INSTITUT FÜR INFORMATIONS- UND KOMMUNIKATIONSTECHNIK (IIKT)

Accessing the Interlocutor

Recognition of Interaction-related Interlocutor States in Multiple Modalities

DISSERTATION

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

von
Olga Egorow, M.Sc. geb. am 14.04.1989 in Tomsk

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

Gutachter: Prof. Dr. rer. nat. Andreas Wendemuth
Prof. Dr.-Ing. Christian Diedrich
Prof. Dr.-Ing. Stefan Kopp

Promotionskolloquium am 03.07.2020

The Arcadian dream all fallen through
But the Albion sails on course
– Carl Barât & Peter Doherty

Abstract

The research in the field of human-computer interaction aims at enabling technical systems to interact with humans in the same way that humans do among themselves. One aspect of natural human interaction is implicitly communicating the internal state, such as the current emotions, using voice, gestures and facial expressions. Gaining access to this information is one of the central topics addressed in affective computing.

This thesis focuses on the automatic recognition of three internal interlocutor states highly relevant for the domain of human-computer interaction – namely *trouble*, *satisfaction* and *cooperativeness* – using different interlocutor signals, such as speech or acoustic signals, physiological signals and spatial upper-body movements. Three existing corpora of interaction data provide the empirical base for the investigations.

The aim of the thesis is to enhance the understanding of interaction-related interlocutor states by developing approaches for their automatic recognition. Furthermore, this thesis contributes to the current state of the art by discussing three methodological challenges: finding appropriate data and developing general data requirements, selecting appropriate modalities and features, and implementing appropriate classification and performance evaluation methods.

As a main objective, three recognition tasks were accomplished: the recognition of *trouble*, *satisfaction*, and *cooperativeness*. For these tasks, existing machine learning techniques were applied: random forests, support vector machines and naïve Bayes classification. All three tasks were performed as binary classification tasks. The evaluation of all three classification approaches was done in a subject-independent way to ensure the generalisation ability of the classifiers.

The conducted research leads to the conclusion that the three investigated interlocutor states can be accessed using features obtained from the considered behavioural signals. Depending on data and setting, the recognition accuracy varies between 64% and 87% f-measure. The physiological signals provided the best recognition results, but it can be argued that for certain applications, especially when other signals are not available, speech enables sufficient recognition performance to create systems adapting to their users' current states. In order to further improve the ability of technical systems to access these states, it is necessary to expand the current understanding of both, the expression of human interaction behaviour and its processing.

Zusammenfassung

Die Forschung auf dem Gebiet der Mensch-Computer-Interaktion hat das Ziel, Systeme zu entwickeln, die mit Menschen auf die gleiche Art interagieren können, wie Menschen es untereinander tun. Ein Aspekt von natürlicher menschlichen Interaktion ist die implizite Vermittlung des inneren Zustandes, beispielsweise der Emotionen, mit Hilfe von Stimme, Gestik, Mimik, etc. Der Zugang zu diesen Informationen ist eines der zentralen Themen von Affective Computing.

Diese Arbeit konzentriert sich auf die automatische Erkennung von drei Gesprächspartner-Zuständen, die für die Mensch-Computer-Interaktion von großer Bedeutung sind – nämlich *Anstrengung*, *Zufriedenheit* und *Kooperativität*. Dabei werden unterschiedliche Gesprächspartner-Signale benutzt, wie Sprache oder akustische Signale, physiologische Signale und Bewegungen des Oberkörpers. Drei bestehende Korpora liefern die empirische Grundlage für diese Untersuchungen.

Das Ziel der Arbeit ist es, das Verständnis von interaktionsrelevanten Gesprächspartner-Zuständen durch Entwicklung von Ansätzen zu ihrer automatischen Erkennung zu verbessern. Weiterhin trägt diese Arbeit zum aktuellen Stand der Wissenschaft in drei methodischen Herausforderungen bei: die Suche nach geeigneten Daten und die Entwicklung von allgemeinen Datenanforderungen, die Auswahl von geeigneten Modalitäten und Merkmalen und die Implementierung von geeigneten Klassifikations- und Evaluationsmethoden.

Zur Erreichung des Ziels wurden drei Erkennungsaufgaben bewerkstelligt: die Erkennung von *Anstrengung*, *Zufriedenheit* und *Kooperativität*. Dabei wurden existierende Methoden des maschinellen Lernens angewandt: Random Forests, Support Vector Machines und Naïve Bayes Klassifikation. Alle drei Aufgaben wurden als binäre Klassifikationsaufgaben aufgefasst. Die Evaluierung aller drei Klassifikationsansätze erfolgte personenunabhängig, um die Generalisierungsfähigkeit der Klassifikatoren zu garantieren.

Die durchgeführte Forschungsarbeit lässt den Schluss zu, dass die untersuchten Gesprächspartner-Zustände mit Hilfe der aus den betrachteten Verhaltenssignalen extrahierten Merkmale erkannt werden können. Dabei variiert die Erkennungsgenauigkeit in Abhängigkeit von Daten und Setting zwischen 64% und 87% F-Measure. Die physiologischen Signale liefern die besten Erkennungsergebnisse, jedoch kann argumentiert werden, dass für bestimmte Anwendungen, insbesondere bei Nichtverfügbarkeit von anderen Signalen, Sprache eine ausreichende Erkennungsleistung ermöglicht, um Systeme zu entwickeln, die sich auf den aktuellen Nutzerzustand einstellen können. Um die Fähigkeit von technischen Systemen zu verbessern, diesen Zustand zu erfassen, ist

es notwendig, das derzeitige Verständnis sowohl von Ausdruck menschlichen Verhaltens als auch von dessen Verarbeitung auszubauen.

List of Authored Publications

1. Contributions to international peer-reviewed journals:

   Egorow, O. & Wendemuth, A. (2019). "On Emotions as Features for Speech Overlaps Classification". To appear in *IEEE Transactions on Affective Computing*.

2. Contributions to national peer-reviewed journals:

   Egorow, O.; Siegert, I. & Wendemuth, A. (2017). "Prediction of User Satisfaction in Naturalistic Human-computer Interaction". *Kognitive Systeme* 2017.1, s.p.

3. Contributions to peer-reviewed conference proceedings:

   Böck, R.; **Egorow, O.** ; Siegert, I. & Wendemuth, A. (2017). "Comparative Study on Normalisation in Emotion Recognition from Speech". In: *Intelligent Human Computer Interaction*. Cham: Springer International Publishing, pp. 189–201.

   Böck, R.; **Egorow, O.** & Wendemuth, A. (2017). "Speaker-group Specific Acoustic Differences in Consecutive Stages of Spoken Interaction". In: *Elektronische Sprachsignalverarbeitung 2017: Tagungsband der 28. Konferenz*, Dresden: TUDpress, pp. 211–218.

   Böck, R.; **Egorow, O.** & Wendemuth, A. (2018). "Acoustic Detection of Consecutive Stages of Spoken Interaction Based on Speaker-group Specific Features". In: *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*, Dresden: TUDpress, pp. 252–254.

   Egorow, O. & Wendemuth, A. (2016). "Detection of Challenging Dialogue Stages Using Acoustic Signals and Biosignals". In: *Proc. of the 24th Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'16)*, Plzen: Vaclav Skala – Union Agency, pp. 137–143.

   Egorow, O. & Wendemuth, A. (2017). "Emotional Features for Speech Overlaps Classification". In: *Proc. of the Interspeech-2017*. International Speech Communication Association, pp. 2356–2360.

   Egorow, O.; Siegert, I. & Wendemuth, A. (2018). "Improving Emotion Recognition Performance by Random-forest-based Feature Selection". In: *Speech and Computer*. Cham: Springer International Publishing, pp. 134–144.

Egorow, O.; Mrech, T.; Weisskirchen, N. & Wendemuth, A. (2019). "Employing Bottleneck and Convolutional Features for Speech-Based Physical Load Detection on Limited Data Amounts". In: *Proc. of the Interspeech-2019*. International Speech Communication Association, pp. 1666–1670.

4. Contributions to book chapters:

   Böck, R.; **Egorow, O.**; Höbel-Müller, J.; Requardt, A. F.; Siegert, I. & Wendemuth, A. (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. Cham: Springer International Publishing, pp. 203–233.

Contents

List of Figures

List of Tables

List of Abbreviations

CNN	Convolutional Neural Network
DC	Davero Corpus
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyogram
FN	False Negative
FP	False Positive
HCI	Human-Computer Interaction
HHI	Human-Human Interaction
LLD	Low-Level Descriptor
LMC	Last Minute Corpus
LOSO	Leave-One-Subject-Out
LSP	Linear Spectral Pair
LSTM	Long-Short Term Memory
MFCC	Mel-Frequency Cepstral Coefficient
NB	Naïve Bayes
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Network
RSP	Respiration
SC	Skin Conductivity
SVM	Support Vector Machine
TDT	Train-Dev-Test
TN	True Negative
TP	True Positive
UAF	Unweighted Average F-Measure
UAP	Unweighted Average Precision
UAR	Unweighted Average Recall

VAD Voice Activity Detection

WoZ Wizard-of-Oz

Chapter 1

Introduction

Humans speak volumes without words. A smile, a shrug, rolling eyes – these unspoken cues guide our everyday interactions and relationships with others. For machines to be the companions and helpers of our dreams, they would need to use and understand the same social cues that we do. This remains a huge challenge, as human facial expression and body language are often subtle, their meaning based on context and subject to interpretation.

Wall text, *Robots – making machines human*, Tekniska Museet, Stockholm.

Contents

NATURAL communication between humans is influenced by different factors besides the actual content of the message being sent and perceived. The influence can be rather "static" such as sex and age as well as personal traits, but it can also depend on the current situation – the social context and the current affective state of the interlocutors. These factors are not only present in Human-Human Interaction (HHI), they also shape and change Human-Computer Interaction (HCI).

This thesis is dedicated to the topic of accessing the internal state of an interlocutor. In particular, we will engage ourselves with three types of interactional circumstances that will be described in detail further below: challenging interaction, satisfying interaction and cooperative interaction. For this, we will rely on different modalities – in most cases, we will analyse the interlocutor's speech, but also physiological signals and bodily movements.

But before presenting the investigated research questions, we should discuss why the internal states of the participants of an interaction are an important issue to be investigated. In this chapter, Section 1.1 presents the motivation behind the thesis, Section 1.2 describes the current state of the relevant research field and points out the problems to be solved, Section 1.3 states the added value of the presented research, Section 1.4 defines the aim that shall be achieved, and finally, Section 1.5 presents an overview of the whole thesis.

1.1 Accessing the Interaction Experience

When we speak about communication of messages, there are two channels to be considered: The first channel is used for explicit messages, the second for implicit ones [Cowie et al. 2001]. As everyone knows, an interlocutor telling that "it is raining" is often not only pointing out a fact, she might also show how this information affects her, e.g. whether she likes it or not and what she wants to do about it. The same words can be used to convey a multitude of messages depending on the intonation, context or body language. These relations between the form and the meaning of language are long known and well researched, most prominently by communication theorists Watzlawick and Schulz von Thun [Watzlawick 1964; Watzlawick et al. 2011; Schulz von Thun 2013].

This is especially true for interactions between humans, or HHI. But can this be taken for granted for interactions between humans and technical systems, or HCI? In HHI, humans interact exclusively with other humans, presuming that all interlocutors are able (at least to some extent) to read both the explicit and the implicit interaction channels. In HCI, humans interact with a system while supposing that the system is operated in an automatic way. The important question is now whether the interlocutors also presume that the system can understand the implicit channel – and whether they are able and willing to deliberately stop using this channel if it is not the case. At the same time, computers are seen as social actors, rendering HCI "fundamentally social", where the human interaction partners show "a wide range of social behaviours" such as politeness norms, the notions of *self* and *others* and gender stereotypes [Nass et al. 1994]. This means that HCI is indeed very similar to HHI. Aiming at this similarity, the field of *affective computing*

believes that emotions and other affective states are essential to human cognition and perception, being linked to memory, perception, learning and even decision making [Sloman 1987; Damasio 1994; Picard 1997]. Therefore, accepting affective states as a key feature of cognition, affective computers could provide better performance in assisting humans. The recognition of affective states might be an important step towards making the interaction experience as pleasant and effective as possible – and in this way, also as similar to HHI as possible.

