Human Emotions.* Emotions, Personality, and Psychotherapy. Springer US.

Izard, C. E.; Libero, D. Z.; Putnam, P. & Haynes, O. M. (1993). 'Stability of Emotion Experiences and Their Relations to Traits of Personality'. *Journal of Personality and Social Psychology* 64 (5), pp. 847–860.

Jack, R. E.; Blais, C.; Scheepers, C.; Schyns, P. G. & Caldara, R. (2009). 'Cultural Confusions Show that Facial Expressions Are Not Universal'. *Current Biology* 19 (18), pp. 1543–1548.

James, W. (1884). 'What is an Emotion?' *Mind* 9, pp. 188–205.

Jeon, M. (2015). 'Towards affect-integrated driving behaviour research'. *Theoretical Issues in Ergonomics Science* 16 (6), pp. 553–585.

Jeon, M.; Walker, B. N. & Yim, J.-B. (2014). 'Effects of specific emotions on subjective judgment, driving performance, and perceived workload'. *Transportation Research Part F: Traffic Psychology and Behaviour* 24, pp. 197–209.

Jones, C. M. & Jonsson, I.-M. (2005). 'Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses'. In: *Proc. of the 17th Australia conference on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future.* OZCHI. Canberra, Australia: ACM, pp. 1–10.

— (2007). 'Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces'. In: *Universal Access in Human-Computer Interaction*

- *Ambient Interaction.* Ed. by Stephanidis, C. Vol. 4555. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 411–420.

Juslin, P. N. & Sloboda, J. A. (2001). *Music and Emotion: Theory and Research.* Series in Affective Science. Oxford University Press.

Kaiser, J. F. (1990). 'On a simple algorithm to calculate the 'energy' of a signal'. In: *Proc. of the 1990 International Conference on Acoustics, Speech, and Signal Processing.* ICASSP. Albuquerque, New Mexico, USA: IEEE, pp. 381–384.

Katsis, C. D.; Katertsidis Nikolaos nd Ganiatsas, G. & Fotiadis, D. I. (2008). 'Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach'. *IEEE Transactions on Systems, Man, and Cyberenetics - Part A: Systems and Humans* 38 (3), pp. 502–512.

Kaya, H. & Salah, A. A. (2016). 'Combining modality-specific extreme learning machines for emotion recognition in the wild'. *Journal on Multimodal User Interfaces* 10 (2), pp. 139–149.

Keogh, E. & Mueen, A. (2017). 'Curse of Dimensionality'. In: *Encyclopedia of Machine Learning and Data Mining.* Ed. by Sammut, C. & Webb, G. I. Springer, Boston, MA. Chap. C, pp. 314–315.

Kim, E. H.; Hyun, K. H.; Kim, S. H. & Kwak, Y. K. (2009). 'Improved Emotion Recognition With a Novel Speaker-Independent Feature'. *IEEE/ASME Transactions on Mechatronics* 14 (3), pp. 317–325.

Kim, J.; Englebienne, G.; Truong, K. P. & Evers, V. (2017). 'Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning'. In: *Proc. of the 18th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA), pp. 1113–1117.

Kim, J. & André, E. (2008). 'Emotion recognition based on physiological changes in music listening'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30 (12), pp. 2067–2083.

Klasmeyer, G. & Sendlmeier, W. F. (1995). 'Objective Voice Parameters to Characterize the Emotional Content in Speech'. In: *Proc. of the 13th International Congress of Phonetic Sciences.* Vol. 1. ICPhS. Stockholm, Sweden: International Phonetic Association, pp. 182–185.

Klein, J.; Moon, Y. & Picard, R. W. (2002). 'This computer responds to user frustration: Theory, design, and results'. *Interacting with Computers* 14 (2), pp. 119–140.

Knower, F. H. (1941). 'Analysis of Some Experimental Variations of Simulated Vocal Expressions of the Emotions'. *The Journal of Social Psychology* 14, pp. 369–372.

Kohavi, R. (1995). 'A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection'. In: *Proc. of the 14th International Joint Conference on Artificial Intelligence (II)*. IJCAI. Montreal, Quebec, Canada: IJCAI, pp. 1137–1143.

Kondo, K. (2012). *Subjective Quality Measurement of Speech - Its Evaluation, Estimation and Applications*. Signals and Communication Technology. Springer-Verlag Berlin Heidelberg.

Konečni, V. J. (2008). 'Does music induce emotion? A theoretical and methodological analysis'. *Psychology of Aesthetics, Creativity, and the Arts* 2 (2), pp. 115–129.

Koo, J.; Kwac, J.; Ju, W.; Steinert, M.; Leifer, L. & Nass, C. (2015). 'Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance'. *International Journal on Interactive Design and Manufacturing* 9 (4), pp. 269–275.

Koo, T. K. & Li, M. Y. (2016). 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research'. *Journal of Chiropractic Medicine* 15 (2), pp. 155–163.

Kords, M. (2021). *Anteil der Pkw mit Fahrassistenzsystemen in Deutschland 2019*. URL: `https://de.statista.com/statistik/daten/studie/1083873/umfrage/anteil-der-pkw-mit-fahrassistenzsystemen-in-deutschland/` (visited on 19/06/2021).

Kramer, E. (1964). 'Elimination of verbal cues in judgments of emotion from voice'. *Journal of Abnormal and Social Psychology* 68 (4), pp. 390–396.

Krauss, R. M.; Curran, N. M. & Ferleger, N. (1983). 'Expressive Conventions and the Cross-Cultural Perception of Emotion'. *Basic and Applied Social Psychology* 4 (4), pp. 295–305.

Kreibig, S. D. (2010). 'Autonomic nervous system activity in emotion: A review'. *Biological Psychology* 84.3, pp. 394–421.

Kreutz, G.; Ott, U.; Teichmann, D.; Osawa, P. & Vaitl, D. (2008). 'Using music to induce emotions: Influences of musical preference and absorption'. *Psychology of Music* 36 (1), pp. 101–126.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology.* 2nd edition. Sage Publications.

Kunche, P. & Reddy, S. M. (2016). *Metaheuristic Applications to Speech Enhancement.* Ed. by Neustein, A. & Lee, F. SpringerBriefs in Speech Technology. Springer International Publishing.

Kvam, P. H. & Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering.* Hoboken, New Jersey: A John Wiley & Sons, Inc.

Kwon, O.-W.; Chan, K.; Hao, J. & Lee, T.-W. (2003). 'Emotion Recognition by Speech Signals'. In: *Proc. of the 8th European Conference on Speech Communication and Technology.* EUROSPEECH/INTERSPEECH. Geneva, Switzerland: International Speech Communication Association (ISCA), pp. 125–128.

Labelle, F.; Lefebvre, R. & Gournay, P. (2016). 'A Subjective Evaluation of the Effects of Speech Coding on the Perception of Emotions'. In: *Proc. of the 2016 International Symposium on Intelligent Signal Processing and Communication Systems.* ISPACS. Phuket, Thailand: IEEE, pp. 1–6.

Laflamme, C.; Adoul, J.-P.; Su*, H. & Morissette, S. (1990). 'On Reducing Computational Complexity of Codeboook Search in CELP Coder Through the Use of Algebraic Codes'. In: *Proc. of the 1990 International Conference on Acoustics, Speech, and Signal Processing.* ICASSP. Barcelona, Spain: IEEE, pp. 177–180.

Lahaie, O.; Lefebvre, R. & Gournay, P. (2017). 'Influence of audio bandwidth on speech emotion recognition by human subjects'. In: *Proc. of the 2017 IEEE Global Conference on Signal and Information Processing.* GlobalSIP. Montreal, Canada: IEEE, pp. 61–65.

Landgraf, R. (2018). 'Die Effekte kommunikationsunterstützender Systeme auf natürliche Dialogsprache im Auto - eine phonetisch-linguistische Analyse'. PhD thesis. Liel: Christian-Albrechts-Universität zu Kiel.

Landis, J. R. & Koch, G. G. (1977). 'The Measurement of Observer Agreement for Categorical Data'. *Biometrics* 33 (1), pp. 159–174.

Lang, P. J. (1980). 'Behavioral treatment and bio-behavioral assessment: Computer applications'. In: *Technology in Mental Health Care Delivery Systems.* Ed. by

Sidowski, J. B.; Johnson, J. H. & Williams, T. A. Ablex Publishing Corporation, pp. 119–137.

Lang, P. J. (1985). 'The Cognitive Psychophysiology of Emotion: Fear and Anxiety'. In: *Anxiety and the Anxiety Disorders.* Ed. by Tuma, A. H. & Maser, J. 1st edition. Taylor & Francis, pp. 131–170.

Lang, P. J.; Bradley, M. M. & Cuthbert, B. N. (2008). *International Affective Picture System (IAPS): Instruction manual and affective ratings.* Technical Report A-8. Gainesville: The Center for Research in Psychophysiology, University of Florida.

Larradet, F.; Niewiadomski, R.; Barresi, G.; Caldwell, D. G. & Mattos, L. S. (2020). 'Toward Emotion Recognition From Physiological Signals in the Wild: Approaching the Methodological Issues in Real-Life Data Collection'. *Frontiers in Psychology* 11.1111, pp. 1–24.

Laver, J. (1980). 'The phonetic description of voice quality'. *Journal of the International Phonetic Association* 11 (2), pp. 78–84.

— (1994). *Principles of phonetics.* Cambridge University Press.

Lazarus, R. S. (1991). 'Progress on a cognitive-motivational-relational theory of emotion'. *American Psychologist* 46.8, pp. 819–834.

— (2006). *Stress and Emotion: A New Synthesis.* Springer Publishing Company, Inc.

Lazarus, R. S. & Folkman, S. (1984). *Stress, Appraisal, and Coping.* Springer, New York, NY.

Lee, C. M.; Narayanan, S. & Pieraccini, R. (2001). 'Recognition of Negative Emotions from the Speech Signal'. In: *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding.* ASRU. Madonna di Campiglio, Italy: IEEE, pp. 240–243.

Lee, C. M. & Narayanan, S. S. (2005). 'Toward Detecting Emotions in Spoken Dialogs'. *IEEE Transactions on Speech and Audio Processing* 13 (2), pp. 293–303.

Lefter, I.; Rothkrantz, L. J. M. & Burghouts, G. J. (2012). 'Aggression Detection in Speech Using Sensor and Semantic Information'. In: *Text, Speech and Dialogue.* Ed. by Sojka, P.; Horák, A.; Kopeček, I. & Pala, K. Vol. 7499. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 665–672.

Lerch, R.; Sessler, G. & Wolf, D. (2009). *Technische Akustik - Grundlagen und Anwendungen.* Springer Dordrecht Heidelberg London New York.

Lewis, D. D. & Gale, W. A. (1994). ‘A Sequential Algorithm for Training Text Classifiers’. In: *Proc. of the 17th annual international ACM SIGIR conference on Research and development in information retrieval.* Ed. by Croft, W. B. & Rijsbergen, C. J. van. SIGIR. Springer, London, pp. 3–12.

Liaw, A. & Wiener, M. (2002). ‘Classification and regression by Random Forest’. *R news* 2.3, pp. 18–22.

Lieberman, P. & Michaels, S. B. (1962). ‘Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech’. *Journal of the Acoustical Society of America* 34 (7), pp. 922–927.

Light, R. J. (1971). ‘Measures of response agreement for qualitative data: Some generalizations and alternatives’. *Psychological Bulletin* 76 (5), pp. 365–377.

Lin, Y.-L. & Wei, G. (2005). ‘Speech emotion recognition based on HMM and SVM’. In: *Proc. of the 4th international conference on Advances in Machine Learning and Cybernetics.* ICMLC. Guangzhou, China: IEEE, pp. 4898–4901.

Liscombe, J.; Vendetti, J. & Hirschberg, J. (2003). ‘Classifying Subject Ratings of Emotional Speech Using Acoustic Features’. In: *Proc. of the EUROSPEECH 2003.*

Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P. & Tan, G.-Z. (2018). ‘Speech emotion recognition based on feature selection and extreme learning machine decision tree’. *Neurocomputing* 273, pp. 271–280.

Löcken, A.; Ihme, K. & Unni, A. (2017). ‘Towards Designing Affect-Aware Systems for Mitigating the Effects of In-Vehicle Frustration’. In: *Proc. of the 9th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications.* AutomotiveUI. Oldenburg, Germany: ACM, pp. 88–93.

Loizou, P. C. (2007). *Speech Enhancement - Theory and Practice.* CRC Press.

— (2011). ‘Speech Quality Assessment’. In: *Multimedia Analysis, Processing and Communications.* Ed. by Weisi, L.; Tao, D.; Kacprzyk, J.; Li, Z.; Izquierdo, E. & Wang, H. Vol. 346. Studies in Computational Intelligence. Springer-Verlag Berlin Heidelberg, pp. 623–654.

Lourenço, A.; Alves, A. P.; Carreiras, C.; Duarte, R. P. & Fred, A. (2015). ‘CardioWheel: ECG Biometrics on the Steering Wheel’. In: *Machine Learning* *and Knowledge Discovery in Databases.* Ed. by Bifet, A.; May, M.; Zadrozny, B.; Gavalda, R.; Pedreschi, D.; Bonchi, F.; Cardoso, J. & Spiliopoulou, M. Vol. 9286. Lecture Notes in Computer Science (LNCS). Springer, Cham, pp. 267–270.

Lu, J.; Xie, X. & Zhang, R. (2013). 'Focusing on appraisals: How and why anger and fear influence driving risk perception'. *Journal of Safety Research* 45, pp. 65–73.

Luengo, I.; Navas, E.; Hernáez, I. & Sánchez, J. (2005). 'Automatic Emotion Recognition Using Prosodic Parameters'. In: *Proc. of the 9th European Conference on Speech Communication and Technology.* EUROSPEECH/INTERSPEECH. Lisbon, Portugal: International Speech Communication Association (ISCA), pp. 493–496.

Ma, J.; Hu, Y. & Loizou, P. C. (2009). 'Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions'. *Journal of the Acoustical Society of America* 125 (5), pp. 3387–3405.

Maffei, A. & Angrilli, A. (2019). 'E-MOVIE - Experimental MOVies for Induction of Emotions in neuroscience: An innovative film database with normative data and sex differences'. *Research Article* 14.e0223124 (10), pp. 1–22.

Maier, A.; Haderlein, T.; Stelzle, F.; Nöth, E.; Nkenke, E.; Rosanowski, F.; Schützenberger, A. & Schuster, M. (2009). 'Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer'. *EURASIP Journal on Audio, Speech, and Music Processing* 2010.926951, pp. 1–7.

Malik, M. N.; Hafeez, T. & Hussain, F. (2020). 'Emotion Perception in Artificially Created Noisy Environment Utilizing Nonsense Speech'. In: *Proc. of the 1st International Conference on Communication, Electrical and Computer Networks.* ICCECN. Kuala Lumpur, Malaysia: IEEE, pp. 1–5.

Malta, L.; Miyajima, C.; Kitaoka, N. & Takeda, K. (2011). 'Analysis of Real-World Driver's Frustration'. *IEEE Transactions on Intelligent Transportation Systems* 12 (1), pp. 109–118.

Marsland, S. (2015). *Machine Learning - An Algorithmic Prespective.* Ed. by Herbrich, R. & Graepel, T. 2nd edition. Machine Learning & Pattern Recognition Series. Boca Raton: CRC Press.

Martin, O.; Kotsia, I.; Macq, B. & Pitas, I. (2006). 'The eNTERFACE' 05 Audio-Visual Emotion Database'. In: *Proc. of the 22nd International Conference on Data Engineering Workshops.* ICDEW. Atlanta, GA, USA: IEEE, pp. 1–8.

McDougall, W. (1908). *An Introduction to Social Psychology.* Psychology Press.

McGraw, K. O. & Wong, S. P. (1996). 'Forming inferences about some intraclass correlation coefficients'. *Psychological Methods* 1 (1), pp. 30–46.

McKeown, G.; Valstar, M. F.; Cowie, R. & Pantic, M. (2010). 'The SEMAINE corpus of emotionally coloured character interactions'. In: *Proc. of the 2010 IEEE International Conference on Multimedia and Expo.* ICME. IEEE, pp. 1079–1084.

Mehler, B.; Kidd, D.; Reimer, B.; Reagan, I.; Dobres, J. & McCartt, A. (2015). 'Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems'. *Ergonomics* 59 (3), pp. 344–367.

Mehrabian, A. (1996). 'Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament'. *Current Psychology* 14 (4), pp. 261–292.

Mehrabian, A. & Russell, J. A. (1974). *An approach to environmental psychology.* MIT Press.

Melhart, D.; Liapis, A. & Yannakakis, G. N. (2019). 'PAGAN: Video Affect Annotation Made Easy'. In: *Proc. of the 8th International Conference on Affective Computing and Intelligent Interaction.* ACII. Cambridge, United Kingdom: IEEE, pp. 130–136.

Mencattini, A.; Martinelli, E.; Costantini, G.; Todisco, M.; Basile, B.; Bozzali, M. & Di Natale, C. (2014). 'Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure'. *Knowledge-Based Systems* 63, pp. 68–81.

Microsoft (31st May 2018). *About the Windows Media Codecs.* Microsoft. URL: `https://docs.microsoft.com/en-us/windows/win32/medfound/about-the-windows-media-codecs#windows-media-audio-9` (visited on 07/02/2020).

Morris, J. D. (1995). 'Observations: SAM: The self-assessment manikin: An efficient cross-cultural measurement of emotional response'. *Journal of Advertising Research* 35 (6), pp. 63–68.

Murray, I. R. & Arnott, J. L. (1993). 'Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion'. *Journal of the Acoustical Society of America* 93 (2), pp. 1097–1108.

Nasoz, F.; Alvarez, K.; Lisetti, C. L. & Finkelstein, N. (2004). 'Emotion recognition from physiological signals using wireless sensors for presence technologies'. *Cognition, Technology & Work* 6 (1), pp. 4–14.

NCAP, E. (2017). *Euro NCAP 2025 Roadmap - In Pursuit of Vision Zero*. Technical Paper. Leuven, Belgium.

Niewiadomski, R.; Mancini, M.; Varni, G.; Volpe, G. & Camurri, A. (2016). 'Automated Laughter Detection From Full-Body Movements'. *IEEE Transactions on Human-Machine Systems* 46 (1), pp. 113–123.