Affective states comprise a vast variety of behaviour and experience patterns, and therefore, they can be described differently depending on the individual case. An affective state or an emotion can even be defined in a pervasive way, leading to the definition of emotion as "whatever is present in most of life, but absent when people are emotionless" [Cowie et al. 2011]. As a starting point, we can look at the six basic emotions that were introduced by Ekman: These emotions are thought to be pan-cultural since they lead to the same facial behaviours in every investigated culture, namely happiness, sadness, anger, fear, surprise, and disgust [Ekman 1970]. Since the original investigation, this list has been further expanded leading to the development of the "affective space", trying to capture all emotions known from everyday life. The "affective space" can be seen as a three-dimensional space, with the dimensions of pleasure or valence (positive or negative), arousal or activation (high or low), and dominance or control (high or low) [Mehrabian 1996]. This space can be further enriched to four dimensions by adding unpredictability [Fontaine et al. 2007] or even seen as an "hourglass" with four concomitant dimensions [Cambria et al. 2012]. This hourglass of emotions was developed especially in the context of HCI to measure to what degree the user of a HCI system is amused, interested, comfortable and confident, resulting in the four dimensions of pleasantness, attention, sensitivity and aptitude, respectively.

In this thesis, we focus on affective states especially relevant in the context of interaction. Obviously, we cannot investigate all of them in the scope of a thesis, and therefore we will limit our considerations to three distinctive internal states. The first state we will refer to as *trouble*. By that we mean the state of an interlocutor experiencing a mismatch between her expectations and the real course of the interaction. This affective state is defined by a complex mix of dissatisfaction, irritation and frustration occurring when the interaction becomes challenging. The second state we want to investigate is the state of *satisfaction*. In this state, the interlocutor is content and at ease with the current situation. The last state we will encounter in this thesis is the state of *cooperativeness* or agreeableness. It is defined by the interlocutor acting in a sympathetic and considerate way. We will further elaborate on these states and their definitions in Section 2.1.1.

But why these states? The answer for this question can be found when looking at the desirable qualities of technical systems used in everyday life. In order to make the interaction with a computer more similar to interactions with other humans, we need systems which adapt themselves to our requests, anticipate our behaviour and provide tailored solutions for our problems – just as a good personal assistant or *companion* would [Wendemuth & Biundo 2012]. Especially in times of an ageing population and a shortage of skilled healthcare workers, technical systems attending to elderly people can be a solution. In this case, such systems should be able to detect trouble in order to prevent it, to ensure the user's satisfactory experience as well as cooperation. Of course, there are other abilities necessary for a good assistant, but the investigation of these three states is a good starting point towards a truly adaptive assistive system.

The next section provides a general overview of the state of the art in the field of internal state modelling and recognition in HCI, allowing us to analyse where the current research in this field is heading. A more detailed overview of existing works regarding the states investigated in this thesis will be provided in the chapters dedicated to the respective states.

1.2 Internal State Recognition in Human-Computer Interaction

The field of HCI is based on the idea that "rather than just using machines, we interact with them" [Suchman 1987], which results in the necessity to establish mutual intelligibility. In other words, the link between observable behaviour and the underlying mental processes must be made clear [Suchman 1987]. This means that, somehow, the technical systems must learn to infer the internal affective or cognitive states of their interaction partners from the signals they perceive from them.

In HHI, humans communicate naturally, through "gestures, expressions, movements, and discover the world by looking around and manipulating physical stuff" [Valli 2008] – if we want to make HCI more natural, we need systems that can be interacted with as humans are "used to interact with the real world in everyday life" [Valli 2008]. This means that the systems must be able to process unscripted, unforced, not acted and possibly multimodal interaction with their users [Valli 2008]. Again, accessing the internal state of the participants of such an interaction is a crucial part in this endeavour.

The task of internal state recognition can be described as the mapping of the behavioural cues of the interaction participants (i.e. interlocutors) to their inner experience. The behavioural cues, for instance speech, gestures or facial

expressions, are seen as signals emitted by the underlying processes, i.e. the true state of the interlocutor or the ground truth. This point of view allows for the application of methods from signal processing and pattern recognition to solve the task at hand. In order to access the information on the interlocutor state, specific features are extracted from the interlocutor signals as content-carrying units, as is standard procedure in signal processing. Here, two possibilities must be distinguished, namely uninformed information extraction (e.g. via filters) or an information extraction based on a generation model (i.e. the model of the underlying process generating the signal). Analogously, a recognition model is developed using empirical data (i.e. exemplary interactions) applying methods from pattern recognition. It is assumed that there is a mapping between certain patterns in the extracted features that can be recognised by the recognition model, and the underlying processes described by the generation model. The recognition model is then used to "recognise" or classify the interlocutor state based on the features extracted from the interlocutor signal. The result of this process is the probability of the interlocutor being in a certain state given the observed realisation of the features. This recognition process is called *classification* and will be explained in more detail in Section 2.2.3.

In this section, we will briefly review the current state of the art in the field of affective state recognition, beginning with recent development in this field and highlighting two challenges, namely the search for consistent feature sets and evaluation methods.

1.2.1 General Development of Internal State Recognition

Affective computing has been concerned with the internal state of humans in general, and specifically computer users, for around thirty years now, since the coining of the term *affective computing* itself [Picard 1997]. The aim of affective computing is to provide technical systems with human-like abilities of "observation, interpretation and generation of affect" to implement more "harmonious" HCI [Tao & Tan 2005]. One fundamental challenge on the way towards such technical system is the recognition of affective states. The progress in this field is eminent: Starting with the recognition of acted emotions under highly controlled conditions [Schuller et al. 2003; Schuller et al. 2007], the current research is aiming at improving the recognition rates in "in the wild" real-life experiments [Truong et al. 2012; Kim et al. 2017]. The importance of systems applicable to real-life scenarios is reflected, for example, in the annual Emotion Recognition in the Wild (EmotiW) Challenge with its latest

seventh instalment in 2019 focussing on audio-visual emotion recognition in unconstrained conditions[1].

Another direction of development in the field of affective computing is the shift from "classical" machine learning methods such as Support Vector Machine (SVM) and Random Forest (RF) to deep learning modelling techniques such as Convolutional Neural Network (CNN), Long-Short Term Memory (LSTM) and Recurrent Neural Network (RNN). In the early days of speech-based emotion recognition, SVM seemed to be the means of choice, for speech [Kwon et al. 2003; Grimm et al. 2007; Chandaka et al. 2009], facial emotions [Michel & El Kaliouby 2003; Lozano-Monasor et al. 2014], as well as biophysiological data [Takahashi 2004; Wang et al. 2011]. Nowadays, more and more researchers employ CNN and LSTM for the classification of emotions, since these techniques enabled a massive boost in performance for most pattern recognition tasks [Arel et al. 2010]. Such deep architectures are applied throughout different fields of emotion recognition, for instance, from speech [Lim et al. 2016], video data [Fan et al. 2016] and biophysiological data [Alhagry et al. 2017]. However, the application of deep learning is not always possible due to a lack of reliably annotated data and interindividual differences in expressing emotions – this development leads to a discussion on the future of big data use in this field [Böck et al. 2019]. So far, at least for speech data, the performance of both CNNs and LSTMs needs improvement [Kurpukdee et al. 2017; Etienne et al. 2018a]. Another problem concerning such systems is their high complexity and low explanatory power regarding the modelled phenomena [Gilpin et al. 2018].

1.2.2 The Quest for Suitable Features

One especially crucial point in speech-based emotion recognition is the fact that emotions are "piggybacking" on top of the actual speech signal: Classifying emotions always means disentangling the variations induced by the emotional content from the variations induced by the semantic content of the processed speech. The same is true for facial expressions of emotions: In order to recognise an angry glance, the computer system has to first recognise the facial features that can be different depending on the a multitude of factors, from individual differences to lighting conditions. That is the reason emotion recognition is a difficult task: We need to find the necessary information that is "hidden" in the signal. Therefore, the question of finding the most informative parts of the signal is so important. This explains the multitude of modalities and feature sets investigated in this field. A well-chosen feature set conveys all discrimination-relevant aspects of the signal: In the case of the speech sig-

[1] `https://sites.google.com/view/emotiw2019/home`, retrieved December 15, 2019.

nal, certain features such as the Mel-Frequency Cepstral Coefficients (MFCCs) have been shown to be a good choice for a variety of speech-related tasks, for example speech and speaker as well as emotion recognition [Zhen et al. 2000; Kwon et al. 2003]. But they are also used for other auditory tasks, such as the recognition of the music genre [Li et al. 2003]. This can be seen as a proof that they contain meaningful information for different tasks.

For the modality of speech, there is an ongoing discussion on the "most important" features, leading to a variety of investigations on prosodic and spectral features [Luengo et al. 2005; Bitouk et al. 2010; Li et al. 2015] as well as a comparison of different feature sets [Vogt & Andre 2005]. The relevance of this issue is highlighted by well established annual challenges such as the Computational Paralinguistics Challenge ComParE[2]. Besides introducing new feature sets, e.g. based on wavelets, spectrograms, saliency and prominence as well as "audio words" [Kishore & Satish 2013; Badshah et al. 2017; Mao et al. 2014; Jing et al. 2018; Schmitt et al. 2016], there is also a significant amount of research done on feature selection [Ververidis & Kotropoulos 2008; Rong et al. 2009; Kim et al. 2013; Trabelsi & Bouhlel 2015; Liu et al. 2018]. However, recently this discussion also involves end-to-end approaches, where instead of extracting hand-crafted features, the process of feature extraction is left to deep learning [Trigeorgis et al. 2016].

Other modalities struggle with the question of finding the most suitable features, too. Besides the audio signal, in this thesis we will encounter physiological signals and three-dimensional (3D) data of bodily motions. For the physiological domain, we will focus on Electrocardiogram (ECG), Electroencephalogram (EEG), Electromyogram (EMG), Respiration (RSP), Skin Conductivity (SC). Some of these signals are more difficult to obtain than others (e.g. EEG that requires a high number of gel-based electrodes versus SC measurements with a single sensor). There is a growing number on investigations including the mentioned as well as other physiological signals, such as cardiorespiratory activity alone or in combination with SC [Rainville et al. 2006; Yannakakis & Hallam 2008], EEG [Kothe et al. 2013; Velchev et al. 2016], EMG and ECG alone [Naji et al. 2015; Shin et al. 2017] and in combination with SC and body temperature [Katsis et al. 2008; Canento et al. 2011], etc. Since there are various physiological signals that can be used, there is no general set of features. A generally accepted tool for processing data obtained from EMG, ECG, SC and RSP changes is the Augsburg Biosignal Toolbox (AuBT) [Wagner et al. 2005]. More recent tools are the Toolbox for Emotional feAture Extraction from Physiological signals (TEAP) enabling the processing of the EEG, galvanic skin response, EMG, skin temperature, RSP pattern and blood

[2] `http://www.compare.openaudio.eu/`, retrieved December 15, 2019.

volume pulse [Soleymani et al. 2017], and PhysioLab for processing ECG and EMG signals, as well as electrodermal activity [Muñoz et al. 2018].

Choosing suitable features for the processing of 3D data is similarly difficult. One of the problems is that the field of emotion recognition focuses mostly on speech, facial expressions and gestures instead of body expressions and body movements [De Gelder 2009]. Although there are known links between body movements and emotional states, there is no "standard feature set" available. Human raters can recognise five acted basic emotions with around 75% to 91% accuracy for dynamic emotional displays and 34% to 66% for static emotional displays [Atkinson et al. 2004] – this allows us to conclude that the feature set in question should represent the dynamics of bodily motions. The dynamics can be captured, for example, by using the amount of detected motion, but also the velocity and acceleration of motion [Castellano et al. 2007; Saha et al. 2014]. Especially the features based on velocity, acceleration and movement strength, but also those based on the extension of the body or spatial extent seem to be relevant, as shown by feature selection [Ahmed & Gavrilova 2019]. Although the dynamics should be captured in the features, static postures and geometric features are also used for the recognition of both basic and continuous emotions [Kaza et al. 2016; Wang et al. 2013].

One way to benefit from the redundant and complementary information contained in different communication channels is to incorporate multiple modalities at the same time, since we know that affective states are expressed by the whole body and processed in a holistic way [De Meijer 1989; Witkower & Tracy 2018]. This is also true for the human emotion recognition process – for human judges, using a combination of both, face and body images can improve the recognition of the conveyed emotions [Ambady & Rosenthal 1992]. The same applies for automatic recognition, for example for the bi-modal recognition of face expressions and body gestures [Gunes & Piccardi 2007] as well as facial expressions and speech [Busso et al. 2004]. We will return to the possibilities offered by data from different modalities in Section 4.1.

1.2.3 The Quest for Consistent Evaluation Methods

Besides the search for the best performing features, the evaluation of the emotion recognition systems is another important issue deserving special attention. Although there is no doubt that the systems should be evaluated in way allowing to directly compare their performance, both the evaluation procedures and metrics used in the literature are not consistent. This can be explained by the different backgrounds of the researchers in the field of affective computing, for instance, psychology, neuroscience, computer science, electrical engineering, etc. Recently, this problem has been acknowledged, leading to a strive for

"more standardized evaluations", which is "of crucial importance" in order to compare different approaches [Weninger et al. 2015].

Returning to the approaches mentioned above, we can find a lot of different evaluation procedures and metrics. Unfortunately, many authors report the classification results in terms of "classification performance", "recognition rate" or "error rate", without further specification of its exact calculation. Another problem is that often the only reported metric is the classification accuracy. Accuracy counts the correctly identified instances compared to the total number of instances, regardless of their class. This metric is easy to calculate for binary and multi-class problems, but it does not account for the class imbalance, favouring approaches recognising the majority class [Chawla 2010; Hossin & Sulaiman 2015]. It also impairs the comparability of the approaches, if their performance is measured using different metrics such as accuracy, recall, precision, area under curve, error rate, and so on.

The next point that needs our attention is the evaluation setup that can be either subject-dependent or subject-independent. For the latter, the performance of the recognition system should be tested on data of subjects unseen during the training process. Unfortunately, the published descriptions do not always clearly state if the reported experiments are indeed subject-independent or whether data from the same subject (albeit different samples) are included in the training and evaluation processes. This is a problem – especially for the domain of physiological signals. We know from the literature that in subject-dependent studies, the recognition rates are higher than in subject-independent studies [Healey et al. 2001; Kim et al. 2004; Jatupaiboon et al. 2015]. The same is true for speech data, leading to the use of the term "speaker-independent" as early as 1984 [Leonard 1984]. For bodily movements, the same applies: Here, this is an even bigger problem, since the universality of bodily expression is under discussion, since the affective body language seems to depend on the culture as well as the the current situation [Kleinsmith & Bianchi-Berthouze 2013; Ting-Toomey & Dorjee 2018].

Although this issue has been appreciated, there still seems to be a lack of consistency in the published work. For example, for the domain of emotion recognition from EEG signals, 46.8% of the surveyed approaches use subject-independent setups and 43.5% subject-dependent setups, with 1.7% providing no information on the setup [Alarcão & Fonseca 2019]. Regarding affective body expressions, subject-independent and subject-dependent setups are used side by side, with some investigations reporting a comparison between both setups [Castellano et al. 2007; Karg et al. 2010; Savva et al. 2012]. For speech data, the awareness of this problem seems to have penetrated the research field due to its relevance for speech recognition, as we have already observed above. Most recent approaches are indeed implemented in a subject-independent (i.e.

speaker-independent) way. In general, it can be stated that "it is particularly difficult to determine whether a developed approach is subject independent and can work well with any context" [Poria et al. 2017].

We will return to evaluation procedures and metrics and discuss the available methods and their implementation in detail in Section 2.2.7.

Furthermore, there are two additional issues regarding the topic of evaluation. First, since we evaluate the congruence between the labels obtained during the annotation process (i.e. the assumed ground truth) and the labels obtained in the classification process, the evaluation greatly depends on the quality of the annotated data. Therefore, the reliability of the annotation must be counted as an influence. Second, the sample size is also important, since a certain amount of data must be available to warrant the statistical significance of the results. We will return to these questions in Section 3.1.

This section presented an overview of the development and challenges in the field of affective state recognition in HCI. We will review the existing work in detail in the subsequent chapters for each of the investigated interlocutor states. In the next section, we will identify the problems to be solved in this thesis.

1.3 Identifying the Research Subject

Around a decade ago, three trends in the field of affect recognition have been pointed out: the "striving for more natural and real-life data", "a thorough exploitation of the feature space", and the focus on emotion-related affective states instead of prototypical emotions [Batliner et al. 2011].

The first trend has become the mainstream direction after the introduction of large naturalistic data sets such as the IEMOCAP corpus [Busso et al. 2008], which now acts as a benchmark for testing novel recognition approaches as well as feature sets. However, this does not mean that the problem of emotion recognition has been solved – on the contrary, there is a remarkable gap in recognition performance on acted and naturalistic data. As already mentioned in Section 1.2.1, for acted speech data, recognition accuracies of around 90% for seven emotions are reported [Schuller et al. 2003], whereas for naturalistic data, accuracies of around 64% for four emotions can be achieved so far [Heracleous et al. 2019]. Caused by fundamental differences between acted and naturalistic data (such as higher amount noise and lower expressiveness in the latter case), this problem remains challenging. Regarding the exploitation of the feature space, a shift from hand-crafted features to end-to-end techniques can be observed, as we have already discussed in Section 1.2.2. The third trend, concerning the broader conception of emotion as an affective state, has lead

to protean new application fields of affective computing, for instance, assistive driving systems and empathetic car interfaces [Lotz et al. 2018; Zepf et al. 2019], pain management and stress level evaluation [Thiam et al. 2019; Slavich et al. 2019] as well as recognition of affective disorders [Anis et al. 2018].

This thesis regards emotion as a pervasive concept influencing interaction between humans as well as between humans and computers. We will focus on three specific affective states relevant for HCI as a harbinger of truly adaptive systems, aiming at a better understanding of the relations between the interlocutor's "communicative signals" (be it speech, body movement or physiological signals) and her internal state. In Section 1.1, we have already elaborated on the importance of the three selected states. Furthermore, this thesis aims to contribute to the current state of the art in three methodological questions elaborated below.

1. The first question that we aim to answer is how to collect, select and process data suitable for the task of internal state recognition. There are several requirements that should be covered by the data. First of all, it is necessary to define an appropriate amount of data that not only includes the phenomena that we are investigating, but also acknowledges possible influences such as the distribution of subject characteristics and recording conditions. The data should also reflect real-life and naturalistic situations and surroundings, since we are interested in natural interaction. We will contribute to the solution of this problem in Chapter 3, by analysing the general requirements in detail in Section 3.1 and presenting the solutions found for this thesis in Sections 3.2 and 3.3.

2. The second methodological question that we consider is which modalities and features are suitable for the recognition of the investigated interlocutor states. In Section 1.2.2, we have already discussed that there is no consensus on the selection of modalities and features for certain applications. Each modality has its own advantages and disadvantages. Furthermore, even for the same modality, different features can be used: For example, for the speech modality, we can either use spectral features calculated from very short frames of speech or prosodic features capturing whole sentences. In Chapter 4 and especially in Section 4.6, we will compare different modalities and the implications of the modality choice.

3. The third methodological question we want to address is how to correctly evaluate the developed recognition systems. As we have seen in Section 1.2.3, this question has not yet obtained the deserved attention, resulting in different evaluation setups and metrics being used concurrently. In Section 4.5.2, we will analyse different evaluation setups in order to investigate their advantages and disadvantages. In Section 6.5.2, we will develop evaluation setups and metrics that can be used in es-

pecially difficult scenarios in order to ensure subject-independence and accounting for imbalanced data.

After stating the expected scientific contributions and the added value, we can define the aim and objectives of this thesis in the next section.

1.4 Aim of the Thesis

The sovereign aim of investigations in the field of HCI is, without doubt, the improvement of the interaction experience – making the interaction more efficient and effective, but also more pleasant and rewarding. This aim can only be achieved by partial contributions developing over decades.

The aim of this thesis is to enable a deeper understanding of HCI by investigating three exemplary interaction-related interlocutor states based on specific cues from different modalities. This aim can be further refined to the following objectives:

- selecting and generating the data necessary for data-driven modelling,
- finding the right features and developing a model for the recognition of the three states of interest: the state of *trouble*, *satisfaction* and *cooperativeness*,
- analysing and interpreting the achieved results to identify the next steps necessary for further development.

The technical purpose of this thesis is the recognition of pre-defined interlocutor states in naturalistic, real-life interactions using different modalities. Regarding the acoustic modality, the focus lies on the spectral and prosodic features of the speech as well as specific events such as speech overlaps. From the biophysiological modality, we will use features derived from EMG signals, SC and RSP. Furthermore, we will investigate features based on the interlocutor's 3D movements. As the basis for these investigations, we will use three existing corpora: Two versions of the Last Minute Corpus (LMC), a collection of naturalistic multimodal HCI recordings, and the Davero Corpus (DC), a collection of real-life call centre HHI telephone conversations. For the recognition, we will apply existing classification methods depending on the challenges of the specific state to be recognised. We will especially focus on the evaluation of the classification, striving for a truly subject-independent evaluation and meaningful metrics.

The scientific purpose is to investigate the relationship between the signals sent by the interlocutor and her inner state, and to find out which signals can be used for the recognition of the state. Furthermore, this thesis contributes to the ongoing discussion on methodological questions.

The presented research is bound to a certain framework consisting of existing methodology and the available resources in terms of software, data, and also hardware. The first technical problem to be solved is proceeding from raw data to real insight – this means choosing the right data and processing the data in an appropriate way, as well as establishing the ground truth as a basis for the subsequent recognition experiments. Furthermore, choosing the relevant signals from the available data, especially with limited resources in terms of limited data points, annotations and channels, is also an issue to be solved. Despite these restrictions, the generality of the presented findings must be ensured – this issue demands special attention by implementing interlocutor-independent evaluation of the recognition models and interpreting the classification results within these constraints. The small sample size must be taken into account when deriving statements on the transferability of the investigated approaches. These are the main technical difficulties to be solved, besides implementation problems such as computational time and the availability and usability of toolkits.

Based on this, we can derive the limitations of the research presented in this thesis. Obviously, choosing only three interlocutor states leaves all other possible states outside of the scope of the investigations. Furthermore, data-driven modelling always relies on the availability and quality of data, since only the relations contained in the data can be learned and recognised. For example, the small size of the data handled in this thesis does not allow to apply deep learning methods. Furthermore, there is only a limited number of classification approaches with their respective settings and feature sets that can be tested, making the search for the "perfect" classifier a never-ending endeavour that we will have to stop to pursue at some point.

Having established the aim as well as the limitations of this thesis, we can take a look at the overall structure before beginning to "access the interlocutor" in the following chapters.

1.5 Structure of the Thesis

The current chapter introduced the topic of the thesis by explaining how the presented research can be integrated into the field of HCI research and especially the field of affective computing. The next chapter, Chapter 2, provides background information necessary for understanding the remaining part of the thesis, including the definition of important concepts and details on the employed methods. Chapter 3 explains the importance of data in general, defines the requirements placed on data for the field of affective computing and describes the data used throughout the thesis. Chapters 4, 5 and 6 focus on the three previously mentioned interlocutor states. Each of the chapters starts

with an overview of existing work on the respective state before presenting the developed approaches for its recognition, constituting the main scientific contribution of this thesis. Chapter 4 is dedicated to the detection of *trouble* in HCI using different modalities. Chapter 5 deals with the recognition of *satisfaction* in spoken HCI. Chapter 6 introduces speech overlaps as markers for *cooperativeness* in interaction and presents a classification approach for cooperative and competitive overlaps in HHI based on emotional features. Chapter 7 summarises and discusses the thesis and evaluates the achieved results before providing insight on open questions and possible future developments.