Nordholm, S. E.; Dam, H. H.; Lai, C. C. & Lehmann, E. A. (2014). 'Broadband Beamforming and Optimization'. In: *Array and Statistical Signal Processing*. Ed. by Zoubir, A. M.; Viberg, M.; Chellappa, R. & Theodoridis, S. Vol. 3. Academic Press Library in Signal Processing. Elsevier. Chap. 13, pp. 553–598.

Ntalampiras, S. & Fakotakis, N. (2011). 'Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition'. *IEEE Transactions on Affective Computing* 3 (1), pp. 116–125.

Oates, C.; Triantafyllopoulos, A.; Steiner, I. & Schuller, B. (2019). 'Robust Speech Emotion Recognition under Different Encoding Conditions'. In: *Proc. of the 20th Annual Conference of the International Speech Communication Association*. INTERSPEECH. Graz, Austria: International Speech Communication Association (ISCA), pp. 3935–3939.

Oehl, M.; Ihme, K.; Pape, A.-A.; Vukelić, M. & Braun, M. (2020). 'Affective Use Cases for Empathic Vehicles in Highly Automated Driving: Results of an Expert Workshop'. In: *HCI in Mobility, Transport, and Automotive Systems - Automated Driving and In-Vehicle Experience Design*. Ed. by Krömker, H. Vol. 12212. Lecture Notes in Computer Science (LNCS). Springer, Cham, pp. 89–100.

Olson, D. L. & Delen, D. (2008). *Advanced Data Mining Techniques*. Springer-Verlag Berlin Heidelberg.

Ortony, A. & Turner, T. J. (1990). 'What's Basic About Basic Emotions?' *Psychological Review* 97 (3), pp. 315–331.

Oshiro, T. M.; Perez, P. S. & Baranauskas, J. A. (2012). 'How Many Trees in a Random Forest?' In: *Machine Learning and Data Mining in Pattern Recognition*. Ed. by Perner, P. Vol. 7376. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 154–168.

Özseven, T. (2019). 'A novel feature selection method for speech emotion recognition'. *Applied Acoustics* 146, pp. 320–326.

Padmanabhan, J. & Premkumar, M. J. J. (2015). 'Machine Learning in Automatic Speech Recognition: A Survey'. *IETE Technical Review* 32 (4), pp. 240–251.

Pao, T.-L.; Liao, W.-Y.; Chen, Y.-T.; Yeh, J.-H.; Cheng, Y.-M. & Chien, C. S. (2007). 'Comparison of Several Classifiers for Emotion Recognition from Noisy Mandarin Speech'. In: *Proc. of the 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing.* IIH-MSP. Kaohsiung, Taiwan: IEEE, pp. 23–26.

Parada-Cabaleiro, E.; Baird, A.; Batliner, A.; Cummins, N.; Hantke, S. & Schuller, B. (2017). 'The Perception of Emotions in Noisified Nonsense Speech'. In: *Proc. of the 18th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA), pp. 3246–3250.

Pasupathi, M. (2003). 'Emotion regulation during social remembering: Differences between emotions elicited during an event and emotions elicited when talking about it'. *Memory* 11 (2), pp. 151–163.

Pearce, D. & Hirsch, H.-G. (2000). 'The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions'. In: *Proc. of the 6th International Conference on Spoken Language Processing.* ICSLP. Beijing, China: International Speech Communication Association (ISCA), pp. 29–32.

Pêcher, C.; Lemercier, C. & Cellier, J.-M. (2009). 'Emotions drive attention: Effects on driver's behaviour'. *Safety Science* 47 (9), pp. 1254–1259.

— (2010). 'The Influence of Emotions on Driving Behavior'. In: *Traffic Psychology: An International Perspective.* Ed. by Hennessy, D. Psychology Research Progress. Nova Science Publishers. Chap. 9, pp. 145–158.

Pell, M. D. & Kotz, S. A. (2011). 'On the Time Course of Vocal Emotion Recognition'. *Research Article* 6.e27256 (11), pp. 1–16.

Pérez-Espinosa, H.; Reyes-García, C. A. & Villaseñor-Pineda, L. (2012). 'Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model'. *Biomedical Signal Processing and Control* 7 (1), pp. 79–87.

Picard, R. R. & Cook, R. D. (1984). 'Cross-Validation of Regression Models'. *Journal of the American Statistical Association* 79.387, pp. 575–583.

Picard, R. W. (1997). *Affective Computing.* MIT Press.

Picard, R. W. & Klein, J. (2002). 'Computers that recognise and respond to user emotion: theoretical and practical implications'. *Interacting with Computers* 14 (2), pp. 141–169.

Planet, S. & Iriondo, I. (2012). 'Comparative Study on Feature Selection and Fusion Schemes for Emotion Recognition from Speech'. *International Journal of Interactive Multimedia and Artificial Intelligence* 1.6 (Special Issue on Intelligent Systems and Applications), pp. 44–51.

Plutchik, R. (1958). 'SECTION OF PSYCHOLOGY: OUTLINES OF A NEW THEORY OF EMOTION'. *Transactions of the New York Academy of Sciences.* 2nd ser. 20 (5), pp. 394–403.

— (1980). 'A General Psychoevolutionary Theory of Emotion'. In: *Theories of Emotion.* Ed. by Plutchik, R. & Kellerman, H. 1st ed. Emotion: Theory, Research, and Experience. Academic Press. Chap. 1, pp. 3–33.

— (2001). 'The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice'. *American Scientist* 89, pp. 344–350.

Pohjalainen, J.; Ringeval, F.; Zhang, Z. & Schuller, B. (2016). 'Spectral and Cepstral Audio Noise Reduction Techniques in Speech Emotion Recognition'. In: *Proc. of the 24th ACM international conference on Multimedia.* MM. Amsterdam, The Netherlands: ACM, pp. 670–674.

Pollack, I.; Rubenstein, H. & Horowitz, A. (1960). 'Communication of Verbal Modes of Expression'. *Language and Speech* 3 (3), pp. 121–130.

Powers, D. M. W. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.* Technical Report SIE-07-001. Adelaide, Australia: School of Informatics and Engineering, Flinders University.

Probst, P.; Boulesteix, A.-L. & Bernd, B. (2019). 'Tunability: Importance of Hyperparameters of Machine Learning Algorithms'. *Journal of Machine Learning Research* 2019 (20), pp. 1–32.

Prodeus, A.; Didkovskyi, V.; Didkovska, M.; Kotvytskyi, I.; Motorniuk, D. & Khrapachevskyi, A. (2018). 'Objective and Subjective Assessment of the Quality and Intelligibility of Noised Speech'. In: *Proc. of the 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology.* PIC S&T. Kharkov, Ukraine: IEEE, pp. 71–74.

Purves, D.; Augustine, G. J.; Fitzpatrick, D.; Katz, L. C.; LaMantia, A.-S.; McNamara, J. O. & Williams, S. M. (2001). *Neuroscience.* 2nd edition. Sinauer Associates.

Rammstedt, B. & John, O. P. (2007). 'Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German'. *Journal of Research in Personality* 41.1, pp. 203–212.

Rasch, B.; Friese, M.; Hofmann, W. & Neumann, E. (2014). *Quantitaive Methoden 2.* 4th edition. Springer-Verlag Berlin Heidelberg.

Revelle, W. R. & Scherer, K. R. (2009). 'Personality and emotion'. In: *Oxford Companion to Emotion and the Affective Sciences.* Ed. by Sander, D. & Scherer, K. R. Oxford University Press, pp. 304–305.

Ringeval, F.; Sonderegger, A.; Sauer, J. & Lalanne, D. (2013). 'Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions'. In: *Proc. of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition.* FG. Shanghai, China China: IEEE, pp. 1–8.

Rong, J.; Li, G. & Chen, Y.-P. P. (2009). 'Acoustic feature selection for automatic emotion recognition from speech'. *Information Processing and Management* 45 (3), pp. 315–328.

Rooney, B.; Benson, C. & Hennessy, E. (2012). 'The apparent reality of movies and emotional arousal: A study using physiological and self-report measures'. *Poetics* 40 (5), pp. 405–422.

Ross, M.; Duffy, R. J.; Cooker, H. S. & Sargeant, R. L. (1973). 'Contribution of the Lower Audible Frequencies to the Recognition of Emotions'. *American Annals of the Deaf* 118 (1), pp. 37–42.

Russell, J. A. (1980). 'A circumplex model of affect'. *Journal of Personality and Social Psychology* 39 (6), pp. 1161–1178.

Russell, J. A. & Barrett, L. F. (1999). 'Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant'. *Journal of Personality and Social Psychology* 76 (5), pp. 805–819.

Russell, J. A. & Lemay, G. (2000). 'Emotion Concepts'. In: *Handbook of Emotions.* Ed. by Lewis, M. & Haviland-Jones, J. M. 2nd ed. Guilford Press, pp. 491–503.

SAE (2018). *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.* Standard J3016_201806. Warrendale, PA, USA: Society of Automotive Engineers.

Salomon, D. (2007). *Data Compression - The Complete Reference.* 4th edition. Springer-Verlag London.

Satt, A.; Rozenberg, S. & Hoory, R. (2017). 'Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms'. In: *Proc. of the 18th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Stockholm, Sweden: International Speech Communication Association (ISCA), pp. 1089–1093.

Sauert, B.; Enzner, G. & Vary, P. (2006). 'Near End Listening Enhancement with Strict Loudspeaker Output Power Constraining'. In: *Proc. of the 10th International Workshop on Acoustic Echo and Noise Control.* IWAENC. Paris, France, pp. 1–4.

Schaefer, A.; Nils, F.; Sanchez, X. & Philippot, P. (2010). 'Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers'. *Cognition and Emotion* 24 (7), pp. 1153–1172.

Schaffer, C. (1993). 'Selecting a Classification Method by Cross-Validation'. *Machine Learning* 13 (1), pp. 135–143.

Scheffer, T.; Decomain, C. & Wrobel, S. (2001). 'Active Hidden Markov Models for Information Extraction'. In: *Advances in Intelligent Data Analysis.* Ed. by Hoffmann, F.; Hand, D. J.; Adams, N.; Fisher, D. & Guimaraes, G. Vol. 2189. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 309–318.

Scheibner, P. (2012). *Ökonomische Objektivierung von elektromechanischen Funktionsgeräuschen.* Logos Verlag Berlin.

Scherer, K. R. (1986a). 'Vocal Affect Expression: A Review and a Model for Future Research'. *Psychological Bulletin* 99 (2), pp. 143–165.

— (1986b). 'Voice, Stress, and Emotion'. In: *Dynamics of Stress: Physiological, Psychological and Social Perspectives.* Ed. by Appley, M. H. & Trumbull, R. The Plenum Series on Stress and Coping. Springer, Boston, MA. Chap. 9, pp. 157–179.

— (1987). 'Toward a dynamic theory of emotion: The component process model of affective states'. *Geneva Studies in Emotion and Communication* 1.

Scherer, K. R. (2005a). ‘Unconscious Processes in Emotion: The Bulk of the Iceberg’. In: *Emotion and Consciousness.* Ed. by Barrett, L. F.; Niedenthal, P. M. & Winkielman, P. The Guilford Press. Chap. 13, pp. 312–334.

— (2005b). ‘What are emotions? And how can they be measured?’ *Social Science Information* 44 (4), pp. 695–729.

— (2009). ‘The dynamic architecture of emotion: Evidence for the component process model’. *Cognition and Emotion* 23 (7), pp. 1307–1351.

Scherer, K. R.; Johnstone, T. & Klasmeyer, G. (2003). ‘Vocal expression of emotion’. In: *Handbook of affective sciences.* Ed. by Davidson, R. J.; Scherer, K. R. & Goldsmith, H. H. Series in affective science. Oxford University Press, pp. 433–456.

Scherer, K. R.; Shuman, V.; Fontaine, J. J. R. & Soriano, C. (2013). In: *Components of emotional meaning: A sourcebook.* Ed. by Fontaine, J. J. R.; Scherer, K. R. & Soriano, C. Series in affective science. Oxford University Press. Chap. 18, pp. 281–298.

Schiel, F.; Steininger, S. & Türk, U. (2002). ‘The SmartKom Multimodal Corpus at BAS’. In: *Proc. of the 3rd international conference on Language Resources and Evaluation.* LREC. Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA), pp. 200–206.

Schimmack, U. (1997). ‘Das Berliner-Alltagssprachliche-Stimmungs-Inventar (BASTI): Ein Vorschlag zur kontentvaliden Erfassung von Stimmungen [The Berlin Everyday Language Mood Inventory (BELMI): Toward the content valid assessment of moods]’. *Diagnostica* 43 (2), pp. 150–173.

Schirmer, A. & Adolphs, R. (2017). ‘Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence’. *Trends in Cognitive Science* 21 (3), pp. 216–228.

Schmidt, K.; Patnaik, P. & Kensinger, E. A. (2011). ‘Emotion’s Influence on Memory for Spatial and Temporal Context’. *Cognition and Emotion* 25 (2), pp. 229–243.

Schmidt-Daffy, M. (2013). ‘Fear and anxiety while driving: Differential impact of task demands, speed and motivation’. *Transportation Research Part F: Traffic Psychology and Behaviour* 16, pp. 14–28.

Schölkopf, B. & Smola, A. J. (2003). *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond.* Ed. by Dietterich, T. Adaptive Computation and Machine Learning. MIT Press.

Schuller, B. (2008). ‘Speaker, Noise, and Acoustic Space Adaptation for Emotion Recognition in the Automotive Environment’. In: *Proc. of the ITG Conference on Voice Communication [8. ITG-Fachtagung].* Aachen, Germany: VDE Verlag GmbH, pp. 1–4.

— (2011). ‘Recognizing Affect from Linguistic Information in 3D Continuous Space’. *IEEE Transactions on Affective Computing* 2 (4), pp. 192–205.

— (2018). ‘Speech Emotion Recognition - Two Decades in a Nutshell, Benchmarks, and Ongoing Trends’. *Communications of the ACM* 61.5, pp. 90–99.

Schuller, B.; Arsic, D.; Wallhoff, F. & Rigoll, G. (2006). ‘Emotion Recognition in the Noise Applying Large Acoustic Feature Sets’. In: *Proc. of the 3rd International Conference on Speech Prosody.* SP. Dresden, Germany: International Speech Communication Association (ISCA), pp. 1–4.

Schuller, B.; Lang, M. & Rigoll, G. (2005). ‘Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers’. In: *Fortschritte der Akustik - DAGA’05.* Ed. by Fastl, H. & Fruhmann, M. Vol. 31. Dautesche Jahrestagung für Akustik. Deutsche Gesellschaft für Akustik e.V. (DEGA), pp. 329–330.

Schuller, B.; Müller, R.; Lang, M. & Rigoll, G. (2005). ‘Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features Within Ensembles’. In: *Proc. of the 9th European Conference on Speech Communication and Technology.* EUROSPEECH/INTERSPEECH. Lisbon, Portugal: International Speech Communication Association (ISCA), pp. 805–808.

Schuller, B.; Reiter, S. & Rigoll, G. (2006). ‘Evolutionary Feature Generation in Speech Emotion Recognition’. In: *Proc. of the IEEE International Conference on Multimedia and Expo.* ICME. Toronto, Ontario, Canada: IEEE, pp. 5–8.

Schuller, B.; Rigoll, G.; Grimm, M.; Kroschel, K.; Moosmayr, T. & Ruske, G. (2007). ‘Effects of In-Car Noise-Conditions on the Recognition of Emotion within Speech’. In: *Fortschritte der Akustik - DAGA 2007.* Ed. by Mehra, S.-R. & Leistner, P. Vol. 33. Dautesche Jahrestagung für Akustik. Deutsche Gesellschaft für Akustik e.V. (DEGA), pp. 305–306.

Schuller, B.; Rigoll, G. & Lang, M. (2003). ‘Hidden Markov model-based speech emotion recognition’. In: *Proc. of the 2003 IEEE International Conference on Acoustics, Speech and Signal Processing.* ICASSP. Hong Kong: IEEE, pp. 1–4.

Schuller, B.; Seppi, D.; Batliner, A.; Maier, A. & Steidl, S. (2007). ‘Towards More Reality in the Recognition of Emotional Speech’. In: *Proc. of the 2007 IEEE*

International Conference on Acoustics, Speech and Signal Processing. ICASSP. Honolulu, Hawaii, USA: IEEE, pp. IV-941–IV-944.

Schuller, B.; Steidl, S. & Batliner, A. (2009). 'The INTERSPEECH 2009 Emotion Challenge'. In: *Proc. of the 10th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Brighton, United Kingdom: International Speech Communication Association (ISCA), pp. 312–315.

Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Müller, C. & Narayanan, S. (2010). 'The INTERSPEECH 2010 Paralinguistic Challenge'. In: *Proc. of the 11th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Makuhari, Chiba, Japan: International Speech Communication Association (ISCA), pp. 2794–2797.

Schuller, B.; Vlasenko, B.; Arsic, D.; Rigoll, G. & Wendemuth, A. (2008). 'Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition'. In: *Proc. of the 2008 IEEE International Conference on Multimedia and Expo.* ICME. Hannover, Germany: IEEE, pp. 1333–1336.

Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G. & Wendemuth, A. (2009). 'Acoustic Emotion Recognition: A Benchmark Comparison of Performances'. In: *Proc. of the IEEE Workshop on Automatic Speech Recognition & Understanding.* ASRU. Merano, Italy: IEEE, pp. 552–557.

Schuller, B. et al. (2013). 'The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism'. In: *Proc. of the 14th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Lyon, France: International Speech Communication Association (ISCA), pp. 148–152.

Schuster, M.; Maier, A.; Haderlein, T.; Nkenke, E.; Wohlleben, U.; Rosanowski, F.; Eysholdt, U. & Nöth, E. (2006). 'Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition'. *International Journal of Pediatric Otorhinolaryngology* 70 (10), pp. 1741–1747.

Schwan, B. (15th Nov. 2011). *Details zum Siri-Protokoll.* URL: `https://www.heise.de/mac-and-i/meldung/Details-zum-Siri-Protokoll-1379044.html` (visited on 12/09/2020).