Chapter 2

Important Concepts and Methods

Contents

BEFORE diving into the topic of interlocutor behaviour and the recognition of internal states, we should first illuminate the concepts used throughout this thesis and find coherent descriptions that we can refer to in the following chapters. First of all, we need decisive definitions of terms such as interlocutor states in interaction, but we should also elaborate on the concepts of interaction channels and interlocutor signals. We will do so in Section 2.1. Furthermore, we will provide an overview of methods available for the task of interlocutor state recognition, focussing on machine learning and related questions such as feature extraction, evaluation scenarios and classification methods. This will be done in Section 2.2.

2.1 Definition of Important Concepts

In Chapter 1, we have already seen an overview of the field of affective computing in general and especially internal state modelling. Now, we want to precisely define what user states we are interested in as well as what we mean when we speak of an interlocutor signal.

2.1.1 Internal Interlocutor States

As already stated in Section 1.1, we are interested in three distinctive interlocutor states that can occur in interaction, be it HCI or HHI. In the context of HHI, affective or emotion-related states are defined as an "affective stance taken towards another person in a specific interaction, colouring the interpersonal exchange in that situation" [Scherer 2003]. Extending this definition to HCI, the term *interlocutor state* refers to the mental and affective state experienced by a participant of an interaction during this interaction. Such a state can be described situationally, which means by describing the situation inducing this state.

The first state we are interested in is *trouble*. The participant of an interaction experiences trouble when there is a mismatch between her expectations and the real course of the interaction. This state is characterised by a complex mix of dissatisfaction, irritation, anger and frustration occurring when the interaction becomes challenging. It is important to recognise this state, since it can be seen as a "critical phase of the dialogue" [Batliner et al. 2003]. In the scope of this thesis and the investigated data, this state is induced by posing an increasingly challenging task. We will return to this state and examples for interaction in this state in Section 4.2.

The second state of interest for this thesis is *satisfaction*. The participant of an interaction experiences satisfaction when she is content with the current interaction course. Loosely following the definition of "user information satisfaction" [Ives et al. 1983], we define that the interlocutor is in the state of satisfaction when she believes that the current interaction and situation meets her requirements in an adequate way. This definition touches similar aspects as ISO 9241-11, defining satisfaction as "freedom from discomfort and positive attitudes towards the use of the product" [ISO9241-11:1998 1998], or, more recently, "positive attitudes, emotions and/or comfort resulting from the use of a system" [ISO9241-11:2015 2015]. We will see how the concept of the state of satisfaction can be grasped in the context of HCI in Section 5.2.

The last state for the scope of this thesis is the state of *cooperativeness*. The participant of an interaction can be seen as cooperative or experiencing cooperativeness when she coordinates her behaviour with other interaction participants "to achieve mutual goals" [Johnson 1975]. Regarding this state, we are mainly interested in the interlocutor's speaking behaviour. We assume that the interlocutor is experiencing cooperativeness when she speaks cooperatively, without interrupting the interaction partner in a competitive way. We will analyse this state in more detail using the example of cooperative and competitive speech overlaps in Section 6.1.

2.1.2 Interaction in Multiple Modalities

The interaction between humans is always holistic, conveying different "communicative signals in a complementary and redundant manner" [Jaimes & Sebe 2007], in a joint process [Chen 2000]. In other words, we send and receive messages using different channels or *modalities*. A modality is defined as a "mode of communication according to human senses" of vision, hearing, touch, smell and taste [Jaimes & Sebe 2007]. However, not all modalities are used to the same extent in both, HHI and HCI. Therefore, the focus in HCI research lies on the most common ones, such as speech, gestures, touch, as well as facial expressions.

When a system uses different modalities for input and/or output, we speak of a *multimodal* system. A common example of such a system is a computer with a keyboard and a mouse: We can interact with the computer via the two modalities of text and mouse movements, and the computer can interact with us using visual output via the monitor and audio output via loudspeakers. In the scope of this thesis, we will encounter the modalities of audio signals in the form of speech and speech-related characteristics, physiological signals in the form of electromyogram, respiration and skin conductivity, as well as movement signals in the form of three-dimensional upper-body postures.

When multiple modalities are processed in parallel, we speak of *multimodal fusion* [Corradini et al. 2005]. This can take place on different levels. Fusion on feature level, which means the combination of the input signals resulting in a joint feature space, is called early fusion. Fusion on semantic or decision level, which means the combination of the individual processing results, is called late fusion. It is also possible to fuse the different modalities on an intermediate level. For human sensory input processing, the famous McGurk effect confirms that auditory and visual information is processed in a joint manner [McGurk & MacDonald 1976]. Furthermore, there is evidence that the audiovisual integration occurs prior to word identification, which would correspond to early fusion [Barutchu et al. 2008]. Although seeming to be the "natural way" of multimodal fusion, early fusion also poses difficult challenges, such as the large dimensionality of the feature space, different temporal resolution of the fused signals and their synchronicity [Jaimes & Sebe 2007].

However enriching multimodal processing might be, even one modality can offer different kinds of information. The audio signal carries the speech with its linguistic content, but also additional paralinguistic information, such as the implicit messages we discussed in Section 1.1. Such additional information can be contained in certain speech phenomena such as discourse particles and filled or silent pauses, speech overlaps and even the speech rate. The same is also true for bodily movements: For instance, a certain gesture itself sends an explicit message, but the way it is performed (slowly, repeatedly, etc.) sends an

implicit one. Such phenomena can be captured during the feature extraction process which we will discuss in Section 2.2.4.

2.2 Applied Methods and Technical Background

This section aims at providing a general overview of the methods applied throughout the thesis. Roughly speaking, a large part of the presented work deals with pattern recognition. The basic idea of pattern recognition is surely to detect anomalies in a collection of data samples. From the vast amount of approaches to solve this task, we will focus on two areas. First, we will apply statistical methods to show that there are significant differences in the data. Second, we will use classification to assign different classes to the data based on the statistical differences contained within. Therefore, this section will focus on statistical and classification methods. But first, we need a consistent definition and description of data.

2.2.1 Data Description

For the scope of this thesis, we define *data* as a collection of samples described by their characteristics or features. In this way, the data D can be seen as an $(n \times m)$-matrix, with each of the m rows corresponding to a sample of the data, which in turn consists of n features:

$$D = \begin{pmatrix} x_1^1 & x_2^1 & ... & x_n^1 \\ x_1^2 & x_2^2 & ... & x_n^2 \\ ... & ... & ... & ... \\ x_1^m & x_2^m & ... & x_n^m \end{pmatrix}, D \in \mathbb{R}^{n \times m} \tag{2.1}$$

We write $\mathbf{x^i}$ when we refer to an individual sample with its realisation, e.g. $\mathbf{x^1} = (x_1^1, ..., x_n^1)$ and $\mathbf{x_j}$ when we refer to the realisation of a feature over all samples, e.g. $\mathbf{x_1} = (x_1^1, ..., x_1^m)^T$. Furthermore, to refer to the features, we define a feature set F with n elements, with each feature F_i corresponding to the concept of the ith feature instead of its specific realisation in the ith column of D.

2.2.2 Statistical Methods

As already mentioned above, the first step towards the recognition of certain phenomena is the detection of significant anomalies in the data. Therefore, we need to employ statistical methods to analyse a problem prior to classification. The assumption here is that the observable differences in the data are caused by the differences in the underlying processes generating the classes of the data.

This enables the classification of the data based on the observable characteristics (i.e. feature realisations). One direction to find relations between the data and the underlying processes is to search for correlations between certain characteristics, or more precisely, between differences in characteristics depending on certain events. For this, we can use *Pearson's correlation coefficient*. Another direction is to test a hypothesis describing the data (or the distribution of samples in the data) – for example, to compare the observed distribution of samples to their expected distribution in order to find significant differences. For this, we will use *Fisher's exact test*.

Pearson's Correlation Coefficient

One important tool for assessing the correlation between two characteristics is the *Pearson's correlation coefficient* r [Lee Rodgers & Nicewander 1988]. The value of r can vary between 1 for a perfect correlation, 0 for no correlation and -1 for an inverse correlation.

Simply speaking, given a number of samples, the values of two features of these samples are correlated if certain values of one feature occur together with certain values of the other feature. For example, the features *size* F_s and *age* F_a of humans are correlated up to a certain degree, since for small values of F_a, small values of F_s can be expected. However, these two features are not perfectly correlated, since people stop growing at a certain age.

For two features $\mathbf{x_j}$ and $\mathbf{x_k}$ over m samples, given $cov(\mathbf{x_j}, \mathbf{x_k})$ as their covariance, $\sigma(\mathbf{x_j})$ and $\sigma(\mathbf{x_k})$ as their standard deviations, and $\mu(\mathbf{x_j})$ and $\mu(\mathbf{x_k})$ as their mean values, we define $r(\mathbf{x_j}, \mathbf{x_k})$ as follows:

$$r(\mathbf{x_j}, \mathbf{x_k}) = \frac{cov(\mathbf{x_j}, \mathbf{x_k})}{\sigma(\mathbf{x_j})\sigma(\mathbf{x_k})} = \frac{\sum_{i=1}^{m}(x_j^i - \mu(\mathbf{x_j}))(x_k^i - \mu(\mathbf{x_k}))}{\sqrt{\sum_{i=1}^{m}(x_j^i - \mu(\mathbf{x_j}))^2}\sqrt{\sum_{i=1}^{m}(x_k^i - \mu(\mathbf{x_k}))^2}} \quad (2.2)$$

We will use Pearson's correlation coefficients in Section 4.3.1 to find significant correlations between features of speech signals and the interaction stage.

Fisher's Exact Test

The first step in statistical hypothesis testing is to define a hypothesis about an observable phenomenon, i.e. a statement describing a population. The subsequent test shall then prove whether the hypothesis should be accepted or rejected based on an observed sample.

An important statistical test that we will encounter in this thesis is *Fisher's exact test* [Andrés & Tejedor 1995]. This test is a statistical significance test

for categorial data used to find out whether there are significant differences between the expected and the observed outcome of an experiment. In this way, the test is similar to the χ^2 test as well as the G-test and is designed for small sample sizes. In general, it is a method to analyse contingency tables in order to find whether the expected distribution deviates significantly from the assumed distribution.

In the original setup, the test was developed by Fisher for a "Lady Tasting Tea" experiment [Fisher 1956]. In this experiment, a person claims that she can discriminate whether the milk or the tea was poured in the cup first based on the taste of the resulting beverage. In a setting with a number of cups of tea, some of which are in the "pouring condition" *tea-first* and others in the condition *milk-first*, this person has to decide (i.e. classify) whether she thinks that a particular cup belongs to the first or to the second condition (i.e. its class) based on the perceived taste (we will refer to it as *tea-ish* vs. *milk-ish*). What shall be tested is whether the classification performed by the person based on the taste is associated with the real class, the pouring condition. The null hypothesis for this test is that both classifications are independent, i.e. that the taste is independent of the pouring condition. The rejection of this hypothesis leads to the conclusion that the classifications are associated, meaning that the taste depends on the pouring condition. For the test, we must generate two contingency tables: The first one corresponds to the expected (a priori) distribution of the samples over the classes (e.g. an equal distribution of *tea-first* and *milk-first*), the second one to the distribution of the observed classes (*tea-ish* and *milk-ish*). The test now calculates how much the observed distribution deviates from the expected distribution. If the deviation is significant, then the hypothesis that the classifications are independent (i.e. the person is merely guessing) is rejected. Being an exact test, this test also delivers the significance value p. Table 2.1 shows a general contingency table referring to the previous tea tasting experiment.

Table 2.1: A contingency table for the tea tasting experiment mentioned above. The variables a, b, c, d correspond to the numbers of the observed instances of the respective classes.

Condition	Tea-first	Milk-first	Total
Tea-ish	a	b	a + b
Milk-ish	c	d	c + d
Total	a + c	b + d	a + b + c + d = n

Given this table, p can be calculated as follows:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{2.3}$$

We will apply Fisher's exact test to a real example in Section 6.4 in order to decide whether the types of speech overlaps are associated with the emotional content of the surrounding utterances.