Scott, W. A. (1955). 'Reliability of Content Analysis: The Case of Nominal Scale Coding'. *Public Opinion Quarterly* 19 (3), pp. 321–325.

Sedaaghi, M. H.; Kotropoulos, C. & Ververidis, D. (2007). 'Using Adaptive Genetic Algorithms to Improve Speech Emotion Recognition'. In: *Proc. of the 9th Workshop on Multimedia Signal Processing.* MMSP. Crete, Greece: IEEE, pp. 461–464.

Sedaaghi, M. H.; Ververidis, D. & Kotropoulos, C. (2007). 'Improving speech emotion recognition using adaptive genetic algorithms'. In: *Proc. of the 15th European Signal Processing Conference.* EUSIPCO. Poznań, Poland: IEEE, pp. 2209–2213.

Settles, B. (2010). *Active Learning Literature Survey.* Technical Report 1648. Wisconsin, Madison, USA: Computer Sciences, University of Wisconsin-Madison.

Shafaei, S.; Hacizade, T. & Knoll, A. (2019). 'Integration of Driver Behavior into Emotion Recognition Systems: A Preliminary Study on Steering Wheel and Vehicle Acceleration'. In: *Computer Vision – ACCV 2018 Workshops.* Ed. by Carneiro, G. & You, S. Vol. 11367. Lecture Notes in Computer Science (LNCS). Springer, Cham, pp. 386–401.

Sharma, D.; Hilkhuysen, G.; Gaubitch, N. D.; Brookes, M. & Naylor, P. (2010). 'C-Qual—A Validation of PESQ Using Degradations Encountered in Forensic and Law Enforcement Audio'. In: *Proc. of the 39th International Conference: Audio Forensics: Practices and Challenges.* Hillerød, Denmark: Audio Engineering Society (AES), pp. 177–181.

Sheikhan, M.; Bejani, M. & Gharavian, D. (2013). 'Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method'. *Neural Computing and Applications* 23 (1), pp. 215–227.

Shen, Y.; Wu, L.; Li, Y.; Liu, S. & Wen, Q. (2016). 'Big Data Overview'. In: *Big Data: Concepts, Methodologies, Tools, and Applications.* Ed. by Khosrow-Pour, M.; Clarke, S.; Jennex, M. E.; Becker, A. & Anttiroiko, A.-V. Vol. 1. IGI Global. Chap. 1, pp. 1–29.

Shinar, D. (1998). 'Aggressive driving: the contribution of the drivers and the situation'. *Transportation Research Part F: Traffic Psychology and Behaviour* 1 (2), pp. 137–160.

Shrout, P. E. & Fleiss, J. L. (1979). 'Intraclass correlations: Uses in assessing rater reliability'. *Psychological Bulletin* 86 (2), pp. 420–428.

Sidorov, M.; Brester, C.; Minker, W. & Semenkin, E. (2014). 'Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm'. In: *Proc. of the 9th International Conference on Language Resources* *and Evaluation.* LREC. Reykjavik, Iceland: European Language Resource Association (ELRA), pp. 3481–3485.

Sidorov, M.; Brester, C. & Schmitt, A. (2015). ‘Contemporary Stochastic Feature Selection Algorithms for Speech-Based Emotion Recognition’. In: *Proc. of the 16th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Dresden, Germany: International Speech Communication Association (ISCA), pp. 2699–2703.

Siegert, I. (2015). ‘Emotional and User-Specific Acoustic Cues for Improved Analysis of Naturalistic Interactions’. PhD thesis. Magdeburg: Otto-von-Guericke Universität Magdeburg.

Siegert, I.; Böck, R.; Vlasenko, B. & Wendemuth, A. (2015). ‘Exploring Dataset Similarities using PCA-based Feature Selection’. In: *Proc. of the 2015 International Conference on Affective Computing and Intelligent Interaction.* ACII. Xi’an, China: IEEE, pp. 387–393.

Siegert, I.; Böck, R. & Wendemuth, A. (2014). ‘Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements’. *Journal on Multimodal User Interfaces* 8, pp. 17–28.

Siegert, I. & Wendemuth, A. (2017). ‘ikannotate2 – A Tool Supporting Annotation of Emotions in Audio-Visual Data’. In: *Elektronische Sprachsignalverarbeitung 2016. Tagungsband der 28. Konferenz.* Ed. by Trouvain, J.; Steiner, I. & Möbius. Vol. 86. Studientexte zur Sprachkommunikation. Saarbrücken, Germany: TUDpress, pp. 17–24.

Sinder, D. J.; Varga, I.; Krishnan, V.; Rajendran, V. & Villette, S. (2015). ‘Recent Speech Coding Technologies and Standards’. In: *Speech and Audio Processing for Coding, Enhancement and Recognition.* Ed. by Ogunfunmi, T.; Togneri, R. & Narasimha, M. S. Springer-Verlag New York. Chap. 4, pp. 75–109.

Slavik, K. M. (2008). ‘Anschlusstechnik, Interfaces, Vernetzung’. In: *Handbuch der Audiotechnik.* Ed. by Weinzierl, S. VDI-Buch. Springer-Verlag Berlin Heidelberg. Chap. 18, pp. 945–1033.

Sogon, S. (1975). ‘A study of the personality factor which affects the judgment of vocally expressed emotions’. *Japanese Journal of Psychology* 46 (5), pp. 247–254.

Solera-Ureña, R.; Padrell-Sendra, J.; Martín-Iglesias, D.; Gallardo-Antolín, A.; Peláez-Moreno, C. & Díaz-de-María, F. (2007). ‘SVMs for Automatic Speech Recognition: A Survey’. In: *Progress in Nonlinear Speech Processing.* Ed. by

Stylianou, Y.; Faundez-Zanuy, M. & Esposito, A. Vol. 4391. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 190–216.

El-Solh, A.; Cuhadar, A. & Goubran, R. (2007). 'Evaluation of Speech Enhancement Techniques for Speaker Identification in Noisy Environments'. In: *Proc. of the 9th IEEE International Symposium on Multimedia Workshops.* ISMW. Taichung, Taiwan: IEEE, pp. 235–239.

Solomon, R. C. (2000). 'The Philosophy of Emotions'. In: *Handbook of Emotions.* Ed. by Lewis, M. & Haviland-Jones, J. M. 2nd ed. The Guilford Press, New York, London. Chap. 1, pp. 3–15.

Son, R. van (2005). 'A Study of Pitch, Formant, and Spectral Estimation Errors Introduced by Three Lossy Speech Compression Algorithms'. *Acta Acustica united with Acustica* 91.4, pp. 771–778.

Song, P.; Jin, Y.; Zha, C. & Zhao, L. (2015). 'Speech emotion recognition method based on hidden factor analysis'. *IET Electronics Letters* 51 (1), pp. 112–114.

Spearman, C. (1904). 'The Proof and Measurement of Association Between Two Things'. *American Journal of Psychology* 15, pp. 88–103.

Spruyt, V. (2014). *The Curse of Dimensionality in classification.* URL: https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/ (visited on 07/03/2021).

Stańczyk, U.; Zielosko, B. & Jain, L. C. (eds.). *Advances in Feature Selection for Data and Pattern Recognition.* Vol. 138. Intelligent Systems Reference Library. Springer, Cham.

Staroniewicz, P. & Majewski, W. (2009). 'Polish Emotional Speech Database – Recording and Preliminary Validation'. In: *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions.* Ed. by Esposito, A. & Vích, R. Vol. 5641. Lecture Notes on Artificial Intelligence (LNAI). Springer, Berlin, Heidelberg, pp. 42–49.

Statista, R. D. (2021a). *Market size of ADAS systems worldwide 2015-2023.* URL: https://www.statista.com/statistics/591579/adas-and-ad-systems-in-light-vehicles-global-market-size/ (visited on 19/06/2021).

— (2021b). *Size of the global autonomous car market 2019-2023.* URL: https://www.statista.com/statistics/428692/projected-size-of-global-autonomous-vehicle-market-by-vehicle-type/ (visited on 19/06/2021).

Stearns, P. N. (2000). 'History of Emotions: Issues of Changes and Impact'. In: *Handbook of Emotions.* Ed. by Lewis, M. & Haviland-Jones, J. M. 2nd ed. The Guilford Press, New York, London. Chap. 2, pp. 16–29.

Steeneken, H. J. M. & Houtgast, T. (1980). 'A physical method for measuring speech-transmission quality'. *Journal of the Acoustical Society of America* 67 (1), pp. 318–326.

Steinhauser, K.; Leist, F.; Maier, K.; Michel, V.; Pärsch, N.; Rigley, P.; Wurm, F. & Steinhauser, M. (2011). 'Effects of emotions on driving behavior'. *Transportation Research Part F: Traffic Psychology and Behaviour* 59 (Part A), pp. 150–163.

Steinwart, I. & Christmann, A. (2008). *Support Vector Machines.* Ed. by Jordan, M.; Kleinberg, J. & Schölkopf, B. Information Science and Statistics. Springer Science+Business Media.

Sun, L.; Fu, S. & Wang, F. (2019). 'Decision tree SVM model with Fisher feature selection for speech emotion recognition'. *EURASIP Journal on Audio, Speech, and Music Processing* 2019.2, pp. 1–14.

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

Taal, C. H.; Hendriks, R. C.; Heusdens, R.; Jensen, J. & Kjems, U. (2009). 'An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction'. In: *Proc. of the 10th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Brighton, United Kingdom: International Speech Communication Association (ISCA), pp. 1947–1950.

Taubman-Ben-Ari, O. (2012). 'The effects of positive emotion priming on self-reported reckless driving'. *Accident Analysis & Prevention* 45, pp. 718–725.

Tawari, A. & Trivedi, M. (2010a). 'Speech Based Emotion Classification Framework for Driver Assistance System'. In: *Proc. of the 2010 IEEE Intelligent Vehicles Symposium.* IV. San Diego, CA, USA: IEEE, pp. 174–178.

— (2010b). 'Speech Emotion Analysis in Noisy Real-World Environment'. In: *Proc. of the 20th International Conference on Pattern Recognition.* ICPR. Istanbul, Turkey: IEEE, pp. 4605–4608.

Teager, H. M. & Teager, S. M. (1990). 'Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract'. In: *Speech Production and Speech Modelling.*

Ed. by Hardcastle, W. J. & Marchal, A. Vol. 55. NATO Science Series. Springer, Dordrecht, pp. 241–261.

Theodoridis, S. & Koutroumbas, K. (2009). *Pattern Recognition.* 4th edition. Academic Press.

Tognetti, S.; Garbarino, M.; Bonanno, A. T.; Matteucci, M. & Bonarini, A. (2010). 'Enjoyment Recognition From Physiological Data in a Car Racing Game'. In: *Proc. of the 3rd international workshop on Affective interaction in natural environments.* AFFINE. Firenze, Italy: ACM, pp. 3–8.

Tomkins, S. S. (1962). *Affect, imagery, consciousness, Vol. 1: The positive affects.* American Psychological Association.

Triantafyllopoulos, A.; Keren, G.; Wagner, J.; Steiner, I. & Schuller, B. (2019). 'Towards Robust Speech Emotion Recognition Using Deep Residual Networks for Speech Enhancement'. In: *Proc. of the 20th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Graz, Austria: International Speech Communication Association (ISCA), pp. 1691–1695.

Truong, K. P.; Leeuwen, D. A. van & Jong, F. M. de (2012). 'Speech-based recognition of self-reported and observed emotion in a dimensional space'. *Speech Communication* 54 (9), pp. 1049–1063.

Valin, J.-M. (2006). 'Speex: A Free Codec For Free Speech'. In: *Proc. of the 7th linux.conf.au (LCA'06).* Dunedin, New Zealand.

— (2007). *The Speex Codec Manual.* Version Version 1.2 Beta 3.

Valin, J.-M.; Vos, K. & Terriberry, T. (2012). *Definition of the Opus Audio Codec.* RFC 6716. Internet Engineering Task Force (IETF).

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory.* Ed. by Jordan, M.; Lauritzen, S. L.; Lawless, J. F. & Nair, V. 2nd edition. Information Science and Statistics. Springer-Verlag New York.

Verma, B. & Choudhary, A. (2018a). 'A Framework for Driver Emotion Recognition using Deep Learning and Grassmann Manifolds'. In: *Proc. of the 21st International Conference on Intelligent Transportation Systems.* ITSC. Maui, Hawaii, USA: IEEE, pp. 1421–1426.

— (2018b). 'Deep Learning Based Real-Time Driver Emotion Monitoring'. In: *Proc. of the 2018 IEEE International Conference on Vehicular Electronics and Safety.* ICVES. Madrid, Spain: IEEE, pp. 1–6.

Ververidis, D.; Kotropoulos, C. & Pitas, I. (2004). ‘Automatic emotional speech classification’. In: *Proc. of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing.* ICASSP. Montreal, Quebec, Canada: IEEE, pp. 593–596.

Ververidis, D.; Kotsia, I.; Kotropoulos, C. & Pitas, I. (2008). ‘Multi-modal emotion-related data collection within a virtual earthquake emulator’. In: *Proc. of the 2nd Workshop on Corpora for Research on Emotion and Affect.* LREC. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 57–60.

Vetter, P.; Leicht, L.; Leonhardt, S. & Teichmann, D. (2017). ‘Integration of an electromagnetic coupled sensor into a driver seat for vital sign monitoring: Initial insight’. In: *Proc. of the 2017 IEEE International Conference on Vehicular Electronics and Safety.* ICVES. Vienna, Austria: IEEE, pp. 185–190.

Vogt, T. & André, E. (2005). ‘Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition’. In: *Proc. of the IEEE International Conference on Multimedia and Expo.* ICME. Beijing, China: IEEE, pp. 474–477.

Volkening, N.; Unni, A.; Rieger, J. W.; Fudickar, S. & Hein, A. (2018). ‘Development of a Mobile Functional Near-Infrared Spectroscopy Prototype’. In: *Internet of Vehicles. Technologies and Services Towards Smart City.* Ed. by Skulimowski, A. M.; Sheng, Z.; Khemiri-Kallel, S.; Cérin, C. & Hsu, C.-H. Vol. 11253. Lecture Notes in Computer Science (LNCS). Springer, Cham, pp. 146–161.

Volkswagen AG (2019). *Microsoft and Volkswagen jointly develop the Automotive Cloud.* URL: https://www.volkswagenag.com/en/news/stories/2019/02/microsoft-and-volkswagen-jointly-develop-software.html (visited on 13/05/2021).

Vrana, S. R. (1993). ‘The psychophysiology of disgust: Differentiating negative emotional contexts with facial EMG’. *Psychophysiology* 30 (3), pp. 279–286.

Vuoskoski, J. K. & Eerola, T. (2011). ‘Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences’. *Musicae Scientiae* 15 (2), pp. 159–173.

Walter, S.; Scherer, S.; Schels, M.; Glodek, M.; Hrabal, D.; Schmidt, M.; Böck, R.; Limbrecht, K.; Traue, H. C. & Schwenker, F. (2011). ‘Multimodal Emotion Classification in Naturalistic User Behavior’. In: *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments.* Ed. by Jacko,

J. A. Vol. 6763. Lecture Notes in Computer Science (LNCS). Springer, Berlin, Heidelberg, pp. 603–611.

Wang, Y. & Guan, L. (2004). 'An investigation of speech-based human emotion recognition'. In: *Proc. of the 6th Workshop on Multimedia Signal Processing.* MMSP. Siena, Italy: IEEE, pp. 15–18.

— (2008). 'Recognizing Human Emotional State From Audiovisual Signals'. *IEEE Transactions on Multimedia* 10 (5), pp. 936–946.

Wang, Y.; Guan, L. & Venetsanopoulos, A. N. (2012). 'Kernel Cross-Modal Factor Analysis for Information Fusion With Application to Bimodal Emotion Recognition'. *IEEE Transactions on Multimedia* 14 (3), pp. 597–607.

Watson, D.; Clark, L. A. & Tellegen, A. (1988). 'Development and validation of brief measures of positive and negative affect: The PANAS scales'. *Journal of Personality and Social Psychology* 54 (6), pp. 1063–1070.

Wendemuth, A. (2004). *Grundlagen der stochastischen Sprachverarbeitung.* Oldenbourg Wissenschaftsverlag.

Weninger, F.; Schuller, B.; Batliner, A.; Steidl, S. & Seppi, D. (2011). 'Recognition of Nonprototypical Emotions in Reverberated and Noisy Speech by Nonnegative Matrix Factorization'. *EURASIP Journal on Audio, Speech, and Music Processing* 2011.838790, pp. 1–16.

Whaiduzzaman, M.; Sookhak, M.; Gani, A. & Buyya, R. (2014). 'A survey on vehicular cloud computing'. *Journal of Network and Computer Applications* 40, pp. 325–344.

Widrow, B.; Glover, J. R.; McCool, J. M.; Kaunitz, J.; Williams, C. S.; Hearn, R. H.; Zeidler, J. R.; Dong, E. J. & Goodlin, R. C. (1975). 'Adaptive Noise Cancelling: Principles and Applications'. *Proc. of the IEEE* 63 (12), pp. 1692–1716.

Wilhelm, F. H. & Grossman, P. (2010). 'Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment'. *Biological Psychology* 84 (3), pp. 552–569.

Wu, S.; Falk, T. H. & Chan, W.-Y. (2011). 'Automatic speech emotion recognition using modulation spectral features'. *Speech Communication* 53 (5), pp. 768–785.

Xiaoqing, J.; Kewen, X.; Yongliang, L. & Jianchuan, B. (2017). ‘Noisy speech emotion recognition using sample reconstruction and multiple-kernel learning’. *The Journal of China Universities of Posts and Telecommunications* 24 (2), pp. 1–9.

Xu, X.; Deng, J.; Zheng, W.; Zhao, L. & Schuller, B. (2015). ‘Dimensionality Reduction for Speech Emotion Features by Multiscale Kernels’. In: *Proc. of the 16th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Dresden, Germany: International Speech Communication Association (ISCA), pp. 1532–1536.

Xu, Y.; Hübener, I.; Seipp, A.-K.; Ohly, S. & David, K. (2017). ‘From the lab to the real-world: An investigation on the influence of human movement on Emotion Recognition using physiological signals’. In: *Proc. of the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops.* PerCom Workshops. Kona, Big Island, Hawaii, USA: IEEE, pp. 345–350.