2.2.3 Classification Pipeline

In Section 1.2, we have already briefly described how classification is used for internal state recognition. Now, we want to attend to the details of the classification process. In general, it can be defined as "inferring a boolean-valued function from training examples of its inputs and outputs" [Mitchell 1997]. As depicted in Figure 2.1, the process of classification consists of several stages that we will discuss below.

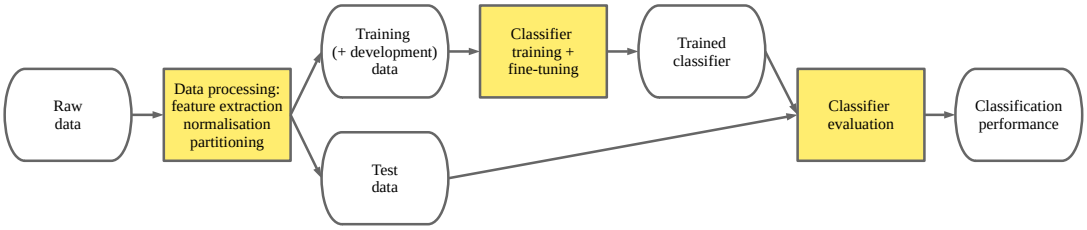

Figure 2.1: A schematic depiction of the classification pipeline.

First, we need to acquire the necessary data and process them appropriately. This processing consists of *data selection* and *data annotation*. Data selection is required to ensure that the data contain a sufficient amount of variation to provide an objective representation of reality. The process of data annotation is necessary to establish the "ground truth", which is a term borrowed from remote sensing referring to the real world conditions obtained on site – in other words, the verified class of a sample.

Having obtained annotated data, we need to break it down to specific features describing the data, which is called *feature extraction*. As we have already discussed in Section 1.2, these features must contain the information necessary for the classification, i.e. they must convey the class differences. This information is already present in the raw signal and is not generated during feature extraction – however, this step is necessary to make it available to the classifier. After the feature extraction, the data can be prepared for classification by specific pre-processing, such as normalisation, data partitioning, data labelling, etc.

In the next step, the classifier – the particular method performing the classification – must be trained on a partition of the data, the training data. During this *training process*, the classifier generates a recognition model mentioned in Section 1.2 by "learning" the differences in the features of the data with respect to the classes and adapting its internal parameters. The best configuration of the classifier – referred to as hyperparameters – can be fine-tuned by repeatedly testing different hyperparameter settings on another partition of the data, often called the development data.

In the final step, *evaluation*, the classifier is applied to a previously unseen portion of the data, the test or evaluation data. The performance of the classifier on these data, i.e. how the assigned classes match the true classes of the data, constitutes the classification performance that can be measured using different evaluation metrics. In order to ensure the generalisation ability of the classification, the evaluation must be performed in a way rendering the results independent of the individual subject characteristics, or subject-independently.

These concepts will be further elaborated one by one below.

2.2.4 Feature Extraction and Normalisation

As already discussed in Section 1.2.2, finding suitable features is a crucial point for recognising certain phenomena. Features of physical objects can be viewed as their attributes – such as size, shape, colour, etc. When we speak of features in the scope of this thesis, we mean characteristics describing an interlocutor's signal in a meaningful way, minimising the possible redundancy and maximising the contained information with respect to differences in data of different classes. The process of calculating such features is referred to as feature extraction. During this process, we transform the raw data recorded in the real world into features. We have already briefly introduced the data consisting of samples and their features in Section 2.2.1. As a reminder, each data sample $\mathbf{x^m}$ is an n-dimensional vector defined by n feature values, with each feature corresponding to a dimension in the feature space.

The feature extraction process consists of several steps that we explain exemplarily for the raw audio signal – for other signals, this procedure can be performed in an analogous way. First, the continuous input signal is divided into computationally manageable short portions called frames, which usually overlap to capture the temporal changes in the signal. Each frame must lie within a temporal segment when the process generating the data is (or is assumed to be) stationary with respect to the data class. For each frame, characteristics describing the signal are calculated, for instance, the spectrum of the signal, the fundamental frequency, the energy, etc. This process results in a feature vector for each frame that can be enriched by further values such as

statistical functionals (e.g. minimum and maximum value of each feature over a certain number of the original frames). The functionals convey additional information on the temporal changes of the signal, since they are calculated over a longer period of time. Furthermore, a label referring to the ground truth described by the signal is usually added at this stage – this issue will be addressed in Section 2.2.5. Finally, we obtain the sample data that can be used for the classification process.

In the next chapters, we will repeatedly encounter the so-called *emobase* feature set which is one of the baseline feature sets of the openSMILE toolkit [Eyben et al. 2014]. It is widely known in the community of speech processing and is used in a broad variety of domains, such as acoustic scene classification [Marchi et al. 2016], humour detection [Bertero & Fung 2016], physical pain detection [Oshrat et al. 2016], spontaneous speech recognition [Toyama et al. 2017], dialogue performance evaluation [Ramanarayanan et al. 2017], etc. It contains 988 spectral, prosodic and voice quality features based on 26 Low-Level Descriptors (LLDs) and their deltas with 19 functionals. In the original version of this feature set, the LLDs are extracted from a 25 ms window with a 10 ms shift, with the functionals calculated for a whole utterance. The full list of features is shown in Table 2.2.

Table 2.2: An overview of the 988 features in the *emobase* feature set (courtesy of A. Requardt [Requardt et al. 2019]).

LLDs & their Δ	**Functionals**
Intensity	Minimum value
Loudness	Maximum value
12 mel-frequency cepstrum coefficients	Position index
Pitch (F_0)	Range
F_0 envelope	Mean value
Voicing probability	2 linear regression coefficients
8 line spectral frequencies	Linear & quadratic error
Zero-crossing rate	Standard deviation
	Skewness
	Kurtosis
	Quartile 1-3
	3 inter-quartile ranges

We will return to the feature sets specific to our research questions in the respective chapters in order to describe them in detail.

The features extracted from the interlocutor signals are highly dependent on the individual interlocutor characteristics, e.g. the voice frequency in case of audio signals. In the classification process, we need to find the differences between the classes in question, and therefore it is necessary to account for the

inter-individual differences in the data. This can be done by a process called *normalisation*, which can be implemented using several different methods. In this thesis, we will use only standardisation, an approach that "uses the mean and variance values to transform the given input [...] to achieve [...] zero mean and variance of one" [Böck et al. 2017a].

Given a sample $\mathbf{x^i}$ and a feature $\mathbf{x_j}$, its mean value $\mu(\mathbf{x_j})$ and standard deviation $\sigma(\mathbf{x_j})$, we obtain the standardised sample values s_j^i as follows:

$$s_j^i = \frac{x_j^i - \mu(\mathbf{x_j})}{\sigma(\mathbf{x_j})} \tag{2.4}$$

We will use only standardisation, since it was shown that any kind of normalisation improves the classification results, with standardisation achieving the largest performance boost, by my colleagues and myself [Böck et al. 2017a].

2.2.5 Data Annotation, Labelling and Partitioning

Simply obtaining the features from raw data is not enough – the feature vectors also need a target value, the label describing their designated class. In cases when the underlying process is known, classes can be assigned (i.e. annotated) directly. If it is not known, the labels can be generated during the annotation process. In our tea example used for Fisher's exact test above, the classes *tea-first* and *milk-first* were known by design. But if we were interested in the taste preferences regarding the tea, we would have to ask the subjects, since there is no other possibility to assess the ground truth. We use the term *annotation* to describe the process of defining the ground truth in the raw data, for instance, by listening to a recording containing a person's laughter and assigning it the class label "laughter". The mapping of the feature vectors obtained from this recording to the class of the recording is referred to as *labelling* – in this process, all feature vectors obtain an additional attribute called the *class attribute*, containing the class label. Without loss of generality, we consider only problems with two classes.

Given two classes a and b as well as the data matrix D as defined in Equation 2.1, we define a set of class labels $C = \{c_a, c_b\}$ with c_a indicating a membership in class a and c_b indicating a membership in class b. Given I as the set of all sample indices $\{1, 2, ..., m\}$, the labelling function χ can now be defined as follows for class a:

$$\chi : I \to C, \quad \chi(i) = c_a \Leftrightarrow \mathbf{x^i} \quad \text{is of class } a \tag{2.5}$$

The same can be defined for class b in an analogous way.

Another important processing step is *data partitioning*. Since the aim of pattern recognition is to find general patterns describing the investigated phenomenon (*to generalise*), the recognition algorithm must be able to recognise the phenomenon in previously unseen data. Therefore, the data must be partitioned prior to classification to ensure the generalisation ability of the classifier. We will address the problem of measuring the performance of a classifier in detail in Section 2.2.7. In general, the data must be divided into at least two sets: The training set is used for the training of the classifier, whereas the test set is used to evaluate the generalisation ability or performance of the classifier on previously unseen data. Furthermore, a development set can be used to "develop" the classifier by testing different classifier hyperparameters. As we are interested in the recognition of states of individuals, we need to pay special attention to generalisation ability over different individuals – therefore, it is not enough to use different samples in the different data sets, it is also necessary that these samples are from different individuals, reflecting the characteristics of the sampled data in a balanced way. In this case, we speak of *subject-independent* sets, since the different sets contain data from strictly different subjects and the classification is based on differences independent of individual subjects.

2.2.6 Classification Methods

In this thesis, we will encounter three different classifiers, Support Vector Machine(SVM), Random Forest(RF) and Naïve Bayes (NB). All three classifiers are widely known and well established tools for pattern recognition tasks, therefore we will focus especially on those aspects of these techniques that are of interest in the scope of this thesis.

Support Vector Machine

One of the most widely used classification methods are SVMs originally developed in 1982 [Vapnik 1982] and further improved by adding the so-called "kernel trick" [Boser et al. 1992] as well as the "soft margin" [Cortes & Vapnik 1995]. An SVM is a so-called large margin classifier, typically used for two-class problems. The idea behind it is to separate the samples belonging to two different classes by introducing a hyperplane between them. We have already defined a sample as a vector in the n-dimensional feature space in Section 2.2.4. The hyperplane is constructed or "supported" by taking into account the vectors closest to this hyperplane, the so-called "support vectors", hence the name of the classifier. These vectors are chosen based on the assumption that they are the hardest to separate into classes, since they populate the same area in the feature space and thus are similar. This is also the reason for opting for a "large margin" between the samples of different classes during the construction

of the hyperplane. The maximisation of the margin shall guarantee the generalisation ability of the classifier, again based on the assumption that samples from the same class populate the same area in the feature space.

In the original, relatively simple, setting, only linear classification can be performed. But SVMs can be adapted to be used for non-linear problems by introducing two further techniques, the soft margin and the kernel function. The idea behind the soft margin is to allow some misclassifications, i.e. samples being inside the margin or even on the "wrong" side of the hyperplane. The degree of tolerance against such misclassifications is determined by the cost parameter C, which defines the penalty of a misclassification – the higher C, the "harder" is the margin. The idea behind the kernel function is to "map the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori" [Cortes & Vapnik 1995], assuming that "in this space a linear decision surface [can be] constructed with special properties" to ensure the generalisation ability [Cortes & Vapnik 1995].

For this, we need a definition of linear separability for two sets D_a and D_b which are defined as follows:

$$D_a \subset D, D_a = \{\mathbf{x^i} | \chi(i) = c_a\}, \quad \text{analogously for } D_b. \tag{2.6}$$

Two sets D_a and D_b are linearly separable in an n-dimensional space if there exists a vector of weights $\mathbf{w} = (w_1, w_2, ..., w_n) \in \mathbb{R}^n$ as well as a threshold h such that the following is true:

$$\sum_{j=1}^{n} w_j x_j^i > h \quad \forall \mathbf{x^i} \in D_a \quad \text{and} \quad \sum_{j=1}^{n} w_j x_j^i < h \quad \forall \mathbf{x^i} \in D_b \tag{2.7}$$

Given two sets D_a and D_b that are not linearly separable, the transform ϕ is defined as a function which yields two sets D'_a and D'_b that are linearly separable:

$$\begin{aligned} \phi : \mathbb{R}^n \to \mathbb{R}^t, t > n \\ D'_a = \{\phi(\mathbf{x^i}) | \forall \mathbf{x^i} \in D_a\} \\ \text{and} \quad D'_b = \{\phi(\mathbf{x^i}) | \forall \mathbf{x^i} \in D_b\} \end{aligned} \tag{2.8}$$

such that D'_a and D'_b linearly separable in the sense defined above.