Yang, N.; Yuan, J.; Zhou, Y.; Demirkol, I.; Heinzelman, W. & Sturge-Apple, M. L. (2014). ‘How Does Noise Impact Speech-based Emotion Classification?’ In: *Proc. of the Designing Speech and Language Interactions Workshop of the 2014 Conference on Human Factors in Computing Systems.* CHI. Toronto, Canada: ACM, pp. 1–4.

Yeh, L.-Y. & Chi, T.-S. (2010). ‘Spectro-Temporal Modulations for Robust Speech Emotion Recognition’. In: *Proc. of the 11th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Makuhari, Chiba, Japan: International Speech Communication Association (ISCA), pp. 789–792.

You, M.; Chen, C.; Bu, J.; Liu, J. & Tao, J. (2006a). ‘A Hierarchical Framework for Speech Emotion Recognition’. In: *Proc. of the IEEE International Symposium on Industrial Electronics.* ISIE. Montreal, Quebec, Canada: IEEE, pp. 515–519.

— (2006b). ‘Emotion Recognition from Noisy Speech’. In: *Proc. of the IEEE International Conference on Multimedia and Expo.* ICME. Toronto, Ontario, Canada: IEEE, pp. 1653–1656.

Yu, D. & Deng, L. (2015). *Automatic Speech Recognition- A Deep Learning Approach.* Signals and Communication Technology. Springer-Verlag London.

Zajonc, R. B.; Murphy, S. T. & Inglehart, M. (1989). ‘Feeling and facial efference: Implications of the vascular theory of emotion’. *Psychological Review* 96 (3), pp. 395–416.

Zander, T. O.; Lehne, M.; Ihme, K.; Jatzev, S.; Correia, J.; Kothe, C.; Picht, B. & Nijboer, F. (2011). 'A dry EEG-system for scientific research and brain–computer interfaces'. *Frontiers in Neuroscience* 5.53, pp. 1–10.

Zepf, S.; Stracke, T.; Schmitt, A.; van de Camp, F. & Beyerer, J. (2019). 'Towards Real-Time Detection and Mitigation of Driver Frustration using SVM'. In: *Proc. of the 18th IEEE International Conference On Machine Learning And Applications.* ICMLA. Boca Raton, Florida, USA: IEEE, pp. 202–209.

Zhang, M.; Ihme, K. & Drewitz, U. (2018). 'Discriminating Drivers' Fear and Frustration through the Dimension of Power'. In: *Proc. of the 6th Humanist Conference.* The Hague, Netherlands.

Zhang, Z.; Coutinho, E.; Deng, J. & Schuller, B. (2015). 'Cooperative Learning and its Application to Emotion Recognition from Speech'. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23 (1), pp. 115–126.

Zhang, Z.; Han, J.; Deng, J.; Xu, X.; Ringeval, F. & Schuller, B. (2018). 'Leveraging Unlabelled Data for Emotion Recognition with Enhanced Collaborative Semi-Supervised Learning'. *IEEE Access* 6, pp. 22196–22209.

Zhang, Z. & Schuller, B. (2012). 'Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition'. In: *Proc. of the 13th Annual Conference of the International Speech Communication Association.* INTERSPEECH. Portland, OR, USA: International Speech Communication Association (ISCA), pp. 362–365.

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

Zhou, G.; Hansen, J. H. L. & Kaiser, J. F. (2001). 'Nonlinear Feature Based Classification of Speech Under Stress'. *IEEE Transactions on Speech and Audio Processing* 9 (3), pp. 201–216.

Zhu, X. (2008). *Semi-Supervised Learning Literature Survey.* Technical Report 1530. Wisconsin, Madison, USA: Computer Sciences, University of Wisconsin-Madison.

Appendices

Appendix A

Machine Learning Algorithms

Contents

APPENDIX A presents a detailed description of the Support Vector Machine (SVM) and Random Forest (RF) classifier. All descriptions are taken analogously from [Theodoridis & Koutroumbas 2009; Clarke et al. 2009] and [Breiman 2001], if not indicated differently.

A.1 Support Vector Machine

SVMs are based on the assumption that samples originating from two different classes (ω_1 and ω_2) can be separated, if necessary in a transformed space, by a linear hyperplane, i.e. the data samples belonging to the two classes are linearly separable. Whenever more than two classes need to be separated the problem is split into binary classification problems either by testing one class against the sum of all other classes (one vs. all) or each class against each other (one vs. one). The separating hyperplane $g(x)$ of the two classes is defined as

$$g(x) = \mathbf{w}^T \mathbf{x} + w_o = 0, \tag{A.1}$$

with $\mathbf{w} = w_1, w_2, ..., w_l$ as weight vector of the feature vector $\mathbf{x} = [x_1, x_2, ..., x_l]$ in the l-dimensional feature space and w_0 as the threshold. However, this hyperplane is not unique and there exist multiple hyperplanes which can be used to separate both classes (cf. Figure A.1). The basic principle of a SVM classifier is to find the optimal hyperplane which maximizes the distance between two classes such that the margin corresponds to $2z$, with z being the distance between the hyperplane and the nearest data samples belonging to class ω_1 and ω_2, respectively. The distance between the hyperplane and the data samples is determined as

$$z = \frac{|g(x)|}{\|\mathbf{w}\|}. \tag{A.2}$$

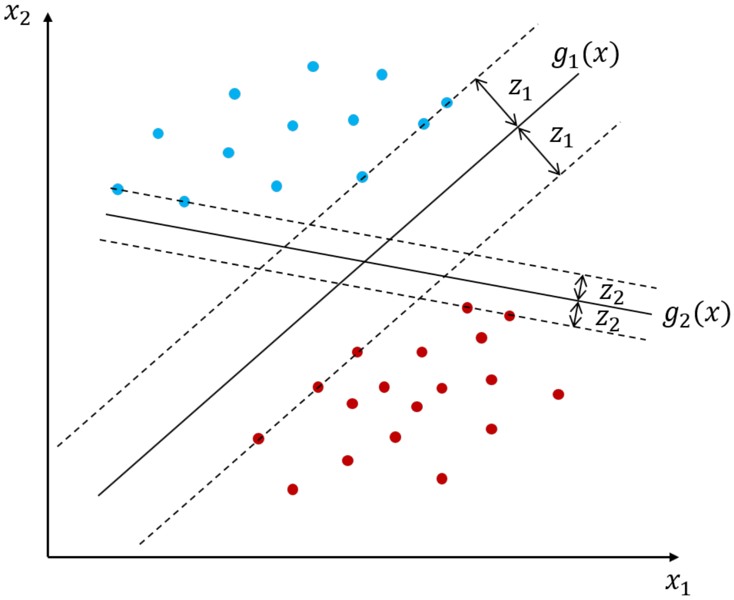


Figure A.1: Data samples of a two-dimensional, two-class classification problem and two possible separating hyperplanes $g_1(x)$ and $g_2(x)$. The solid lines correspond to the separating hyperplanes and the dashed lines to the margin of the corresponding SVM-classifier of distance z_1 and z_2, respectively. The red and blue color indicates the membership of the data samples belonging to the two considered classes.

The nearest sample point belonging to the classes ω_1 and ω_2 are referred to as *support vectors*. To simplify the optimization problem the supporting hyperplanes, defined by the support vectors, are scaled by $\mathbf{w}$ and w_0 such that $g(x) = \pm 1$. By doing so the margin now corresponds to $\frac{1}{\|\mathbf{w}\|} + \frac{1}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|}$ such that

$$\mathbf{w}^T\mathbf{x} + w_o \geq 1, \forall \mathbf{x} \in \omega_1, \text{and} \tag{A.3}$$

$$\mathbf{w}^T\mathbf{x} + w_o \leq -1, \forall \mathbf{x} \in \omega_2. \tag{A.4}$$

This leads to the following constrained non-linear quadratic optimization problem:

$$\begin{aligned} \text{minimize} \quad & J(\mathbf{w}, w_0) = \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1, i = 1, 2, ..., N, \end{aligned} \tag{A.5}$$

with y_i being the class indicator ($y_i = +1$ for ω_1 and $y_i = -1$ for ω_2) and N being the total number of data samples. This optimization problem can be solved as described in Appendix C of [Theodoridis & Koutroumbas 2009]. By considering the so called *Lagrangian duality*, the optimization problem represented in its *Wolfe dual representation* is as follows:

$$\text{maximize} \quad \mathcal{L}(\mathbf{w}, w_0, \lambda) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \lambda_i[y_i(\mathbf{w}^T\mathbf{x}_i + \mathbf{w}_0) - 1] \tag{A.6}$$

$$\text{s.t.} \quad \mathbf{w} = \sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i \tag{A.7}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{A.8}$$

$$\lambda \geq 0 \tag{A.9}$$

with λ being the vector of Lagrangian multipliers. By substituting (A.7) and (A.8) into (A.6) the following optimization task is obtained:

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right) \tag{A.10}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{A.11}$$

$$\lambda \geq 0$$

One major disadvantage of the introduced basic SVM is that it only works properly if the classes are linearly separable. In some cases, however, the data samples are non ideally separable by a linear hyperplane. Therefore, the *soft margin* is introduced which allows data samples to lie inside the margin. When introducing the soft margin there exist three different types of data samples:

1. Samples that are correctly classified and lie outside of the margin $(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1)$,
2. samples that are correctly classified and lie inside of the margin $(0 \leq y_i(\mathbf{w}^T\mathbf{x}_i + w_0) < 1)$ and
3. samples that are misclassified $(y_i(\mathbf{w}^T\mathbf{x}_i + w_0) < 0)$.

By introducing the slack variable ξ these cases can be summarized into one inequation:

$$y_i[\mathbf{w}^T\mathbf{x}_i + w_0] \geq 1 - \xi_i, \tag{A.12}$$

with $\xi_i = 0$ for case 1, $0 < \xi_i \leq 1$ for case 2 and $\xi_i > 1$ for case 3. This leads to an adaption of the previous optimization problem (A.5) to:

$$\begin{aligned} \text{minimize} \quad & J(\mathbf{w}, w_0) = \frac{1}{2} \left\|\mathbf{w}\right\|^2 + C \sum_{i=1}^{N} \xi_i \qquad \text{(A.13)} \\ \text{s.t.} \quad & y_i[\mathbf{w}^T \mathbf{x}_i + w_0] \geq 1 - \xi_i, i = 1, 2, ..., N, \\ & \xi_i \geq 0, i = 1, 2, ..., N. \end{aligned}$$

and hence:

$$\begin{aligned} & \max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right) \qquad \text{(A.14)} \\ \text{s.t.} \quad & \sum_{i=1}^{N} \lambda_i y_i = 0 \qquad \text{(A.15)} \\ & 0 \leq \lambda_i \leq C \end{aligned}$$

The parameter C, also referred to as cost-value, is a positive value which penalizes those samples lying inside and on the wrong side of the margin. High values imply a strong penalization of wrongly separated samples, while low values indicate a weak penalization [Steinwart & Christmann 2008]. C is one of the most commonly used parameters used for hyper-parameter optimization (cf. Section 2.2.8).

In most cases, however, a non linearly separable classification problem will not be solved satisfactorily by introducing a soft margin only. When this is the case, one can make use of the so called *kernel-trick*. By applying the kernel-trick, the original l-dimensional data samples are transformed into a higher k-dimensional feature space where the classification problem becomes linearly separable (cf. Figure A.2):

$$x \mapsto \Phi(x) \in H \qquad \text{(A.16)}$$

with H being a Hilbert Space equipped with an inner product operation $\langle \Phi(x), \Phi(y) \rangle$. Considering the optimization problem as presented in (A.14) and with $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = K(\mathbf{x}_i, \mathbf{x}_j)$ the following optimization problem now needs to be solved

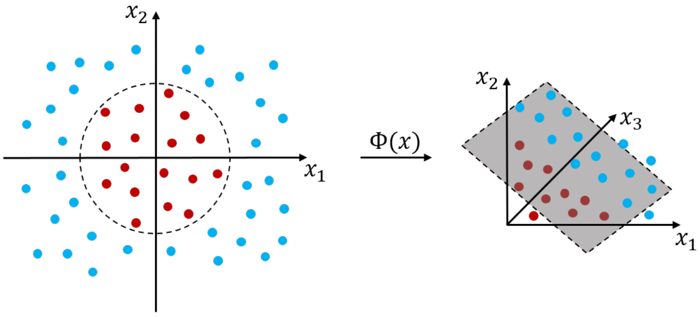


Figure A.2: Data samples of a non linearly separable two-dimensional, two-class classification problem. By applying a kernel ($\Phi(x)$), the two-dimensional classification problem is mapped onto a three-dimensional feature space where the problem is linearly separable by one linear hyperplane. The red and blue color indicates the membership of the data samples belonging to the two considered classes.

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{A.17}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \lambda_i y_i = 0 \tag{A.18}$$

$$0 \leq \lambda_i \leq C$$

where $K(x, y)$ is also referred to as kernel function. Commonly used kernel functions are:

Linear kernel:

$$K(x, y) = \langle \mathbf{x}, \mathbf{y} \rangle \, . \tag{A.19}$$

Polynomial kernel:

$$\begin{aligned} K(x, y) &= \langle \mathbf{x}, \mathbf{y} \rangle^d && \text{(A.20)} \\ &= (\mathbf{x}^T \mathbf{y} + 1)^d && \text{(A.21)} \end{aligned}$$

with d being the degree of the polynomial.

Gaussian radial basis function kernel:

$$\begin{aligned} K(x, y) &= \exp - \frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2} && \text{(A.22)} \\ &= \exp -\gamma \|\mathbf{x} - \mathbf{y}\|^2 && \text{(A.23)} \end{aligned}$$

with $\gamma = \frac{1}{2\sigma^2}$ affecting the width of the Gaussian. Small γ imply a large variance of the Gaussian and vice-versa. This also implies that for large γ

values special attention to overfitting of the training data needs to be drawn [Hsu et al. 2016].

Depending on the utilized kernel, the parameters d or γ are often used for hyperparameter optimization.

A.2 Random Forest

A different machine learning approach used for classification are so-called Random Forests (RFs) [Breiman 2001]. They are based on ensembles of simple binary decision trees, Bootstrap Aggregation (Bagging) and random feature selection at each split.

RFs were first introduced by Breiman in [Breiman 2001]. By combining multiple decision trees, the classification results become more accurate compared to the results of the individual models. One well-established, so called, ensemble method is Bagging [Breiman 1996]. The training set of one tree model is selected randomly from the existing data samples using bootstrapping. Bootstrapping is a re-sampling method to create multiple subsets out of the original data set of same sample size using sampling with replacement. This implies that the new subset can contain the same data sample several times and leave samples of the original set out, respectively. The samples which are left unconsidered in the training set of the individual tree model are called Out-Of-Bag (OOB)-observations. The selected number of decision trees ($numTrees$) defines the number of generated bootstrapped training sets. In the standard Bagging algorithm, the models of the individual trees are trained on each sample of the training set by selecting the best split out of all available features, which is also referred to as Classification and Regression Trees (CART)-approach [Breiman et al. 1984]. In contrast, in case of RF-classification the optimal split is chosen out of a limited number of randomly selected features ($numFeatures$). By limiting the numbers of randomly selected features used at each split the individual tree models are less correlated compared to the standard Bagging algorithm and with a decreasing correlation of the decision trees also the generalization error of the forest decreases. For the evaluation of the RF-classifier a separate test set is not mandatory, as the classifiers performance can be internally evaluated using the OOB-observations. By computing the OOB-error for each tree, which corresponds to the prediction error of the OOB-observations, an independent test set for the individual trees is already present. However, this does not give an information on the global performance of the classifier, but only on the performance of the individual decision trees. An overall evaluation of the classifier is therefore still advisable, also considering subject-independent test results. As the classifier will generate prediction values using the individual decision trees, a majority voting over the individual prediction results is carried out to state the prediction result of the RF-classifier. In addition to the predicted output class, a prediction probability for each test-observation and each class can be calculated by computing the average of the fraction of training-observations of a specific class of the reached tree leaf over all trees. One major advantage of the RF classification approach is that with an increased number of decision trees in the forest the algorithm does not tend to overfit but the generalization error converges towards a limit.

Parameters commonly used for hyper-parameter optimization are $numTrees$ and $numFeatures$. However, there exists no common way to choose the ideal number of $numTrees$ and $numFeatures$, as the parameters are strongly dependent on the number of samples included in the training data set and the number of features in the utilized feature set. During Bagging one training set for each decision tree is generating using sampling with replacement. The number of uniquely generated bootstrapped training sets is therefore limited by the number of training samples and their possible unique combinations. Consequently, with an increasing number of $numTrees$ also the probability of a bootstrapped training sets to occur repetitively in the forest increases. It further needs to be distinguished between small and large training sets. In case of small sets, depending on the research question, the samples are not able to represent the population in a sufficient way. Hence, it is wise to choose large numbers of trees. In case of large sample sizes, however, the samples themselves are more representative and $numTrees$ can be chosen in a much lower region. Furthermore, deciding on the right $numTrees$ is highly dependent on the size of $numFeatures$, as with a high number of features chosen from at each split of the decision tree, the probability of a certain feature to be chosen at this split decreases [Liaw & Wiener 2002]. Hence, with a too low number of trees certain features might not be used in the decision process of the random forest.

As already stated, the idea of choosing features randomly at each split of the decision tree is to de-correlate the individual trees in the forest. By choosing the features on a random basis, all trees utilize a different, random, combination of features. In this process, only the size of the randomly chosen feature subset is given ($numFeatures$) and not the features themselves. For small random feature subsets, the probability of a feature to occur in multiple sets decreases and hence the de-correlation of the individual decision trees increases and the generalization error decreases. In [Hastie et al. 2009] a recommendation on how to choose $numFeatures$ is given. In case of classification problems, the authors recommend to use a default value of $numFeatures = \sqrt{(p)}$, and, in case of regression problems, a value of $numFeatures = p/3$. The parameter p denotes the total number of features included in the feature set. The higher number of features needed for regression problems is already addressed in [Breiman 2001], where it is stated that the generalization error decreases much slower for regression problems than for classification problems. These values, however, should only be used if a tuning of the classifier is not essential for a good performance.