The so-called kernel trick behind this idea is that if there is a kernel function K sufficing to certain properties, ϕ does not need to be computed to find the hyperplane. K is defined as follows:

$$\begin{aligned} K : \mathbb{R}^n \times \mathbb{R}^n &\to \mathbb{R} \\ K(\mathbf{x^i}, \mathbf{x^j}) &= \phi(\mathbf{x^i}) \cdot \phi(\mathbf{x^j}) \end{aligned} \tag{2.9}$$

Each tuple of vectors $(\mathbf{x^i}, \mathbf{x^j})$ can be compared in the original feature space, transforming the result of the comparison by a non-linear transformation, instead of transforming the vectors first [Boser et al. 1992]. Two most often used kernels also provided by numerous SVM implementations are the polynomial kernel K_{pol} with the degree d and the Gaussian Radial Basis Function (RBF) kernel K_{rbf}:

$$K_{\text{pol}}(\mathbf{x^i}, \mathbf{x^j}) = (\mathbf{x^i} \cdot \mathbf{x^j})^d \tag{2.10}$$

$$K_{\text{rbf}}(\mathbf{x^i}, \mathbf{x^j}) = exp(\gamma||\mathbf{x^i} - \mathbf{x^j}||^2) \tag{2.11}$$

The effect of K_{pol} applied to an example is shown in Figure 2.2. In the example's original feature space, there are only two features, F_1 and F_2. As we can see in Figure 2.2 (a), the samples from different classes cannot be separated by a linear hyperplane, but by a parabolic one. Applying the polynomial kernel to the two features, we obtain a third feature F_3. Given a sample $\mathbf{x^i}$, the value of F_1 being x_1^i and of F_2 being x_2^i, the value x_3^i of F_3 can be calculated as follows:

$$x_3^i = (x_1^i)^2 + (x_2^i)^2 \tag{2.12}$$

In Figure 2.2 (b), we can see that the two classes are now separable in this new feature space by a linear hyperplane. We can define this hyperplane $\mathbf{g}$ as follows:

$$\mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} g_1 \\ 22.5 \end{pmatrix} \tag{2.13}$$

By inverting Equation 2.12, $\mathbf{g}$ can be transformed back to the original feature space, resulting in a parabolic hyperplane $\mathbf{h}$:

$$\mathbf{h} = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} h_1 \\ \sqrt{22.5 - (h_1)^2} \end{pmatrix} \tag{2.14}$$

This is shown in Figure 2.2 (a).

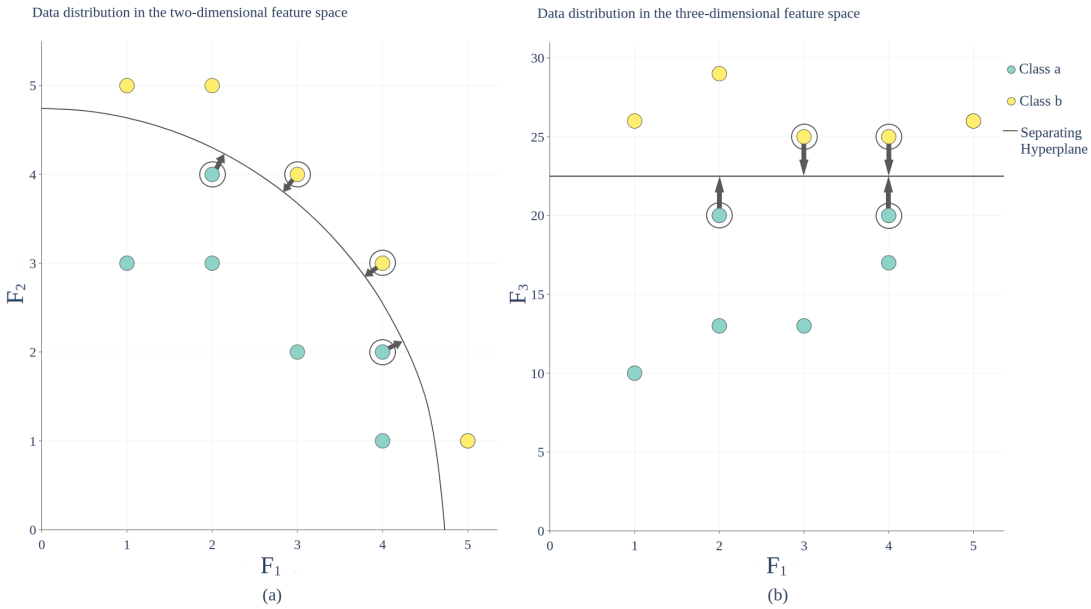

Figure 2.2: Distribution of samples of two classes. (a) shows the distribution in the original two-dimensional feature space, (b) shows the distribution in the three-dimensional feature space that was created using a polynomial kernel. The encircled samples constitute the support vectors.

In most implementations of the SVM classification algorithm, we can choose the hyperparameter C as well as the type of the kernel and its internal hyperparameters (such as the polynomial degree d for the polynomial kernel and the inverse standard deviation γ for the RBF kernel).

We will employ SVMs in Section 5.4 for the task of *satisfaction* classification.

Random Forest

Another classification method applied in this thesis is RF. RF is an ensemble learning method based on decision trees introduced in 1995 [Ho 1995] and further refined in 2001 [Breiman 2001].

The main idea behind a decision tree is to predict the class (or class probability) of a sample $\mathbf{x^i}$ using decision rules based on its feature realisations $(x_1^i, ..., x_n^i)$. In each internal node of such a tree, a decision is made based on a simple rule (e.g. $x_1^i \leq t_1$ for a certain threshold value t_1), before assigning $\mathbf{x^i}$ to a class c in the terminal node (often called the leaf). An exemplary classification is shown in Figure 2.3. Starting from the root on the left-hand side, two decisions are made based on the values of the features F_1, F_2 and F_3 before deciding on c of $\mathbf{x^i}$ in the leaves on the right-hand side.

In the training process, the trees are constructed from the training data using a specific training algorithm (such as ID3, CART, etc.). This is done

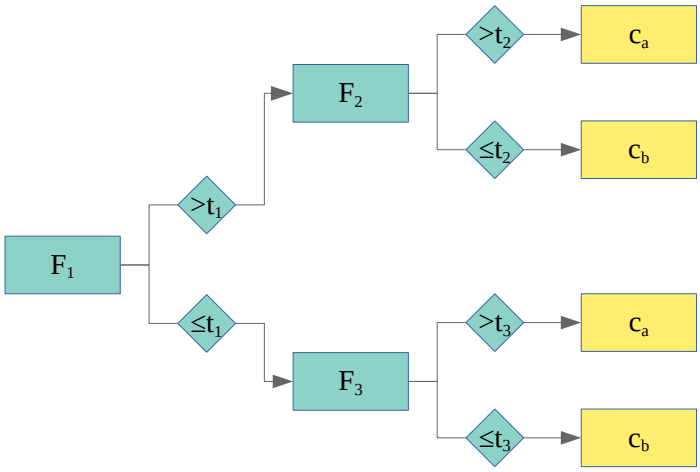

Figure 2.3: An exemplary illustration of a decision tree. The samples are assigned to the classes c_a and c_b based on the values of F_1, F_2 and F_3 compared to the respective threshold values t_1, t_2 and t_3.

in a top-down way: Starting from the root of the tree, the internal nodes use the feature suited best to split the training data into subspaces, using different metrics to define what "best" means [Rokach & Maimon 2005]. One such metric is the Gini impurity, which measures how often a sample from a certain set would be labelled incorrectly if it was labelled randomly with respect to the class distribution in the sample set. Given an n-class problem with $1 < i \leq n$ and $p(i)$ being the prior class probability of class i in the set, the Gini impurity G can be calculated in the following way:

$$G = \sum_{i=1}^{n} p(i) \cdot (1 - p(i)) \tag{2.15}$$

The problem with this setup is that such trees are "prone to be overly adapted to the training data" [Ho 1995]. Before the development of RF, there was no method to grow trees to arbitrary complexity without overfitting [Ho 1995]. For RFs, the decisions are made not on all features but on a randomly chosen subset of the feature space (hence the "random" part of the name), combining a typically high number of trees to cover many of these subspaces (resulting in a "forest"). Each of the trees in the forest casts an independent vote on the class of a sample, before eventually the final class is calculated by a majority over these individual votes. In general, we can define an RF as:

> "[A] classifier consisting of a collection of tree-structured classifiers $\{h(x, \Theta_k), k = 1, ...\}$ where the $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x" [Breiman 2001].

In general, RFs have several hyperparameters that can improve the classification performance, from the configuration of the forest itself (e.g. the number and kind of trees) to the configuration of the individual trees (e.g. the measure for selecting the "best" split, the minimum number of instances per leaf, etc.). In this thesis, we will only optimise the number of trees n_t and the number of features in each of the nodes n_f in order to limit the possibilities – this is a common practice, since these two hyperparameters have the greatest influence [Bernard et al. 2009; Ren et al. 2015; Biau & Scornet 2016].

We will encounter RFs in Sections 4.3.2, 4.4.2 and 4.5.2 for the task of *trouble* classification.

Naïve Bayes

The last classification method that we want to look at in detail is the NB classifier, which is based on the application of the Bayes theorem to pattern classification [Duda & Hart 1973]. The idea behind this classifier is to "learn from training data the conditional probability of each [feature $\mathbf{x_j}$] given the class label c" [Friedman et al. 1997]. Given the prior probability $P(\mathbf{x_j}|c) \quad \forall c \in C$ obtained from the training portion of the data, we calculate the posterior probability $P(C|\mathbf{x^i})$ by application of the Bayes' theorem and choosing the class with the highest posterior probability [Friedman et al. 1997]. The *naïve* characteristic of this classifier is the assumption that each feature contributes to the final classification independently, i.e. both, the value and even presence of a feature does not influence other features. On the one hand, this is a big advantage, since this classifier can work natively with missing feature values. On the other hand, it cannot learn the relationships between features, since all features are seen independently.

In order to understand this classifier, we should look at the Bayes' theorem. Assuming that $P(H)$ is the probability of a hypothesis H being true, and $P(E)$ is the probability of the given evidence E, $P(E|H)$ is the probability of the evidence given that the hypothesis is true, whereas $P(H|E)$ is the probability of the hypothesis being true given the evidence. The conditional probability can now be calculated as follows:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} \tag{2.16}$$

For the NB classification, the hypothesis corresponds to the assumed class c of $\mathbf{x^i}$, and the evidence to its feature realisations $(x_1^i, x_2^i, ..., x_n^i)$. The probability of $\mathbf{x^i}$ belonging to c is calculated as follows:

$$P(c|\mathbf{x^i}) = \frac{P(x_1^i|c) \cdot P(x_2^i|c) \cdot ... \cdot P(x_n^i|c) \cdot P(c)}{P(x_1^i \cdot x_2^i \cdot ... \cdot x_n^i)} = \frac{\prod_{j=1}^{n} P(x_j^i|c)}{P(x_1^i \cdot x_2^i \cdot ... \cdot x_n^i)} \cdot P(c) \tag{2.17}$$

After calculating the probabilities for $\mathbf{x^i}$ belonging to each of the classes $c \in C$, the most probable class is selected. Interestingly, NB has been shown to perform even if the features are not necessarily conditionally independent. It can be optimal in case the dependencies are distributed evenly or cancel each other out [Domingos & Pazzani 1997; Zhang 2004].