Lastly, it needs to be emphasized that the number of decision trees largely affects the computational cost of the classifier, not only during training but also when applied to unknown data samples. Especially for a later real time implementation, a good trade-off between the computational cost and the generalizability of the classifier has to be met. Pre-studies, presented in [Liaw & Wiener 2002], have shown that for large training sets an upper limit of $numTrees = 1000$ leads to a reasonable performance of the classifier while keeping the computational costs in an acceptable region.

Appendix B

Relevant Audio Codecs

APPENDIX B gives an overview on the most commonly used audio codecs, applied in everyday telecommunication and audio streaming applications. A description of the different audio coding technologies, utilized in the now presented audio codecs, is given in Section 2.5.3 of this Thesis.

Waveform Audio File (WAV) is the standard Windows format to store raw audio material. It was introduced by Microsoft and IBM in 1991 and is based on the Research Interchanged File Format (RIFF) container format [IBM Corporation & Microsoft Corporation 1991]. It uses a linear pulse code modulation encoding, where the magnitude of the signal is linearly quantized at regular sample points, given by the sampling rate of the raw audio signal. The bit rate describes the number of bits stored in one second of the signal. This is dependent on the bit depth (bits per sample). Audio signals usually have a bit depth of 16 bit. For a high quality stereo recording with a sampling rate of 44.1 kHz this would result in a bit rate of 1411.2 kbit/s (44.1 kHz·16 bit·2 = 1411.2 kbit/s).

Free Lossless Audio Codec (FLAC) is a non-proprietary, fast and widely supported lossless audio codec [Coalson & Xiph.Org Foundation 2020]. It was developed by the Xiph.Org Foundation in 2001 and is based on linear predictive coding. Unlike other audio codecs the compression is not defined by a given bit rate but by nine compression levels (0–8) from *fast/ low* (0) to *slow/ high* (9) compression. The different compression levels are not achieved by constant bit rates but by variable bit rates, which dynamically adapt to the sound file. The decoded signal is indistinguishable from the original WAV file, therefore, this compression is also referred to as lossless.

MPEG-1/ MPEG-2 Audio Layer-3 (MP3) was developed by the Moving Picture Experts Group (MPEG) and is a lossy audio codec based on perceptual coding using Modified Discrete Cosine Transforms (MDCT) [Brandenburg 1999]. The first development phase already started in 1988 as MPEG-1 audio standard. MPEG-1 consists of three operating modes (layers) for high sampling rates (i.e. 32, 44.2 and 48 kHz). With each layer, the complexity and performance of the codec increased. Layer-3 represents the highest complexity mode, providing the highest quality at low bit rates. It supports bit rates from 32 to 320 kbit/s. This standard was introduced in 1992 as part of the international standard on coding of moving pictures and associated audio (ISO/IEC 11172-3:1993) [ISO/IEC 1993]. The second phase of MPEG was finalized in 1994 as MPEG-2 audio standard (ISO/IEC 13818-3:1998) [ISO/IEC 1998] with a support of additional low sampling rates (i.e. 16, 22.05 and 24 kHz). This standard supports bit rates from 8 to 160 kbit/s. In case of MPEG-1/2 Layer-3 a switch between the bit rates of each audio frame is supported. This enables variable bit rate coding as well as constant bit rates. Furthermore, a backward compatibility between the MPEG-1 and MPEG-2 standards is given.

Advanced Audio Coding (AAC) was developed as the successor of the MP3 audio codec [Brandenburg 1999]. It is a lossy audio codec based on perceptual coding and was introduced by the MPEG in 1997 as enhanced multi-channel coding standard (ISO/IEC 13818-7:2006) [ISO/IEC 2006]. With this new audio standard a backwards compatibility with the MPEG-1 codec was terminated. AAC follows the same coding strategy as MP3 but with enhanced coding efficiency and quality improvement at low bit rates. The codec can be used within a wide range of sampling frequencies (8 to 96 kHz) and bit rates (16 to 128 kbit/s per audio channel). One prominent streaming service provider using AAC is Spotify [**Spotify:2020**]. Depending on the network connection and the player type (i.e. webplayer or desktop version), bit rates from 24 kbit/s (low streaming quality) up to 320 kbit/s (very high streaming quality, only available for Spotify premium users) are utilized.

Windows Media Audio (WMA) was developed as a competitor of MP3 and is a lossy audio codec designed for music compression based on perceptual coding. It uses a proprietary coding technology developed by Microsoft and was first released in 1999. To date there are four different versions of the WMA codec, namely, standard WMA, WMA Professional, WMA Lossless and WMA Voice. All of them have been developed to be used in different application domains [Microsoft 2018]. The standard WMA codec supports bit rates from 64 to 192 kbit/s with constant or variable bit rates. WMA Professional, additionally, supports multiple channel settings as stereo, 5.1 channel and 7.1 channel surround sound at 128 to 768 kbit/s. It is stated that for 5.1 channel surround sound with a compression bit rate of 384 kbit/s no audible difference compared to the original music file is perceived. The WMA Lossless codec is a lossless version of the standard WMA codec that creates a bit-for-bit duplicate of the original audio file. Last, WMA Voice is specially designed for audio files containing speech. It has a mixed mode that can be used to compress audio files containing speech and music and supports low bit rate compression from 4 to 20 kbit/s.

Adaptive Multi-Rate(AMR) was developed by the 3rd Generation Partnership Project (3GPP) and the European Telecommunication Standards Institute (ETSI) as lossy standard speech codec designed for narrowband (200-3400 Hz) mobile communication in 1999 [3GPP/ETSI 2018a]. It is based on Analysis-by-Synthesis (AbS) and utilizes the Multi-Rate Algebraic Codec-Excited Linear Prediction (CELP) (ACELP) (MR-ACELP) technology. The Adaptive Multi-Rate (AMR) codec supports eight compression modes with bit rates ranging from 4.75 to 12.2 kbit/s. There exist two successors of the AMR codec. The first on, AMR-WB, was introduced by 3GPP/ETSI in 2001 [3GPP/ETSI 2018b] and approved by the ITU-T as Recommendation G.722.2[1] [ITU-T 2003c]. As its precursor, it is also based on MR-ACELP, but provides an extended bandwidth of 50 Hz to 7 kHz and supports nine compression modes with bit rates from 6.6 to 23.85 kbit/s. A second successor was introduced as new telecommunication standard AMR-WB+ in 2004 by 3GPP/ETSI [3GPP/ETSI 2018c]. In comparison to its precursors, a hybrid coder is utilized, which is either based on ACELP, for speech parts, or filter bank-based transform coded excitation (perceptual coding), for non-speech parts, that can be switched for each frame of the signal. It supports bit rates from 5.2 to 48 kbit/s.

Speex (SPX) was started as project of the Xiph.Org Foundation in 2002 to address the need of a free, open-source speech codec [Valin 2006]. It is a lossy audio codec and mainly designed for the application in Voice over IP and not for mobile telephony. It utilizes AbS based on the CELP audio coder and supports different quality levels that range from 0 to 10 [Valin 2007]. This quality parameter controls the tradeoff made between the speech quality and the bit rate and is also referred to as compression level. As bit rates a range from 2 to 44 kbit/s is supported by the codec. In case of constant bit rates the quality parameter is denoted as an integer, in case of variable bit rates as a float. The codec can operate in three different modes, namely narrowband (up to 8 kHz), wideband (up to 16 kHz) and ultra-wideband (up to 32 kHz). Early versions of Apples personal assistant Siri (iPhone 4S) are named to have uses SPX as audio codec [Schwan 2011].

OPUS is the successor of SPX and was developed by Skype in cooperation with the Xiph.Org Foundation in 2012 as communication standard [Valin et al. 2012]. It is a lossy audio codec and uses hybrid coding, namely, a modified version of Skypes SILK codec and Constrained Energy Lapped Transform (CELT), developed by the Xiph.Org Foundation. Unlike AMR-WB+, the codec uses either both coding strategies (SILK and CELT encoding) in parallel or one of them solely, based on the utilized bit rate setting. The bit stream is generated through range encoding and contains bits of the SILK and CELT encoders. Depending on the audio stream (speech or music), either SILK, which is based on linear predictive coding (AbS), or CELT, which is based on MDCT (perceptaul coding) can be utilized. The codec can be used in

[1] http://handle.itu.int/11.1002/1000/6506

narrowband (up to 8 kHz), medium-band (up to 12 kHz), wideband (up to 16 kHz), super-wideband (up to 24 kHz) and fullband (up to 48 kHz) operation mode. The hybrid mode can be applied in super-wideband and wideband mode. Here, SILK codes the low frequency bins of the signal and CELT the high frequency bins. The cut of lies at 8 kHz, the maximum wideband speech audio bandwidth. The supported bit rates range from 6 kbit/s in narrowband mode to 510 kbit/s in fullband mode, this also enables an application for surround sound application.

Appendix C

Statistical Analysis of Parametric and Non-Parametric Data

In this Chapter I will give an overview on the statistical analysis methods utilized in this Thesis based on the descriptions presented in [Bortz & Lienert 2008] and [Bortz & Schuster 2010]. A main focus is drawn on the analysis of parametric data (e.g. normally distributed data), as it is the case for naturally formed data like speech, and the method of Analysis of Variance (ANOVA) in combination with post-hoc t-tests, in case of multiple testing. Generally, these measures are used to identify if a factor (e.g. medication), or multiple factors (e.g. medication and gender), has a significant effect on an observed measure (e.g. blood pressure) of a representative number of test subjects per factor group (e.g. placebo, single dose and double dose). Commonly, the ANOVA is used to determine if there exists a general significant effect of the factor on the observed measure. Afterwards, a post-hoc t-test reveals how the different factor groups affect the observed measure. These post-hoc tests can be seen as multiple testing, as the measured data samples of each factor group are tested multiple times against each other (cf. Section C.4). In case of a factor only comprising two factor groups (e.g. gender: female and male) the results of the post-hoc t-test correspond to the results obtained when utilizing an ANOVA under similar testing conditions. As no testing on multiple factors was applied in the scope of this Thesis, these methods will not be addressed.

In general, this kind of testing is also called statistical hypothesis testing and is based on an alternative hypothesis (H_1) and a null hypothesis (H_0). The alternative hypothesis describes a newly made assumption, which has not yet been proven to be correct (e.g. a certain medication leads to differences in the average blood pressure of a patient). It is distinguished between one-sided hypotheses (directional, e.g. the measure increases/ decreases) and two-sided hypotheses (non directional, e.g. it is unknown if the measures increases/ decreases but assumed that there exists a difference). The null hypothesis describes the counter hypothesis of the alternative hypothesis (e.g. the blood pressure stays constant at an average of μ_0). Regarding these hypotheses, three possible hypotheses pairs can be defined considering the observed measure average (μ):

1) $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ (one-sided, the measure increases),

2) $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$ (one-sided, the measure decreases),

3) $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ (two-sided, there exists a difference).

To determine if a null hypothesis gets rejected, the level of significance α is introduced. This level denotes the probability of a null hypothesis being rejected with regard to committing a type I error (the null hypothesis gets incorrectly rejected). Regarding the example introduced above, this could imply a significant effect being incorrectly determined between the double dose and placebo factor group. Especially, in the application domain of medication efficiency, this could result in serious problems when a certain medication treatment is applied, but actually not affecting the concerned health issue. Or even more dramatically, a patient being diagnosed with a serious disease while actually being healthy and vice versa. Commonly used levels of significance range from $\alpha = 0.05$, in case of less serious consequences, to $\alpha = 0.001$, in case of very serious consequences. To decide if the null hypothesis is satisfied regarding the given level of significance, a test statistic needs to be evaluated which is then compared to the critical value of the underlying samples distribution function. Considering a one-sided hypothesis as described in 1) and 2), the decision on the null hypothesis being accepted or rejected is defined as followed:

for 1)

test statistic $\geq$ crit. value $\rightarrow H_0$ rejected
test statistic $<$ crit. value $\rightarrow H_0$ accepted

for 2)

test statistic $\leq$ crit. value $\rightarrow H_0$ rejected
test statistic $>$ crit. value $\rightarrow H_0$ accepted

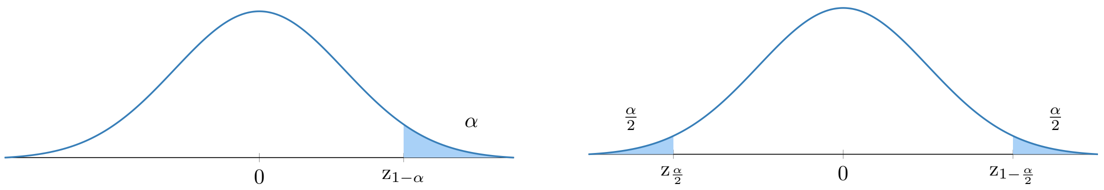
In case of a two-sided hypothesis as described in 3), H_0 is rejected when $|\text{test statistic}| \geq$ crit. value. The main difference when utilizing different statistical analysis tests is the utilized test statistic and corresponding critical value. Mostly, the test statistic is calculated based on the statistics of a set of random samples (average and standard deviation) of the observed measure for each factor group. Whenever huge sets of random samples are available the test statistic is assumed to describe the distribution of the real population. When utilizing parametric data, the samples are assumed to be standard normally distributed. In case of an unknown distribution function of the underlying population, the average and standard deviation of the population needs to be approximated using the random samples statistical values. This distribution is then called a t-distribution. When the distribution of the real population is available the test statistic is described using the z-distribution (normal distribution). Depending on these two factors either a z-test or a t-test is utilized. Assuming that the population is fully described, utilizing a z-test, the critical z-value (z_{crit}) is defined by the percentiles of the corresponding level of significance. For a one-sided z-test this would correspond to the value $z_{1-\alpha}$ or z_α for case 1) and 2), respectively. A visualization of the region of rejection, with the hypothesis that an increase of the considered measure is present (case 1)), is given in Figure C.1 a). For a two-sided z-test (case 3)), the region of rejection is split into a lower and upper region of $\frac{\alpha}{2}$ and the critical z-values would correspond to $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ (cf. Figure C.1 b)). For convenience, there exists tables from which the critical z-values can be easily obtained (e.g. in [Bortz & Schuster 2010] and [Kvam & Vidakovic 2007]). It is now assumed, that the random samples' distribution is consistent for the different factor groups (i.e. $\mu = \mu_0$, the null hypothesis is accepted and there is no significant difference) with

$$x_i \, \mathcal{N}(\mu_0, \sigma^2), \mathrm{for} i = 1, ..., n.$$

The term, x_i represents the measured data of the random sample set of size n. Hence, a z-value can be determined for the random sample set

$$z = \sqrt{n}(\frac{\bar{x} - \mu_0}{\sigma}), \tag{C.1}$$

that corresponds to the test statistics and is later compared to the critical z-value. Furthermore, it can be calculated what percentage of the random sample set lie above the test statistic. This value is called the p-value and can be directly compared to the utilized level of significance. Whenever p exceeds α, the null hypothesis is accepted and no significant difference between the different factor groups is present (H_1 gets rejected).



(a) one-sided z-test with the hypothesis $H_1 : \mu > \mu_0$

(b) two-sided z-test

Figure C.1: Standard normal distribution with critical z-values for a a) one-sided z-test with $H_1 : \mu > \mu_0$ and b) two-sided z-test. The critical z-values correspond to the $1 - \alpha$-percentile in a) and the $\frac{\alpha}{2}$- and $1 - \frac{\alpha}{2}$-percentile in b), of the z-distribution.

To sum up, statistical hypothesis testing determines if there exists a significant difference between the average value of an observed measure and the null hypothesis. It is assumed that the utilized random sample set is normally distributed. With regards to these assumptions, a test statistic and critical value can be determined which is then compared to each other, based on a given level of significance (α). If the random samples' set sufficiently describes the population, the data is assumed to be standard normally distributed (z-distribution) and a simple z-test can be utilized. In most cases the distribution of the population is, however, not sufficiently described by the random sample set and a t-test needs to be applied. To compare two factor groups or more with each other the hypothesis test needs to be adapted. Another aspect which needs to be addressed is whether the random sample set contains paired data samples (e.g. samples originating from the same subject in different factor groups). The just presented general approach is only valid for unpaired data samples. In the following Section C.1 more details on these hypothesis testing methods, utilized in the scope of this Thesis, are presented. These include paired and unpaired testings of factors including two groups and multiple groups. The general idea behind these test methods, however, follows the just presented description of the z-test.

C.1 Parametric Statistics

In this Section I will present all relevant hypothesis testing methods, which are found to be applied to parametric data in the Chapters of this Thesis. A special focus is drawn on paired testing, as the performed experiments are based on comparisons within the utilized subjects (e.g. effect of speech enhancement on the speech emotion recognition performance compared to the performance of the corresponding clean speech experiments). The following descriptions on parametric statistic and hypothesis testing are based on [Bortz & Schuster 2010], if not indicated differently.

C.1.1 t-Test

As already mentioned in the previous Section, the obtained random sample set of an observed measure does, in most cases, not sufficiently describe the real population. Therefore, the data is not assumed to be z-distributed but t-distributed. Now, the test statistic can be calculated by

$$t = \sqrt{n}(\frac{\bar{x} - \mu_0}{s}). \tag{C.2}$$

with s being the approximated standard deviation of the population obtained from the random sample set. Now, as described for the z-test, a critical t-value needs to be obtained by determining the percentile of the corresponding t-distribution under one-sided or two-sided testing conditions. Here, the following aspects of a t-distribution need to be considered:

1. The t-distribution is strongly dependent on its degree of freedom ($df = n-1$), which determines the shape of the distribution.
2. It is unimodal and symmetric. With increase of degree of freedom it converges towards a standard normal distribution.
3. The percentiles of the t-distribution are used as critical values and are referred to as t_{df}.

4. t_{df} can be taken from the corresponding table provided in several books on hypothesis testing (e.g. [Bortz & Schuster 2010]).

The only requirement to apply this, so called, one sample t-test is that the random sample set is normally distributed.

With the presented approach we are now able to compare a t-distributed random sample set to a given average of the null hypothesis. For most applications, however, a comparison between different factor groups is needed. In case of two factor groups the hypothesis testing needs to be re-defined. The null hypothesis now assumes that the there exists no difference between the observed measure of the random sample set obtained for the two factor groups and the alternative hypothesis assumes that there exists a difference:

1) $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 > \mu_2$ (one-sided, the measure increases),

2) $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 < \mu_2$ (one-sided, the measure decreases),

3) $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ (two-sided, there exists a difference).

It now needs to be further distinguished between unpaired (non-related) and paired (related) data samples. In case of two unpaired independent data sets of size n_1 and n_2, the test statistic, as presented in Equation (C.2), can be adapted to

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}, \tag{C.3}$$

and

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{s_p^2 (\frac{1}{n_1} + \frac{1}{n_2})}. \tag{C.4}$$

In this Equation, $\bar{x}_1$ and $\bar{x}_2$ correspond to the average values' observed measures ($x_{i_1,1}$ and $x_{i_2,2}$) of the random sample sets. With regard to the null hypothesis ($\mu_1 = \mu_2$) and the assumption that both random sample sets have common variances, s_p^2 is determined by

$$s_p^2 = \frac{s_1^2 + s_2^2}{2}, \tag{C.5}$$

and the test statistic t is calculated as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}}. \tag{C.6}$$

As two random sample sets with possibly different sample set sizes are utilized, the corresponding t-distribution is now dependent on $df = n_1 + n_2 - 2$ degrees of freedom. As for the one sample t-test, the critical t-value $t_{df;p}$ can be taken from the corresponding table provided in books. Contrarily to the one sample t-test, to apply the unpaired t-test, an additional requirement needs to be fulfilled. This is the homoscedasticity of the random sample sets variances.

In case of paired data samples, the random sample sets are of equal size ($n_1 = n_2$) and the different samples of set 1 can be directly compared to their corresponding pair in set 2. Consequently, there exist $n_d = n_1 = n_2$ pairs of samples and the difference d_i between each sample pair $(x_{i,1}, x_{i,2})$, with $i = 1, ..., n_d$, can be determined as

$$d_i = x_{i,1} - x_{i,2}. \tag{C.7}$$

Accordingly, also the differences of the average sample values can be defined as

$$\bar{d} = \bar{x}_1 - \bar{x}_2. \tag{C.8}$$

Considering these simplifications, the two factor t-test is transformed back into a one sample t-test, with the test statistic defined as

$$t = \sqrt{n_d}(\frac{\bar{d} - \mu_{\bar{d}}}{s_d}). \tag{C.9}$$

The term s_d is defined as the sample standard deviation of the differences d_i. With regard to the null hypothesis ($\mu_1 = \mu_2$) and consequently $\mu_d = 0$, the test statistic is calculated by

$$t = \sqrt{n_d}(\frac{\bar{d}}{s_d}). \tag{C.10}$$

As we reduced the two sample problem to a one sample problem, the degree of freedom of the samples t-distribution is determined by $df = n_d + 1$ and, again, the critical t-value $t_{df,p}$ can be taken from available tables. The requirements to apply a paired t-test are similar to those of a one sample t-test, with the difference pairs d_i being normally distributed.

C.1.2 Analysis of Variance

Whenever a factor contains more than two factor groups, it is not possible anymore to apply a t-test but a, so called, Analysis of Variance (ANOVA) needs to be employed. Here, a factor A contains $i = 1, ..., p$ factor groups. The observed measure of each factor group i and test subjects m is denoted as $y_{i,m}$, with $m = 1, ..., n$. It further is assumed that each factor group contains the same number of observations, such that n is equal for all factor groups and $y_{i,m}$ spans a matrix of size $p \times n$, with $N = n \cdot p$ entries.

As for the hypothesis testing methods presented previously, we want to determine if there exists a difference between the average value of the observed measure of the p factor groups. Following this assumption, the null hypothesis is defined as

$$H_0 : \mu_1 = \mu_2 = ... = \mu_p, \tag{C.11}$$

and the alternative hypotheses states that there exists a difference between the μ_i-value of at least two factor groups.

In case of a single factor ANOVA, the difference in efficacy of each factor group is determined by the sum of squares of its observed measure. It is known that the sum of squares describes the overall variation of the observed measure (i.e. total variance). With this knowledge, an F-test, also used to determine if there exists a significant different between the variance of different factor groups, is utilized (cf. Section C.3). For this test, the test statistic follows an F-distribution and is calculated as

$$F = \frac{MS_A}{MS_e} = \frac{\frac{SS_A}{df_A}}{\frac{SS_e}{df_e}}, \tag{C.12}$$

with MS_A and MS_e denoting the mean square of the sum of squares in between the factor groups (SS_A) and within each factor group (SS_e), respectively. The Terms SS_A and SS_e are determined by

$$SS_A = n \cdot \sum_i (\bar{A}_i - \bar{G})^2, \text{ and} \tag{C.13}$$

$$SS_e = \sum_i \sum_m (y_{i,m} - \bar{A}_i)^2, \tag{C.14}$$

and the correspond degrees of freedom by

$$df_A = p - 1, \text{and} \tag{C.15}$$

$$df_e = p \cdot (n - 1) = N - p. \tag{C.16}$$

The term $\bar{A}_i$ denotes the average observed measure of the factor group i ($\bar{A}_i = \frac{1}{n} \sum_m y_{i,m}$) and $\bar{G}$ the average observed measure over all factor groups ($\bar{G} = \frac{1}{N} \sum_i \sum_m y_{i,m}$). With regard to Equation (C.12), the critical F-value F_{df_A,df_e} is dependent on the two degrees of freedom and the desired level of significance. While for the t-test it was distinguished between a one-sided and a two-sided t-test, in case of the single factor ANOVA we are only interested in if there exist a difference between two of the factor groups. Therefore, the H_0 is rejected whenever

$$F \geq F_{df_A,df_e} \tag{C.17}$$

is satisfied. The critical F-values defined by the percentiles of the F-distribution can, again, be taken from tables provided in several books on hypothesis testing (e.g. [Bortz & Schuster 2010]). Requirements relevant to apply the single factor ANOVA are normally distributed random sample sets and homoscedasticity of the sets.

We now need to further distinguish between unpaired and paired ANOVAs. In case of unpaired data samples, there exist no dependency of the different subjects of the factor groups with each other. Therefore, all factor groups are independent and the total sum of squared is determined as

$$SS_{tot} = \sum_i \sum_m (y_{i,m} - \bar{G})^2 = SS_A + SS_e, \text{ and} \tag{C.18}$$

$$df_{tot} = N - 1 = df_A + df_e. \tag{C.19}$$

The applied single factor ANOVA is also referred to as one-way ANOVA.

In case of paired data samples, there exists an additional dependency of the subjects of each factor group. To evaluate if a certain factor group has a significant effect on the observed measure, however, it is not of interest to include the in between subject differences in the test statistics, as these differences are already known and would distort the effect of the factor groups. Hence, the total sum of squares is split into an in between and within subject variance:

$$SS_{tot} = \sum_i \sum_m (y_{i,m} - \bar{G})^2 = SS_{between} + SS_{within}, \text{ with} \tag{C.20}$$

$$SS_{between} = p \cdot \sum_m (\bar{P}_m - \bar{G})^2 \text{ and} \tag{C.21}$$

$$SS_{within} = \sum_i \sum_m (y_{i,m} - \bar{P}_m)^2 = SS_A + SS_e. \tag{C.22}$$

While SS_A still only takes into account the in between factor group difference (cf. Equation (C.13)), SS_e is now determined by

$$SS_e = \sum_i \sum_m (y_{i,m} - \bar{A}_i - \bar{P}_m + \bar{G})^2. \tag{C.23}$$

The corresponding degrees of freedom are:

$$df_{tot} = N - 1 = df_{between} + df_{within}, \text{ and} \tag{C.24}$$

$$df_{between} = n - 1, \text{and} \tag{C.25}$$

$$df_{within} = df_A + df_e = (p-1) + (n-1) \cdot (p-1) = N - n. \tag{C.26}$$

Considering the adapted term of SS_e, the test statistic F is determined as presented in Equation (C.12) and the critical F-value F_{df_A,df_e} can be taken from available tables. This kind of ANOVA is referred to as repeated-measures ANOVA, as it is originally designed to compare repeated-measures of the same subject under different testing conditions (e.g. blood pressure without medication, with placebo and with medication, respectively). Nevertheless, it is also often used to compare different moments of observation with each other (e.g. blood pressure after one week, two weeks, ... of medication, respectively). To apply the repeated-measures ANOVA, the random sample sets of the different factor groups need to be normally distributed. Contrarily to the one-sided ANOVA, no homoscedasticity, but sphericity (homoscedasticity of the difference between the paired samples) is required.

An overview on how the different requirements of the unpaired and paired t-test, and one-way and repeated-measures ANOVA are determined is presented in Section C.3.

C.2 Non-Parametric Statistics

In case of small or not normally distributed data samples, it is not recommended to utilize parametric statistics. For each of the above presented hypothesis testing methods, there exist designated testing approaches in case of non-parametric data. Again it is distinguished between two factor groups and multiple factor groups, and unpaired and paired data samples. In case of two factor groups and unpaired data samples the Mann-Whitney U-test is applied. For paired data samples of two factor groups Wilcoxon (signed-)rank test can be used (cf. [Bortz & Lienert 2008] and [Bortz & Schuster 2010]). Whenever more than two factor groups need to be evaluated, Kruskal-Wallis H-Test and Friedman-test in case of unpaired and paired data samples, respectively, need to be applied (cf. [Bortz & Lienert 2008]). As in this Thesis only paired samples of two factor groups will be compared, only the Wilcoxon signed-rank test is described in detail in the present Section based in the descriptions in [Bortz & Lienert 2008].

C.2.1 Wilcoxon Signed-Rank Test

This test is specially designed to test the significance of two related samples of non-parametric distributions. Wilcoxon signed-rank test is based on the paired t-test described in the previous Section. As described in Equation C.7 the difference between the observed measure of the two factor groups is calculated, with n_d denoting the number of samples of each factor group. Contrarily to the paired t-test, the test statistic is not based on the difference of each sample pair (d_i) but the rank of the absolute difference $|d_i|$. The rank of all sample pairs i is determined in ascending order from 1 to n_d (e.g. the minimal difference is assigned with rank 1 and the maximal difference assigned with rank n_d). Afterwards, the rank is weighted by the sign of d_i and divided into two sub-classes of positive and negative rank values. The terms T_+ and T_- comprise the sum of ranks of the positive and negative sub-class, respectively. The test-statistic can now be defined as

$$T = min(T_+, T_-). \tag{C.27}$$

When applying an one- or two-sided test, to determine if there exists a significant difference in the considered feature and hence reject the hull hypothesis ($H_0 : \mu_1 = \mu_2$), the test statistic T should not exceed the critical value of the considered level of significance ($T < T_{crit}$). The critical value T_{crit} can be taken from tables provided in several books on non-parametric statistics (e.g. [Bortz & Lienert 2008] and [Kvam & Vidakovic 2007]).

In case of large sample sets $n_d > 50$, T converges to a normal distribution and the z-value can be approximated using the standardized normal distribution (z-distribution). This z-value is then compared to the critical z-value of the considered level of significance. If the z-value exceeds its critical value ($z < z_{crit}$) the null hypothesis is rejected and it can be assumed that there exists a significant difference in the considered feature of the two factor groups.

C.3 Review of Requirements

In the previous Sections C.1 and C.2, different hypothesis testing methods to evaluate whether certain factors have a significant effect on an observed measure are presented. The use of these methods is, however, conditioned by certain requirements on the utilized random sample sets. These requirements are: the random sample sets being normally distributed, and homoscedasticity and sphericity of the sets. For all these requirements there exist designated hypothesis testing methods. I will now only give a brief overview on available testing methods, without a detailed description of the methodological background. Only to name some, the following tests can be utilized:

Normally Distributed Samples: Shapiro-Wilk-Test [D'Agostino 2006]

Homoscedasticity: F-test or Leven-Test [Bortz & Schuster 2010]

Sphericity: Mauchly-Test [Rasch et al. 2014]

If these requirements are not fulfilled by the utilized random sample set there exist ways to modify the applied t-test and ANOVA. In case of not normally distributed sample sets and heteroscedasticity, a non-parametric approach can be chosen (cf. Section C.2).

In case of non-spherical data a correction of degree of freedoms can be applied. However, the reliability of the Mauchly-test is controversial, as, especially for a low number of test subjects the test shows a low statistical power (i.e. data being not spherical, although the Mauchly-test obtained no significant result). At the same time, in case of a large number of test subjects, the test often obtains significant results while the data is actually spherical [Rasch et al. 2014]. Therefore, it is recommended to always apply an adjustment of the degree of freedom of both, numerator ($df_A = p-1$) and denominator ($df_e = (p-1)\cdot(n-1)$), of the critical F-value F_{df_A,df_e}, independent on the outcome of the Mauchly-test. This is done by weighting the degree of freedom with a factor ε, with $\varepsilon < 1$. Depending on the strength of violation of the assumption of sphericity, either a small ε, for strong violations, or a value close to 1, for weak violations, is chosen. In general, the smaller ε is chosen, the stronger the adjustment of the degree of freedom. This adjustment will lead to an increase of the critical F-value and therefore to a progressive decision on H_0.

There exist different ways to determine ε. The most conservative adjustment is obtained when choosing the lowest possible value for ε:

$$\varepsilon = \frac{1}{p-1}. \tag{C.28}$$

The correction of the degree of freedom is obtained by calculating

$$df_A = \varepsilon \cdot (p-1) \text{ and} \tag{C.29}$$

$$df_e = \varepsilon \cdot (p-1) \cdot (n-1). \tag{C.30}$$

This correction is also referred to as the *lowerbound-correction.*

A less conservative method is introduced by Geisser and Greenhouse. Here, ε is estimated based on the covariance matrix C, of size $p \cdot p$, of the random sample sets:

$$\hat{\varepsilon} = \frac{p^2 \cdot (\bar{c}_{ii} - c_{..})^2}{(p-1)\cdot[\sum_{i=1}^{p}\sum_{j=1}^{p} c_{ij}^2 - 2 \cdot p \cdot \sum_{i=1}^{p} \bar{c}_{i.}^2 + p^2 \cdot c_{..}^2]}, \tag{C.31}$$

with $\bar{c}_{ii}$ denoting the average of the main diagonal of C, $\bar{c}_{i.}$ the average of the i-th column of C and $\bar{c}_{..}$ the average over all elements of C. An adjustment utilizing this estimate $\hat{\varepsilon}$ is also referred to as *Greenhouse-Geisser-correction*. In case of $\hat{\varepsilon} < 0.75$ it is recommended to apply a Greenhouse-Geisser-correction (cf. [Rasch et al. 2014] and [Bortz & Schuster 2010]). For larger values ($\hat{\varepsilon} \geq 0.75$) a less conservative correction is recommended, which is called the *Huynh-Feldt-correction*. This approach is dependent on the number of factors evaluated. As only single factor evaluations are performed in the scope of this Thesis, the corresponding Huynh-Feldt-epsilon ($\tilde{\varepsilon}$) is determined by

$$\tilde{\varepsilon} = \frac{n \cdot (p-1) \cdot \hat{\varepsilon} - 2}{(p-1) \cdot [n - 1 - (p-1) \cdot \hat{\varepsilon}]}. \tag{C.32}$$

It is possible that the estimate $\tilde{\varepsilon}$ exceeds a value of $\tilde{\varepsilon} > 1$. In these cases $\tilde{\varepsilon}$ is set to $\hat{\varepsilon} = 1.0$ and no correction of degree of freedom is performed.

C.4 Multiple Testing

Whenever multiple tests are carried out on the same data, this is referred to as multiple testing, for example, when there exist more than two factor groups (e.g. medication groups: placebo, single dose, double dose), which should be tested among each other (placebo vs. single dose, placebo vs. double dose and single dose vs. double dose) in a number of post-hoc t-tests.

C.4.1 α-Inflation

By conducting multiple tests on the same data, the probability of committing a type I error increases with each additional test. This increase is also referred to as α-inflation. The Family-Wise Error Rate (FWER) describes the probability of committing at least one type I error and is determined by

$$FWER = 1 - (1 - \alpha)^M, \tag{C.33}$$

with M being the number of tests carried out on the data set and α being the level of significance [Bortz & Schuster 2010]. In case of the introduced three factor group problem, with a level of significance of $\alpha = 0.05$ this would result in $FWER = 0.143$.

C.4.2 Bonferroni-Correction

One way to prevent α-inflation is by utilizing a Bonferroni-correction. Here, the level of significance gets reduced depending on the number of tests carried out:

$$\alpha' = \frac{\alpha}{m}, \tag{C.34}$$

this reduced α' is substantially smaller than α. Therefore, it is much harder to reach the level of significance [Bortz & Schuster 2010]. For our example, this would results in an corrected α value of $\alpha' = 0.017$ for the carried out post-hoc t-tests.

Appendix D

Big Tables

APPENDIX D contains all the big Tables related to the results presented in Chapter 6.