In the scope of this thesis, we will use NB for discrete values – but it can be used for continuous values as well, by substituting the probability with the probability distribution. We will employ the NB classifier for the task of overlap classification in Section 6.5.

2.2.7 Evaluation Methods

The aim of the evaluation process is to measure the classification success. As previously mentioned, we need to ensure the generalisation ability of the classifier. This requires preparing the data in a certain way that is described in Section 2.2.5. The main objective of this data preparation is to ensure that the test data are not included in the training data. Since our aim is to find general patterns independent from individual subjects whose data we use for the classification, we speak of subject-independent evaluation. For this, the classification performance should remain stable over data of different subjects.

Evaluation Settings

In this thesis, we will encounter two different *evaluation settings*, Train-Dev-Test (TDT) and Leave-One-Subject-Out (LOSO). Both settings are variants of so-called "holdout evaluation", where the main idea is to partition the data in disjoint sets for the development of the classifier and its evaluation [Sammut & Webb 2010].

The first setting, TDT, employs subject-independent training, development and test sets. It can be described as a subject-independent "1-fold-crossvalidation" or subject-independent holdout validation: It divides all available data in a training set, a development set and a test set, each containing data from strictly different subjects. The classifier is trained on the training set and its performance is repeatedly validated on the development set for hyperparameter optimisation. After reaching the best possible results, the classifier is tested on the test data in a final test to determine its performance on previously unseen samples. This setting offers the possibility to develop a system

that is very well suited to the classification problem at hand by optimising the hyperparameters, while still preventing the system from overfitting. However, this setting does not allow for statistical statements on the differences contained in the data by comparing the performance on individual subjects, since it always considers groups with several subjects. Therefore, the individual differences between the subjects are not covered. In other words, this setting is applicable when we want to make statements about the classification, not about the classified data.

These inter-individual differences can be assessed using the second setting, the LOSO setting. This setting can be described as a repetitive version of the TDT setting, or a subject-independent n-fold crossvalidation for n subjects: The classifier is trained on $n-1$ subjects and then tested on the remaining subject, repeating this procedure for all n subjects, which leads to n classifiers and thus n sets of performance measures. On the one hand, by averaging these performance measures and calculating the standard deviation over the subjects, we can make elaborated statements about the classification performance. On the other hand, it delivers n distinct classifiers, which leads to n sets of optimal hyperparameters and therefore n hyperparameter optimisation procedures. Since this is not always feasible, the classifiers are often used with default parameters, resulting in non-optimal classification conditions. Compared to the TDT setting, the LOSO setting enables us to make statements about the classified data, and not so much about the classification procedure.

Evaluation Metrics

Having established the correct evaluation procedure, we can attend to measuring the performance. For this, different *evaluation metrics* can be used. Before explaining these metrics, we should first look at the possible outcomes of a classification. We call a result a True Positive (TP), when an instance of class a is classified as class a. Accordingly, a True Negative (TN) is an instance of class $\neg a$ that is classified as $\neg a$. A False Positive (FP) is an instance of class $\neg a$ that is falsely classified as class a. A False Negative (FN) is an instance of class a that is falsely classified as class $\neg a$. In this sense, there are two possible sources for classification errors. The evaluation metrics are designed to measure these two error types. Recall R measures how many instances of class a were recalled or found by the classifier, evaluating the amount of FNs. Precision P describes how precisely the classifier works: It measures how many of the instances classified as class a are actually of this class, evaluating the amount of FPs. The metrics can be calculated in the following way [Sammut & Webb 2010]:

$$R = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}} \tag{2.18}$$

$$P = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}} \tag{2.19}$$

Naturally, there is a trade-off between these two metrics: If a classifier is optimised for high R of class a classifying as many instances of this class as possible (i.e. decreasing the number of FNs), it probably also will classify some instances of $\neg a$ as a (increasing the number of FPs), resulting in lower P, and vice versa. To balance the objectives of high R and high P, a third metric is introduced, the f-measure F_1, which is the harmonic mean of R and P [Sammut & Webb 2010]:

$$F_1 = 2 \cdot \frac{PR}{P + R} \tag{2.20}$$

These three metrics can be calculated for each of the classes separately for a detailed classification performance analysis. For a general overview of the performance for all n classes, we can calculate the metrics R_i, P_i and F_i for each class i and obtain the unweighted average of all n classes, resulting in the metrics Unweighted Average Recall (UAR), Unweighted Average Precision (UAP) and Unweighted Average F-Measure (UAF):

$$\text{UAR}_n = \frac{\sum_{i=1}^{n} R_i}{n} \tag{2.21}$$

$$\text{UAP}_n = \frac{\sum_{i=1}^{n} P_i}{n} \tag{2.22}$$

$$\text{UAF}_n = \frac{\sum_{i=1}^{n} F_i}{n} \tag{2.23}$$

2.3 Summary of the Chapter

In this chapter, we have discussed the concepts necessary for the understanding of the research presented in the following chapters. We have introduced the concepts related to the field of HHI and HCI and defined the interlocutor states we aim to investigate. Furthermore, we gained an understanding of classification methods including the main techniques for data pre-processing, classification and evaluation.

The next chapter is dedicated to another essential part of this thesis, namely the data that the presented research is based upon.

Chapter 3

The Thirst for Data

Contents

IT is obvious that data are highly important for data-driven modelling – without the right data, we cannot gain insight into the investigated phenomenon. In this chapter, we want to turn our attention to data-related aspects. As already stated in Section 1.1, we are interested in the recognition of three distinctive states of an interlocutor during an interaction – the state of *trouble*, the state of *satisfaction* and the state of *cooperativeness*. In order to investigate how these states can be modelled in a data-driven way, we need appropriate data. First, in Section 3.1, we analyse the data requirements and discuss what kind of data is necessary for the research questions presented in Section 1.4 and addressed in the following chapters. After that, in Sections 3.2 and 3.3, we introduce two corpora suitable for our investigations, the Last Minute Corpus and the Davero Corpus. The LMC will be used in Chapter 4 in order to investigate *trouble* in interaction – the internal state occurring when the expectations and the real course of interaction drift apart. We will also encounter it again in Chapter 5 – here, the state of *satisfaction* using the example of the satisfaction with the interaction participants' own performance will be the object of investigation. The DC will be used for a different question – in Chapter 6, we will work with it to analyse cooperative and competitive speech overlaps in a conversation.

3.1 General Requirements and Challenges

Modelling is usually based on pre-defined expectations concerning the observables: We have certain assumptions about the research subject and can use them for guiding the modelling process itself. In the case of data-driven analysis that we are interested in in this thesis, we rely on data solely. This poses several challenges. First, we need to ensure that we collect data that indeed reflect the research subject – the amount of the necessary data is directly related to the desired effect size, i.e. the degree of statistical significance of the investigated phenomenon. Secondly, the problem is that the acquisition of data always requires a certain amount of resources, especially if the acquired data must be post-processed manually, for example for synchronisation, transcription and annotation.

Since we are interested in certain states experienced during interaction – be it interaction between humans only or humans and computer systems – we also have to consider unknown observables. Such "unknown unknowns" might be critical to understand the relationship between the investigated phenomena and their cause. The best we can do is to cover a broad variety of characteristics of the interaction participants, such as sex, gender, cultural and educational background, age and skills, and so on, since these characteristics can influence the interlocutors' behaviour and attitude, and therefore the interaction itself [Whitley 1997; Cai et al. 2017].

Furthermore, in order to generate a data-driven model for a specific phenomenon, we need a sufficient amount of data reflecting this phenomenon. In other words, a good data set should ideally contain enough samples from each of the investigated phenomena and be balanced in order to avoid structural biases [Torralba & Efros 2011]. For rarely occurring phenomena, such as specific interaction states or the participants' personal traits, this might be a demanding task. Additionally, we need to reliably identify the assumed phenomena, which results in the question of defining the ground truth. We want to turn our attention to these challenges one by one.

3.1.1 Data Generation and "Ground Truth"

The generation of data is closely related to the question of defining the "ground truth", i.e. the real inner experience of the interlocutor. On the one hand, the interaction participants can be asked to "portray" a certain behaviour – this would make both, the generation and annotation of data arguably easy. On the other hand, this procedure would create artificial data, whereas the participants' behaviour under real-life conditions might be entirely different – not only in terms of visible differences such as in posed and spontaneous

smiles [Cohn & Schmidt 2013], but even in terms of different neural pathways [Hager & Ekman 1985].

From Chapter 1, we already know that the field of affective computing is transitioning from such acted scenarios to "in the wild" settings – therefore, we can no longer use prototypical acted representations. We must take another path by inducing the phenomenon we are interested in in the participants without revealing to them the purpose of the experiments. By doing so, we can be sure that their behaviour remains natural, at least as far as possible. But this poses another challenge – now we have to define and obtain the ground truth. This is especially difficult in natural settings, as the participants' behaviour might be not expressive enough to easily observe their internal state by external raters. If it is not possible to access it directly, we have to find other ways, for example by implementing a sophisticated experimental design and preparing detailed annotations.

Certain aspects of an interaction experiment can be ensured by design: The participants of the experiments can be "forced" to experience internal states such as stress and mental load by time constraints and difficult tasks as well as unexpected obstacles [Christodoulides 2017]. An example for such an experimental design is the LMC which we will discuss in detail in Section 3.2. It is also possible to use other means by drawing the user into the emotional state of the interaction partner [Douglas-Cowie et al. 2008] or "priming" the participants by music [Logeswaran & Bhattacharya 2009]. This also aids the annotation process – if certain states are ensured by the experimental design, these can be annotated directly by relying on the annotation of the triggering events. Otherwise, we have to rely on annotators. In the case of natural data with low expressiveness, the annotators have to be trained in a specific way to be able to spot the phenomena in question. Additionally, we need to verify the annotations by employing an "ensemble" of annotators – the final annotation can then be calculated using a majority vote. The inter-rater reliability, i.e. the degree of consensus between the annotators, is an important aspect often addressed in the literature [LeBreton & Senter 2008; Hallgren 2012; Siegert et al. 2014].

Another important issue during the data acquisition process is the quality of the recordings – for interaction data, this is especially difficult, since the recording equipment has to be integrated in the experimental design without overly disturbing the participants' experience. The main challenge is to keep the experimental conditions as natural as possible – since we want to capture natural behaviour – while ensuring the highest possible recording quality. Unfortunately, noise and other artefacts caused by the "in the wild" environment can impair the recognition performance of affect-related states [Lotz et al. 2018]. The topic of storing the data also deserves our attention. The

best method undoubtedly is to store the data in a raw, uncompressed way, since compression has an impact on the data and later recognition performance [Lotz et al. 2017]. But storing the data in this way is resource-demanding and has to be carefully planned.

3.1.2 Data Processing and Annotation

Gathering the right data is a demanding task, but it is still not enough to just collect the data – the processing of the data is at least as important.

In case of multimodal data, the synchronicity of the recorded modalities is an additional challenge. The usual way to achieve synchronous data streams in distributed networks is to align the clocks of the distributed sensors and to exchange specific messages [Sivrikaya & Yener 2004; Sundararaman et al. 2005]. If the synchronicity was not ensured at recording time, it can become technically impossible to establish it afterwards – one can imagine that to align the speech of an interaction participant to her gestures by hand is not an easy task. One solution for this is to use multimodal trigger events such as flash lights or hand claps [Bannach et al. 2009].

Furthermore, outliers[1] must be detected and handled appropriately. A suitable distribution of the characteristics we are interested in must be ensured, the metadata of the participants must be collected and stored in a retrievable way alongside the interaction data. In some cases, detailed transcriptions of the interaction experiments have to be prepared. This can be very expensive in terms of time and human resources, if the transcriptions are done manually. Using automatic speech recognition systems can facilitate this process by providing raw transcripts. Common speech recognitions systems such as provided by Google currently cannot process whole conversations with several interlocutors. Therefore, the transcripts must be pre-processed and segmented into utterances. Furthermore, the transcripts must be post-processed to remove transcription errors and artefacts – in the case of Google and Microsoft speech recognition, 9% and 18% word error rates are reported [Këpuska & Bohouta 2017]. This process requires trained transcribers.