Table D.1: Unweighted Average Recall (UAR) of the Random Forest (RF)-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
numTrees [#]	10	10	61	134	555	1000	516	215	711	446	898	298	1000	894
numFeatures [%]	16	50	24	16	16	16	25	35	32	30	36	39	50	40
Set 1	31.51	31.13	31.58	31.41	30.81	30.90	30.99	31.57	31.07	31.45	30.94	31.19	31.10	30.95
Set 2	31.28	30.93	31.09	31.19	30.62	30.74	30.42	30.90	30.71	31.14	30.80	31.03	30.95	30.81
Set 3	32.17	30.87	31.01	31.32	30.94	30.77	31.26	31.65	30.94	30.84	31.01	30.82	31.32	31.17
Set 4	31.54	31.75	31.74	30.64	30.49	30.45	30.66	31.42	31.05	30.92	30.81	31.12	31.01	31.15
Set 5	30.17	32.04	31.69	30.69	30.86	30.80	31.13	31.28	31.18	31.11	31.14	31.22	31.04	31.09
Set 6	32.53	32.68	31.58	31.46	31.30	31.27	31.15	32.32	31.50	31.90	31.51	32.13	31.69	31.60
Set 7	30.96	30.94	32.10	31.29	31.22	31.35	31.74	31.52	31.83	32.06	31.71	31.40	31.86	31.54
Set 8	31.40	31.53	31.45	32.35	31.68	31.29	31.58	31.47	31.47	31.59	31.82	31.72	31.99	31.88
Set 9	31.76	32.08	32.17	32.36	32.02	31.80	31.69	32.15	32.09	31.63	32.03	31.75	31.81	31.94
Set 10	31.02	32.78	32.32	32.06	31.82	31.67	32.11	32.46	32.18	31.88	32.05	32.35	32.27	32.23
Set 11	32.88	31.92	32.62	31.95	31.63	31.74	32.02	31.84	32.08	32.27	31.98	32.03	32.24	32.08
Set 12	31.80	32.71	31.85	32.00	31.95	31.84	31.52	32.30	32.08	31.65	31.10	31.94	31.99	31.92
Set 13	31.49	31.46	32.13	31.54	31.47	31.40	32.08	32.13	32.14	32.18	32.22	32.38	32.24	32.33
Set 14	31.94	33.03	32.47	32.12	31.74	31.72	32.12	32.34	32.20	32.44	32.32	32.12	32.49	32.23
Set 15	32.83	32.54	32.69	31.65	31.80	31.76	32.14	32.63	32.43	32.35	32.58	32.55	32.49	32.98
Set 16	32.68	31.76	32.88	32.16	32.11	32.04	32.34	32.59	32.62	32.10	32.22	32.68	32.51	32.61
Set 17	32.52	32.61	32.79	31.74	31.81	31.88	32.11	32.26	32.64	32.80	32.20	32.64	32.18	32.43
Set 18	32.10	32.66	32.86	32.10	31.84	31.64	32.16	32.48	32.06	32.48	32.29	32.83	32.59	32.67
Set 19	32.34	32.17	32.67	32.43	31.65	31.58	32.19	32.32	32.27	32.07	32.14	32.26	32.35	32.15
Set 20	32.22	32.62	32.28	32.08	31.94	31.51	31.97	32.11	32.10	32.26	32.14	32.31	32.32	31.96

Table D.2: Unweighted Average Precision (UAP) of the RF-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
numTrees [#]	10	10	61	134	555	1000	516	215	711	446	898	298	1000	894
numFeatures [%]	16	50	24	16	16	16	25	35	32	30	36	39	50	40
Set 1	33.44	32.34	40.96	40.08	39.95	39.94	40.46	39.34	40.68	38.55	39.95	40.65	40.59	40.08
Set 2	32.79	32.37	39.07	39.55	39.42	39.57	38.44	39.27	38.10	40.07	37.98	37.67	39.73	37.93
Set 3	32.78	32.78	37.95	40.88	40.58	39.30	38.70	42.07	38.67	38.33	40.37	41.32	40.48	41.63
Set 4	34.04	33.80	39.69	40.49	41.22	41.18	39.07	41.15	39.45	40.97	40.12	41.86	40.94	40.71
Set 5	30.52	34.42	41.97	39.14	42.97	43.47	45.20	43.51	43.25	42.35	44.60	44.64	43.13	43.51
Set 6	34.87	34.27	41.47	44.06	45.02	45.60	43.54	45.15	43.44	46.81	44.77	44.00	45.35	45.32
Set 7	33.15	33.24	43.54	42.55	42.46	43.68	45.39	43.82	44.38	46.58	46.69	43.65	44.68	44.71
Set 8	34.25	33.90	40.98	46.48	46.01	44.76	45.32	41.87	45.68	45.79	45.84	45.34	45.15	45.91
Set 9	33.60	34.97	43.21	46.84	47.89	47.28	46.00	45.20	47.76	45.43	46.64	43.12	45.17	46.29
Set 10	32.81	34.75	43.53	44.03	44.47	45.06	45.68	45.66	46.90	45.27	45.16	46.19	45.36	47.59
Set 11	34.21	33.12	42.26	42.13	44.64	46.11	46.03	42.84	45.99	45.81	44.69	45.27	46.35	44.06
Set 12	34.82	35.18	38.51	44.83	46.70	47.35	42.69	44.40	45.76	45.54	44.71	44.14	45.10	46.91
Set 13	34.06	33.43	43.25	43.66	42.30	43.30	46.86	45.02	46.61	46.34	45.33	44.97	45.95	46.34
Set 14	33.82	33.93	42.03	43.63	43.79	43.49	46.52	44.01	43.53	45.50	47.65	45.14	47.48	47.31
Set 15	35.47	34.78	44.68	42.46	43.48	43.59	47.84	47.13	46.51	45.52	45.77	45.34	46.17	47.04
Set 16	36.10	33.83	44.19	45.41	46.54	46.22	46.91	44.14	46.73	45.71	45.71	46.39	47.46	46.28
Set 17	35.42	34.74	41.66	43.74	45.16	45.52	45.82	42.19	46.38	44.47	45.85	45.31	43.86	46.62
Set 18	34.85	33.73	43.35	40.62	42.95	44.87	43.76	46.42	44.87	45.54	43.59	44.87	45.89	44.30
Set 19	34.33	33.15	43.02	45.99	43.91	43.29	43.79	43.07	44.49	43.55	45.25	45.58	45.17	45.31
Set 20	34.04	34.42	42.38	43.60	44.57	45.14	42.01	42.65	43.33	43.40	43.67	44.19	43.86	44.01

Table D.3: F1-measure of the RF-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
numTrees [#]	10	10	61	134	555	1000	516	215	711	446	898	298	1000	894
numFeatures [%]	16	50	24	16	16	16	25	35	32	30	36	39	50	40
Set 1	32.45	31.72	35.66	35.22	34.79	34.85	35.10	35.03	35.23	34.64	34.87	35.30	35.22	34.93
Set 2	32.02	31.63	34.63	34.87	34.46	34.60	33.97	34.59	34.01	35.04	34.01	34.03	34.79	34.00
Set 3	32.47	31.80	34.13	35.47	35.11	34.51	34.58	36.13	34.38	34.18	35.08	35.31	35.26	35.65
Set 4	32.74	32.75	35.27	34.88	35.05	35.01	34.36	35.63	34.75	35.25	34.85	35.70	35.29	35.29
Set 5	30.34	33.19	36.11	34.40	35.92	36.05	36.87	36.40	36.24	35.87	36.68	36.74	36.10	36.72
Set 6	33.66	33.46	35.85	36.71	36.93	37.10	36.32	37.67	36.52	37.94	36.99	37.14	37.31	37.24
Set 7	32.02	32.05	36.96	36.06	35.98	36.50	37.36	36.66	37.07	37.98	37.77	36.52	37.19	36.99
Set 8	32.76	32.67	35.59	38.15	37.52	36.83	37.22	35.94	37.26	37.39	37.57	37.33	37.45	37.63
Set 9	32.66	33.47	36.89	38.28	38.38	38.03	37.52	37.58	38.39	37.29	37.98	36.57	37.33	37.80
Set 10	31.89	33.74	37.10	37.11	37.10	37.20	37.71	37.95	38.17	37.41	37.49	38.05	37.72	38.43
Set 11	33.53	32.51	36.82	36.34	37.02	37.60	37.77	36.53	37.80	37.86	37.28	37.52	38.03	37.12
Set 12	33.24	33.90	34.86	37.35	37.94	38.08	36.27	37.40	37.72	37.34	37.37	37.06	37.43	37.99
Set 13	32.72	32.42	36.87	36.62	36.09	36.40	38.09	37.50	38.04	37.99	37.67	37.65	37.89	38.09
Set 14	32.85	33.47	36.64	37.00	36.80	36.68	38.00	37.28	37.02	37.87	38.52	37.53	38.58	38.34
Set 15	34.10	33.62	37.75	36.26	36.73	36.74	38.45	38.56	38.21	37.82	38.07	37.89	38.14	38.77
Set 16	34.30	32.76	37.71	37.65	38.00	37.85	38.29	37.49	38.42	37.72	37.80	38.35	38.59	38.26
Set 17	33.91	33.64	36.70	36.79	37.33	37.50	37.76	36.56	38.32	37.75	37.84	37.94	37.13	38.25
Set 18	33.42	33.18	37.38	35.86	36.57	37.11	37.07	38.22	37.39	37.91	37.10	37.92	38.11	37.60
Set 19	33.30	32.65	37.14	38.04	36.79	36.52	37.10	36.93	37.40	36.94	37.58	37.78	37.70	37.61
Set 20	33.10	33.49	36.64	36.96	37.21	37.11	36.31	36.64	36.88	37.01	37.03	37.33	37.22	37.03

Table D.4: UAR of the Support Vector Machine (SVM)-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{-5}	2^{1}	2^{-5}	2^{7}	2^{3}	2^{15}	2^{7}	2^{1}
γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-15}	2^{-3}	2^{-11}	2^{-13}	2^{-11}	2^{-15}	2^{-11}	2^{-5}
Set 1	25.02	25.02	25.02	25.61	25.71	27.48	26.66	30.09	27.84	28.68	28.96	29.07	29.02	29.49
Set 2	24.97	24.97	24.97	25.41	25.86	27.18	26.88	29.02	28.46	28.80	28.83	29.14	29.05	29.77
Set 3	25.00	25.00	25.00	24.97	25.94	26.90	27.74	28.73	28.88	29.13	29.20	29.34	29.25	30.25
Set 4	25.00	25.00	25.00	25.08	26.03	26.65	27.75	28.81	29.15	29.15	29.24	29.45	29.38	30.28
Set 5	25.00	25.00	25.00	25.10	26.10	26.71	27.97	28.79	29.14	29.23	29.19	29.66	29.60	30.35
Set 6	25.00	25.00	25.00	25.01	26.26	27.17	28.22	28.58	29.51	29.61	29.46	30.14	30.04	31.27
Set 7	25.00	25.00	25.00	24.99	25.92	26.92	28.37	28.49	29.82	29.78	29.68	30.29	30.21	31.28
Set 8	25.00	25.00	25.00	24.99	26.17	27.19	28.79	28.49	29.83	29.98	29.92	30.18	30.18	31.36
Set 9	25.00	25.00	25.00	25.00	25.84	27.32	29.06	27.94	29.98	29.99	29.91	30.24	30.29	31.62
Set 10	25.00	25.00	25.00	25.00	25.59	27.74	29.49	27.93	30.11	30.23	30.14	30.31	30.63	31.63
Set 11	25.00	25.00	25.00	25.00	25.46	27.62	29.75	27.97	30.19	30.31	30.21	30.57	30.61	31.60
Set 12	25.00	25.00	25.00	25.00	25.48	27.75	29.63	28.17	30.12	30.13	30.16	30.36	30.61	31.58
Set 13	25.00	25.00	25.00	25.00	25.14	27.47	30.03	27.39	30.37	30.45	30.66	30.82	30.95	31.51
Set 14	25.00	25.00	25.00	25.00	25.22	27.73	29.70	27.42	30.45	30.54	30.95	31.17	31.39	31.60
Set 15	25.00	25.00	25.00	25.00	25.24	27.78	30.48	27.41	30.37	31.00	30.99	31.14	31.23	31.55
Set 16	25.00	25.00	25.00	25.00	25.19	27.60	30.45	27.60	30.20	30.98	31.09	31.17	31.32	31.83
Set 17	25.00	25.00	25.00	25.00	25.17	28.35	30.23	27.26	30.62	31.41	31.61	31.40	31.61	31.79
Set 18	25.00	25.00	25.00	25.00	25.02	28.38	30.48	27.36	30.67	31.33	31.61	31.29	31.24	31.89
Set 19	25.00	25.00	25.00	25.00	25.01	28.38	30.35	27.19	30.48	31.38	31.72	31.38	31.35	31.91
Set 20	25.00	25.00	25.00	25.00	25.00	29.45	30.63	26.85	31.25	31.89	32.15	31.93	31.97	31.64

Table D.5: UAP of the SVM-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{-5}	2^{1}	2^{-5}	2^{7}	2^{3}	2^{15}	2^{7}	2^{1}
γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-15}	2^{-3}	2^{-11}	2^{-13}	2^{-11}	2^{-15}	2^{-11}	2^{-5}
Set 1	14.96	14.96	14.96	27.02	27.23	31.79	28.78	40.55	30.63	31.97	33.52	32.47	32.42	37.76
Set 2	11.83	11.83	11.83	22.73	28.22	33.71	29.14	37.23	31.13	30.28	30.04	32.25	30.94	35.89
Set 3	11.83	11.83	11.83	17.45	29.48	32.73	30.87	34.80	28.34	32.05	32.00	32.32	32.92	39.03
Set 4	11.83	11.83	11.83	16.70	30.61	30.97	30.76	35.42	31.62	31.16	32.17	33.25	33.50	39.13
Set 5	11.83	11.83	11.83	16.23	30.63	30.97	31.57	36.17	31.19	32.03	31.15	34.30	34.21	41.87
Set 6	11.83	11.83	11.83	14.95	31.57	33.34	34.52	37.72	33.27	35.05	33.34	37.62	39.08	45.31
Set 7	11.82	11.90	11.90	11.83	31.77	31.69	31.52	35.61	33.67	34.59	35.13	36.87	38.65	45.60
Set 8	11.83	11.83	11.82	11.83	30.35	32.59	32.56	37.84	33.99	36.88	34.94	36.42	37.90	46.17
Set 9	11.83	11.83	11.83	11.83	27.43	31.38	34.02	36.39	33.47	36.07	36.55	36.58	37.48	44.65
Set 10	11.83	11.83	11.83	11.83	23.99	33.70	33.36	36.27	36.62	38.99	38.86	39.12	40.82	43.10
Set 11	11.82	11.83	11.83	11.83	23.77	31.55	35.52	38.75	36.33	38.34	39.39	40.26	42.03	43.23
Set 12	11.90	11.83	11.83	11.83	24.75	31.80	34.06	36.99	35.15	36.74	38.50	38.95	39.54	42.75
Set 13	11.82	11.83	11.83	11.83	16.48	30.89	34.74	36.21	34.93	40.08	41.40	42.93	42.40	41.62
Set 14	11.90	11.83	11.83	11.83	17.38	32.73	33.05	34.70	36.75	41.79	44.85	44.98	46.11	42.73
Set 15	11.90	11.83	11.83	11.83	17.38	30.47	33.98	35.51	35.91	41.73	41.16	41.52	42.20	43.30
Set 16	11.83	11.83	11.83	11.83	17.17	31.78	32.95	36.17	35.21	42.85	43.91	42.11	45.03	44.18
Set 17	11.90	11.83	11.83	11.82	17.20	33.42	34.05	34.60	36.53	45.41	44.67	42.32	42.52	42.54
Set 18	11.90	11.83	11.83	11.80	14.51	36.18	34.54	34.13	39.05	46.44	45.93	42.13	44.18	42.76
Set 19	11.90	11.83	11.83	11.83	12.73	36.35	33.82	30.29	37.81	46.79	46.79	44.42	44.00	42.83
Set 20	11.90	11.83	11.83	11.80	11.83	37.83	36.23	27.96	42.42	46.77	46.61	46.83	46.33	42.86

Table D.6: F1-measure of the SVM-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	7	8	9	10	11	12	13	14
C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{-5}	2^{1}	2^{-5}	2^{7}	2^{3}	2^{15}	2^{7}	2^{1}
γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-15}	2^{-3}	2^{-11}	2^{-13}	2^{-11}	2^{-15}	2^{-11}	2^{-5}
Set 1	18.72	18.73	18.73	26.30	26.45	29.48	27.68	34.54	29.17	30.24	31.08	30.68	30.62	33.11
Set 2	16.05	16.05	16.06	24.00	26.99	30.10	27.97	32.62	29.73	29.52	29.42	30.62	29.97	32.54
Set 3	16.06	16.06	16.06	20.54	27.60	29.53	29.23	31.47	28.61	30.52	30.54	30.76	30.98	34.08
Set 4	16.06	16.06	16.06	20.05	28.14	28.65	29.18	31.78	30.34	30.12	30.64	31.24	31.31	34.14
Set 5	16.06	16.06	16.06	19.71	28.18	28.68	29.66	32.07	30.13	30.57	30.14	31.81	31.74	35.19
Set 6	16.06	16.06	16.06	18.71	28.67	29.94	31.05	32.52	31.28	32.10	31.28	33.47	33.97	37.00
Set 7	16.05	16.13	16.13	16.06	28.55	29.11	29.86	31.66	31.63	32.01	32.17	33.26	33.91	37.11
Set 8	16.06	16.06	16.05	16.06	28.10	29.65	30.56	32.51	31.77	33.07	32.24	33.01	33.60	37.35
Set 9	16.06	16.06	16.06	16.06	26.61	29.21	31.35	31.61	31.63	32.75	32.90	33.11	33.51	37.02
Set 10	16.06	16.06	16.06	16.06	24.76	30.43	31.31	31.56	33.05	34.05	33.95	34.16	35.00	36.48
Set 11	16.05	16.06	16.06	16.06	24.58	29.46	32.38	32.49	32.98	33.85	34.19	34.75	35.42	36.51
Set 12	16.13	16.06	16.06	16.06	25.11	29.64	31.69	31.98	32.44	33.11	33.82	34.12	34.50	36.32
Set 13	16.05	16.06	16.06	16.06	19.91	29.08	32.21	31.19	32.49	34.61	35.23	35.88	35.78	35.86
Set 14	16.13	16.06	16.06	16.06	20.58	30.03	31.29	30.63	33.31	35.29	36.63	36.82	37.35	36.33
Set 15	16.13	16.06	16.06	16.06	20.59	29.07	32.14	30.94	32.91	35.58	35.36	35.59	35.90	36.50
Set 16	16.06	16.06	16.06	16.06	20.42	29.54	31.65	31.31	32.52	35.96	36.41	35.82	36.94	37.00
Set 17	16.13	16.06	16.06	16.05	20.43	30.68	32.02	30.49	33.32	37.13	37.02	36.05	36.26	36.38
Set 18	16.13	16.06	16.06	16.03	18.37	31.81	32.38	30.37	34.36	37.42	37.45	35.91	36.60	36.53
Set 19	16.13	16.06	16.06	16.06	16.87	31.87	31.99	28.65	33.75	37.57	37.81	36.78	36.61	36.57
Set 20	16.13	16.06	16.06	16.03	16.06	33.12	33.19	27.39	35.99	37.92	38.05	37.97	37.83	36.41

Table D.7: UAR of the RF-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	2	1	3	4	12	8	13	9	10	6	7	11	5	14
numTrees [#]	10	10	61	134	298	215	1000	711	446	1000	516	898	555	894
numFeatures [%]	50	16	24	16	39	35	50	32	30	16	25	36	16	40
Set 1*	34.52	32.58	33.09	33.96	34.16	34.08	33.79	33.87	33.98	33.75	34.27	34.13	33.98	33.95
Set 2*	32.30	33.99	34.51	34.08	33.90	34.13	33.92	33.62	33.81	34.18	34.33	33.98	34.25	33.89
Set 3*	32.83	33.21	33.97	33.68	34.33	33.64	34.00	33.80	34.15	33.58	34.35	34.21	33.78	34.16
Set 4*	33.30	34.32	34.83	34.65	34.39	34.70	34.39	34.27	34.65	34.29	34.15	34.41	33.98	34.34
Set 5*	31.87	34.00	34.27	33.88	34.90	34.26	34.65	34.86	34.72	34.55	34.91	34.82	34.85	34.83
Set 6*	33.85	32.83	34.13	34.86	34.95	35.11	35.75	34.93	34.95	35.08	34.77	35.47	35.09	35.33
Set 7*	33.50	34.26	34.28	34.60	34.91	34.81	35.15	35.16	35.02	34.92	35.07	35.23	35.04	35.29
Set 8*	34.24	33.24	34.56	34.64	35.43	35.06	35.07	35.09	35.14	34.86	34.96	35.56	34.72	35.44
Set 9*	32.77	34.65	34.89	34.66	34.84	34.99	35.37	35.16	35.01	34.79	35.02	35.34	34.52	35.46
Set 10*	33.81	33.88	34.97	34.72	35.52	35.30	35.41	35.63	35.69	35.30	35.87	35.48	35.16	35.61
Set 11*	34.97	32.86	35.85	35.22	35.47	35.29	35.59	35.44	35.31	35.15	35.44	35.29	35.19	35.23
Set 12*	34.45	33.17	34.99	35.13	35.52	35.21	35.40	35.44	35.37	34.89	35.15	35.57	35.12	35.77
Set 13*	33.33	34.11	34.40	35.46	35.62	35.80	35.27	35.98	35.60	35.05	35.17	35.49	34.93	35.44

Table D.8: UAP of the RF-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	2	1	3	4	12	8	13	9	10	6	7	11	5	14
numTrees [#]	10	10	61	134	298	215	1000	711	446	1000	516	898	555	894
numFeatures [%]	50	16	24	16	39	35	50	32	30	16	25	36	16	40
Set 1*	37.13	34.87	41.46	46.00	44.46	44.75	44.53	44.78	45.21	44.27	44.81	45.08	46.65	45.15
Set 2*	34.58	36.86	45.63	44.81	43.64	45.57	44.52	44.80	44.55	46.57	46.49	43.92	47.46	43.76
Set 3*	35.78	35.60	42.51	43.46	44.39	44.23	44.94	44.55	44.28	44.96	45.54	45.53	45.37	45.53
Set 4*	35.48	36.91	43.77	46.63	45.02	46.28	45.13	45.82	47.35	47.10	45.70	45.89	46.40	45.45
Set 5*	33.88	36.78	44.41	45.74	46.91	45.05	47.80	48.03	47.13	48.89	48.14	48.05	48.91	49.02
Set 6*	36.44	35.60	44.12	46.24	48.74	49.12	50.84	49.24	50.73	50.09	49.63	50.63	50.49	50.31
Set 7*	35.56	37.29	46.13	48.56	49.16	49.51	50.12	50.87	49.24	51.18	50.64	51.76	51.20	50.66
Set 8*	36.68	35.54	45.85	49.32	50.50	49.63	49.97	49.73	51.10	51.01	49.70	50.78	50.27	51.55
Set 9*	36.22	37.01	46.33	49.97	47.95	49.22	49.81	50.24	48.79	50.56	49.15	48.74	50.12	50.88
Set 10*	35.75	36.13	47.46	49.64	51.47	48.83	51.14	52.01	51.58	54.20	52.67	52.49	53.77	52.23
Set 11*	37.03	34.72	47.24	50.06	51.58	52.22	51.60	52.79	51.41	51.33	51.59	51.96	51.47	52.59
Set 12*	36.56	35.53	48.67	50.35	50.54	51.27	49.18	49.71	50.98	52.72	52.59	50.18	51.53	52.86
Set 13*	36.36	36.97	44.76	51.82	50.20	52.73	51.71	52.11	52.25	51.64	52.41	51.10	52.54	51.28

Table D.9: F1-measure of the RF-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	2	1	3	4	12	8	13	9	10	6	7	11	5	14
numTrees [#]	10	10	61	134	298	215	1000	711	446	1000	516	898	555	894
numFeatures [%]	50	16	24	16	39	35	50	32	30	16	25	36	16	40
Set 1*	35.77	33.69	36.81	39.07	38.66	38.69	38.42	38.57	38.80	38.30	38.84	38.85	39.32	38.76
Set 2*	33.40	35.37	39.30	38.71	38.16	39.03	38.51	38.41	38.45	39.43	39.49	38.32	39.79	38.20
Set 3*	34.24	34.36	37.77	37.95	38.72	38.21	38.71	38.44	38.56	38.45	39.16	39.07	38.73	39.04
Set 4*	34.36	35.57	38.79	39.76	39.00	39.66	39.04	39.22	40.02	39.68	39.09	39.33	39.23	39.12
Set 5*	32.85	35.34	38.69	38.93	40.02	38.92	40.18	40.40	39.98	40.49	40.47	40.38	40.70	40.72
Set 6*	35.10	34.16	38.49	39.76	40.71	40.95	41.98	40.87	41.39	41.26	40.89	41.71	41.40	41.51
Set 7*	34.50	35.71	39.33	40.40	40.83	40.88	41.32	41.58	40.93	41.52	41.45	41.92	41.60	41.60
Set 8*	35.42	34.35	39.42	40.69	41.64	41.09	41.21	41.14	41.64	41.41	41.05	41.83	41.07	42.00
Set 9*	34.41	35.79	39.80	40.94	40.35	40.90	41.36	41.37	40.77	41.22	40.90	40.97	40.88	41.80
Set 10*	34.76	34.97	40.27	40.86	42.03	40.98	41.85	42.29	42.19	42.76	42.68	42.34	42.51	42.35
Set 11*	35.97	33.76	40.76	41.35	42.04	42.11	42.12	42.41	41.86	41.73	42.01	42.03	41.80	42.19
Set 12*	35.47	34.31	40.71	41.38	41.72	41.75	41.17	41.38	41.77	41.99	42.14	41.63	41.77	42.67
Set 13*	34.78	35.48	38.90	42.12	41.67	42.64	41.94	42.57	42.35	41.76	42.10	41.89	41.96	41.91

Table D.10: UAR of the SVM-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	8	7	9	12	10	14	13	11
C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{1}	2^{-5}	2^{-5}	2^{15}	2^{7}	2^{1}	2^{7}	2^{3}
γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-3}	2^{-15}	2^{-11}	2^{-15}	2^{-13}	2^{-5}	2^{-11}	2^{-11}
Set 1*	24.95	24.95	24.95	25.21	26.09	29.76	31.92	31.89	31.02	31.54	30.87	32.45	31.70	31.16
Set 2*	24.99	24.99	24.99	25.33	26.09	29.46	32.46	31.98	31.62	32.24	31.84	33.22	32.21	31.65
Set 3*	25.01	25.01	25.02	25.02	26.04	29.16	32.05	31.55	31.46	32.19	31.57	33.27	32.11	31.56
Set 4*	25.00	25.00	25.00	24.96	26.48	27.92	31.04	31.48	31.92	32.13	31.88	33.22	32.34	31.90
Set 5*	25.00	25.00	25.00	24.94	25.69	28.13	29.27	32.47	32.12	33.14	33.27	34.60	33.56	32.74
Set 6*	25.00	25.00	25.00	24.94	25.48	29.79	29.87	32.38	31.98	33.95	33.26	34.72	34.06	33.29
Set 7*	25.00	25.00	25.00	24.99	25.46	28.36	28.99	32.16	32.03	33.89	33.29	34.77	34.39	33.59
Set 8*	25.00	25.00	25.00	25.00	25.13	28.87	29.23	32.10	32.29	34.47	34.17	35.31	34.72	34.02
Set 9*	25.00	25.00	25.00	25.00	25.12	29.55	28.36	32.80	33.00	34.27	33.90	35.31	34.61	34.09
Set 10*	25.00	25.00	25.00	25.00	25.08	29.95	28.29	32.59	33.12	33.92	33.88	34.58	34.11	34.10
Set 11*	25.00	25.00	25.00	25.00	25.00	30.53	27.58	32.90	33.30	34.30	34.33	34.21	34.38	34.91
Set 12*	25.00	25.00	25.00	25.00	25.00	32.01	26.87	33.05	33.93	34.68	35.17	33.45	35.32	35.49
Set 13*	25.00	25.00	25.00	25.00	25.00	32.57	26.20	33.49	34.75	35.69	35.90	32.87	36.33	36.20

Table D.11: UAP of the SVM-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	8	7	9	12	10	14	13	11
C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{1}	2^{-5}	2^{-5}	2^{15}	2^{7}	2^{1}	2^{7}	2^{3}
γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-3}	2^{-15}	2^{-11}	2^{-15}	2^{-13}	2^{-5}	2^{-11}	2^{-11}
Set 1*	11.24	11.24	11.24	23.41	20.96	37.06	41.62	36.65	40.34	38.83	38.92	43.32	40.60	39.75
Set 2*	11.25	11.25	11.25	23.31	22.58	36.07	47.06	38.66	41.14	39.66	40.87	43.03	38.30	40.34
Set 3*	12.72	12.72	12.72	17.83	28.05	35.58	44.26	38.50	40.15	39.24	40.11	45.38	40.53	40.51
Set 4*	11.25	11.25	11.25	15.22	32.23	35.32	43.44	39.78	40.58	43.43	44.69	48.00	44.65	43.99
Set 5*	11.25	11.25	11.25	11.24	26.88	34.67	38.40	40.49	42.54	46.51	48.03	48.10	49.48	49.20
Set 6*	11.25	11.25	11.25	11.24	24.34	35.22	37.87	39.31	39.64	49.42	49.89	49.24	50.32	49.92
Set 7*	11.25	11.25	11.25	11.25	26.69	35.04	37.67	38.91	38.75	51.16	50.54	48.87	50.85	50.96
Set 8*	11.25	11.25	11.25	11.25	19.19	37.70	36.57	40.48	41.34	50.74	50.57	47.56	51.42	51.61
Set 9*	11.25	11.25	11.25	11.25	16.88	38.95	35.64	40.40	43.22	49.28	49.42	46.99	48.30	49.91
Set 10*	11.25	11.25	11.25	11.25	16.04	41.53	36.07	39.87	46.61	47.52	49.75	46.35	48.42	52.17
Set 11*	11.25	11.25	11.25	11.25	11.25	35.33	32.52	41.02	45.64	48.16	48.36	43.52	47.10	51.36
Set 12*	11.25	11.25	11.25	11.25	11.25	40.27	34.50	41.23	49.50	45.65	49.82	43.48	47.47	50.28
Set 13*	11.25	11.25	11.25	11.25	11.25	41.62	27.57	45.30	49.42	47.55	49.37	40.12	47.80	49.64

Table D.12: F1-measure of the SVM-classifier for the evaluated feature sets and parameter sets in descending order of γ

Par. Set	1	2	3	4	5	6	8	7	9	12	10	14	13	11
C	2^{-5}	2^{-3}	2^{15}	2^{11}	2^{3}	2^{13}	2^{1}	2^{-5}	2^{-5}	2^{15}	2^{7}	2^{1}	2^{7}	2^{3}
γ	2^{3}	2^{3}	2^{3}	2^{1}	2^{-1}	2^{-7}	2^{-3}	2^{-15}	2^{-11}	2^{-15}	2^{-13}	2^{-5}	2^{-11}	2^{-11}
Set 1*	15.50	15.50	15.50	24.28	23.25	33.01	36.13	34.11	35.07	34.81	34.43	37.10	35.60	34.94
Set 2*	15.52	15.52	15.51	24.28	24.21	32.43	38.42	35.01	35.76	35.57	35.79	37.49	34.99	35.47
Set 3*	16.87	16.87	16.87	20.82	27.01	32.05	37.18	34.68	35.28	35.37	35.33	38.39	35.83	35.48
Set 4*	15.52	15.52	15.52	18.91	29.07	31.19	36.20	35.15	35.74	36.94	37.21	39.27	37.51	36.98
Set 5*	15.52	15.52	15.52	15.50	26.27	31.06	33.22	36.04	36.60	38.71	39.31	40.25	39.99	39.32
Set 6*	15.52	15.52	15.52	15.50	24.90	31.68	33.40	35.51	35.40	40.25	39.91	40.72	40.62	39.95
Set 7*	15.52	15.52	15.52	15.51	26.06	31.35	32.76	35.21	35.07	40.77	40.14	40.63	41.03	40.49
Set 8*	15.52	15.52	15.52	15.52	21.76	32.69	32.49	35.81	36.26	41.05	40.78	40.53	41.45	41.01
Set 9*	15.52	15.52	15.52	15.52	20.19	33.61	31.59	36.21	35.43	40.43	40.21	40.32	40.32	40.51
Set 10*	15.52	15.52	15.52	15.52	19.57	34.80	31.71	35.87	38.72	39.58	40.31	39.61	40.02	41.25
Set 11*	15.52	15.52	15.52	15.52	15.52	32.75	29.85	36.52	38.50	40.07	40.16	38.31	39.75	41.57
Set 12*	15.52	15.52	15.52	15.52	15.52	35.67	30.21	36.69	40.26	39.42	41.24	37.81	40.50	41.61
Set 13*	15.52	15.52	15.52	15.52	15.52	36.54	26.87	38.51	40.81	40.77	41.57	36.14	41.28	41.87

Appendix E

Real-World Validation

The recognition results presented in this Thesis are all based on empirical research. However, when it comes to a later real-world application further experiments need to be performed to reassure the performance of the system also in unknown real-world scenarios (i.e. field research). This can, for example, be done by performing large scaled field studies where the system is evaluated in real driving situations.

A first small-scaled field study, where frustration was induced to the driver in a similar way as for the data collection in Chapter 3, was performed in the final phase of this Thesis. This was done by prototypically implementing the classifier on an in-car processing unit and communicating the output stream as JSON-message[1] to a central computer via MQTT[2]. The output of the classifier contained a continuous stream of JSON-messages consisting of speech analysis windows of 2 seconds length. These messages also included segments where no speech was present for the analysis of the emotional state. However, the results lack of evidence as only a very small number of eight subjects took part in the study and a ground truth was not assessed. The only available data used to evaluate the algorithm were the output stream of the classifier, field notes of an experimenter, self-reports of the driver and the knowledge of the experimental phase of the test drive. The results, presented as number of speech segments classified by the audio-based emotion algorithms, are stated in Table E.1. The Table also includes the self-reported frustration level on a 10-point Lickert-scale for each subject averaged over the whole test scenario. Bold numbers indicate the number of segments classified as the target state of the driving scenario.

As stated in Chapter 6, the implemented classifier is strongly biased towards detection of a neutral emotional state. A confusion with an emotional state different from neutral is very unlikely. This also explains the high number of detected neutral speech segments for all subjects. The second highest number of detections is, nevertheless, obtained for the state of frustrated driving. A repeated-measures ANOVA revealed that there exists a significant difference between the number of speech samples classified as a certain emotional state (main effect: F(1.6,11.0) = 66.7, p

[1] JavaScript Object Notation; Data communication format

[2] Message Queuing Telemetry Transport; Network protocol for data communication between devices

Table E.1: Number of speech segments classified by the Audio-based emotion classifier, including audio segments where no speech was present, presented per evaluated subject.

Subject No./ Predicted State	1	2	3	4	5	6	7	8	$\sum$
No speech	40	26	11	27	25	19	29	18	195
Neutral	75	67	72	64	48	61	37	47	471
Positive	0	4	1	1	3	0	0	22	31
Frustrated	**2**	**4**	**15**	**0**	**23**	**1**	**10**	**15**	**70**
Anxious	0	0	0	3	0	3	1	0	7
Self-report	1.8	6.25	1.75	6.8	8.25	5.75	6.8	8.25	/

< 0.01, Greenhouse-Geisser-corrected). Nevertheless, a post-hoc one-sided paired t-test revealed that the only significant difference was between the number of samples classified as the neutral state compared to all other emotional states ($p < 0.05$, Bonferroni-corrected). It was further noticed that in most cases a high self-reported frustration level is related to high numbers of detected speech samples classified as frustration. One exception was seen for participant No. 3 for whom a large number of speech samples was detected as frustration while the self-report indicated a very low level of frustration. Another discrepancy was noticed for participant No. 8 for whom a considerably higher number of speech samples classified as positive was present compared to all other participants. To investigate the high number of samples classified as frustration and positive for participant 3 and 8, respectively, a more detailed insight into the conducted evaluations is necessary.

From these results a vague positive tendency on the performance of the audio-based classifier, with focus on detecting frustrated drivers, can be drawn. The results, however, lack of statistical evidence as the sample set does not represent the population in a sufficient way.

Declaration of Honor

I hereby declare that I produced this thesis without prohibited external assistance and that none other than the listed references and tools have been used. I did not make use of any commercial consultant concerning graduation. A third party did not receive any nonmonetary perquisites neither directly nor indirectly for activities which are connected with the contents of the presented thesis.

All sources of information are clearly marked, including my own publications.

In particular I have not consciously:

- Fabricated data or rejected undesired results
- Misused statistical methods with the aim of drawing other conclusions than those warranted by the available data
- Plagiarized data or publications
- Presented the results of other researchers in a distorted way

I do know that violations of copyright may lead to injunction and damage claims of the author and also to prosecution by the law enforcement authorities. I hereby agree that the thesis may need to be reviewed with an electronic data processing for plagiarism.

This work has not yet been submitted as a doctoral thesis in the same or similar form in Germany or in any other country. It has not yet been published as a whole.