The next important step is the annotation of the phenomena we want to investigate. As already mentioned above, especially for natural non-expressive ambiguous data peppered with noise, the annotation process can be very difficult. Finding the right way of annotation – for example choosing appropriate categories to be annotated and ensuring a sufficient number of annotators with a sufficiently high inter-rater reliability – is a question of ongoing research [Afzal & Robinson 2011; Metallinou & Narayanan 2013]. The usual

[1] In our case, outliers are interaction experiments that did not work as expected due to technical difficulties, etc.

way is to employ several annotators in order to reduce the subjectivity of the individuals – the same annotator often judges data differently depending on the current situation [Afzal & Robinson 2011]. Additionally, besides the duration of the annotation process itself, the training of the annotators also takes time. One possible way to make the annotation process less time-consuming is to rely on crowdsourcing [Park et al. 2014]. By employing multiple annotators, it is possible to filter noisy judgements and to achieve annotation results similar to those of experts [Nowak & Rüger 2010]. Nevertheless, introducing crowdsourcing means distributing highly sensitive data, which is a problem of data security.

Data security and protection of privacy is gaining recent attention, especially in the context of the European Union General Data Protection Regulation[2]. For academic research, anonymisation is a common method to ensure privacy protection, although it is not always applicable (e.g. for images of the face). One possible solution is to discard most of the raw, non-anonymised data and to rely on derived features only. But at the same time, especially in fields of ongoing research, it is often not clear what kind of data might be useful in the future. The right way to handle this issue is currently under discussion [Marelli & Testa 2018].

Some of the points mentioned in this section require careful planning before starting the data collection (such as the selection of interaction participants), others can be addressed afterwards (e.g. the training of the annotators). In general, data collection is a complex process deserving special treatment – the mentioned challenges are only a selection of all the challenges to be resolved during this process.

In the following sections, we will look at two previously mentioned data sets that will be used throughout this thesis, the LMC and the DC, and analyse how the requirements described above were met in these corpora.

3.2 The Last Minute Corpus

The data set used in Chapter 4 and Chapter 5 is the LMC. The LMC contains data recorded during interactions between human participants and a computer, which was implemented in a Wizard-of-Oz (WoZ) system, i.e. by an unseen human operator pretending that the computer acts automatically. During the interaction, the participants were seated in front of a computer system in comfortably furnished living-room-like surroundings. In a separate room, the WoZ system was operated by trained operators. The experiments were described in a variety of publications focussing on different aspects of the

[2] `http://eugdpr.org/`, retrieved December 15, 2019.

experimental design and experimental results [Rösner et al. 2012a; Rösner et al. 2012b; Frommer et al. 2012a; Frommer et al. 2012b].

In the course of the recorded interactions, the participating subjects traverse several experimental stages. In the first stage, the "Warm-up" stage, the subject introduces herself to the system by giving information on name, age, profession, recent positive and negative events, experience with technical systems, etc. In the second stage, the "Listing" stage, the subject gets introduced to the actual task: She has to imagine winning a trip to a location known as Waiuku and pack a suitcase for this journey under a rigorous time constraint. For this, various items can be chosen from a pre-defined list of categories (underwear, outerwear, sports equipment, etc.). After going through several categories, the subject learns that the suitcase must conform to weight restrictions, and has to remove several items in order to proceed. Since this poses the first major challenge, this stage is called the "Challenge" stage. The next stage is the "Waiuku" stage – here the subject encounters the next difficulty when the true destination of the trip is revealed. The trip is not a summer trip, as previously assumed by the subject, but a winter trip. Now the packing strategy has to be changed and the subject has to replace the summer items by winter items, making the weight constraint much more severe and further pressing for time. This stage is followed by the "Conclusion" stage, where the subject has to answer a series of questions on the overall experience, including her satisfaction with the packed suitcase.

The experiments were designed to induce high cognitive load, frustration and stress in the subjects by implementing the mentioned weight and time constraints as well as the destination change. The design guarantees that the subjects experience *trouble* in the interaction during the "Challenge" and the "Waiuku" stages – however, their behaviour does not necessarily express this in an obvious way. We will return to this issue in Section 4.2. The other three stages are seen as normal, non-troubled interaction. We will further elaborate on the definition of *trouble* and the differences between troubled and non-troubled stages in Section 4.2. The conclusion stage will be focussed on and further explained in Section 5.2. Table 3.1 presents an overview of all stages, their triggering events and activities.

Table 3.1: Overview of the dialogue stages, their triggers and tasks.

Stage	Trigger	Activity	Troubled?
Warm-up	Introduction request	Self-introduction	No
Listing	Winning the trip	Selecting items	No
Challenge	Weight constraint	Removing items	Yes
Waiuku	Revealing destination	Re-organising items	Yes
Conclusion	Concluding remarks	Experience evaluation	No

The fact that the subjects did experience the experiments in the expected way was ensured in questionnaires and semi-structured interviews [Rösner et al. 2012b]. This also facilitated the annotation of the interaction stages for the experiments, allowing to use the trigger events for the annotation of the stages [Prylipko et al. 2014]. The annotation of the stages is described in detail in Section 4.2.

The recordings were stored in their raw form in a non-anonymised way, including audio and video material as well as other synchronous modalities. The participating subjects gave their consent to this practice as long as the corpus is used for academic research only and not distributed to third parties.

There are two versions of the LMC featuring almost identical experimental design but slightly different recording conditions. These two versions will be referred to as LMCv1 and LMCv2 throughout the thesis.

3.2.1 The Last Minute Corpus Version 1

The LMCv1 comprises data from 133 HCI experiments. The subjects were nearly balanced regarding sex and two age groups (younger subjects under 30 and elderly subjects over 60). The recordings of the LMCv1 consist of acoustic data, physiological data and video data. The video data were captured via two stereo cameras (*Bumblebee2*[3]) and four HD cameras (*Pike F-145C*[4]). The acoustic data were recorded by two directional microphones (*Sennheiser ME 66*[5]) at 44.1 kHz. The quality of the recordings is not stable throughout the experiments, with only 89 subjects achieving a sufficient audio quality. The physiological data were recorded using the *NeXus-32*[6] system and include electromyogram, skin conductivity and respiration data. These physiological signals were recorded for 20 of the 133 subjects, only 19 of which have also appropriate audio recordings.

An overview of the sex and age distribution as well as the duration of the data for the LMCv1 as a whole, and for audio and physiological data subsets is given in Table 3.2.

For the collected data, detailed transcriptions were prepared. Furthermore, the trigger events of the different stages are annotated as well as other phenomena such as speaking behaviour (pauses, breathing, laughing) and offtalk.

[3] `http://www.flir.com/products/bumblebee2-firewire/`, retrieved December 15, 2019.

[4] `http://www.alliedvision.com/en/products/cameras/detail/Pike/F-145.html`, retrieved December 15, 2019.

[5] `http://en-uk.sennheiser.com/directional-microphone-shotgun-film-broadcast-me-66`, retrieved December 15, 2019.

[6] `http://www.mindmedia.com/en/products/nexus-32/`, retrieved December 15, 2019.

Table 3.2: Subject characteristics and data duration in different subsets of the LMCv1.

Data Set	Sex		Age		Duration	
	female	male	elderly	young	total	per subject
Full data set	70	63	61	72	59.0 h	26.6 ± 3.5 min
Acoustic data	49	40	45	44	40.1 h	27.0 ± 3.8 min
Physiological data	9	10	8	11	8.4 h	26.5 ± 4.1 min

3.2.2 The Last Minute Corpus Version 2

The LMCv2 comprises data from 63 subjects, all of them students between 18 and 33 years old. As already mentioned, the differences in the experimental design in comparison to the LMCv1 can be neglected for the scope of the investigations described in this thesis. The experiments feature the same stages with the same tasks to be fulfilled by the subjects.

The main difference is that the subjects had to fill out three questionnaires on their interaction experience, one of them during the "Listing" stage, which resulted in two consecutive parts of this stage. Another difference is that in half of the experiments of the LMCv2, a more natural text-to-speech system was used for the WoZ system [Ferchow et al. 2016].

The recordings contain three data streams: Video (obtained by the same two stereo and four HD cameras as in the LMCv1), audio (obtained by the same two directional microphones as in the LMCv1 and one headset microphone), and 3D data obtained by a Kinect 2 sensor. Unfortunately, only 53 of the recordings are synchronous and thus can be used for further experiments. An overview of the subject distribution and data duration can be found in Table 3.3.

Table 3.3: Subject characteristics and data duration in different subsets of the LMCv2.

Data Set	Sex		Duration	
	female	male	total	per subject
All available data	30	33	24.6 h	22.7 ± 3.1 min
Synchronous data	25	28	19.5 h	22.1 ± 2.9 min

3.3 The Davero Corpus

The DC is a collection of real-life telephone-based human-human dyadic conversations recorded in a German call centre which offers telephone support for a large power supplier [Siegert & Ohnemus 2015; Siegert et al. 2015a]. These dialogues include different interaction styles and revolve around different topics, such as complaints, informational calls, etc. The recordings were conducted

in a single channel containing the speech of both, the agent and the client via telephone. Although the agents' utterances were recorded directly from their headsets, the clients were using common telephones, leading to a high amount of noise with partly incomprehensible speech. In contrast to the LMC, the DC does not feature a specific experimental design but rather a loose setting, since all conversations are centred around a problem to be solved during the call.

The part of the DC used in this thesis contains the recordings of seven days. Each day comprises the dialogues between exactly one of four different agents with a varying number of clients, with 48 dialogues recorded over the seven days. A detailed overview is given in Table 3.4. This part's total duration is 270 minutes, with an average dialogue duration of 5.6 ± 3.3 minutes.

Table 3.4: Overview of the subject distribution in the DC. For each of the days (D1 – D7), the number of dialogues (#Dialogues) and the sex of the agent and the client are shown.

Day	**#Dialogues**	**Sex Agent**		**Sex Client**	
		male	female	male	female
D1	9	9	-	3	6
D2	6	6	-	0	6
D3	5	-	5	0	5
D4	3	-	3	1	2
D5	14	14	-	7	7
D6	8	8	-	7	1
D7	3	-	3	0	3
Overall	48	37	11	18	30

The annotation of the data includes speaker turns, but there are no detailed transcriptions available. Additionally, there are annotations of the emotions expressed in the utterances. The utterances of each speaker are labelled in the emotional dimensions of valence and control, indicating an increase or a decrease in these dimensions compared to the previous utterance of the same speaker. For example, "C+" indicates an increase in the control dimension, "V-" indicates a decrease in the valence dimension, "C0" and "V0" indicate no change.

These annotations were prepared by five trained annotators according to the Geneva Emotion Wheel [Scherer 2005]. The training process for the annotators consisted of three steps. In the first step, the annotators got familiar with the concept of categorial labelling using examples from acted emotional data. This labelling was then extended to the multidimensional approach of the Emotion Wheel. In the third step, the annotators were presented with selected excerpts from the DC previously rated by expert psychologists. The five annotators worked together and discussed the annotation results. Despite this training

process, the inter-rater reliability for the annotations in terms of Krippendorff's α [Hayes & Krippendorff 2007] lies between 0.26 for the control dimension and 0.34 for the valence dimension. Nevertheless, it is consistent with the expectations for such natural material and comparable to those of widely used corpora [Siegert et al. 2014].

The final annotations were obtained using a majority vote among the annotators. Since not all votings resulted in a majority, this procedure led to a sparse annotation, where the utterances without a majority for a label remained unlabelled. In the part of the DC that was used for the investigations described in this thesis, the amount of missing values was around 17%.

The recordings are stored in an anonymised way, with all personal information on the participating subjects removed.

3.4 Summary of the Chapter

In this chapter, we have analysed what challenges are posed by data-driven modelling and discussed the requirements of the "thirst for data".

Furthermore, we have seen how these requirements are dealt with by two very different corpora: The LMC, a WoZ-driven naturalistic yet highly controlled HCI data set in two different versions, and the DC, a real-life HHI data set recorded in a call centre setting. The former was recorded under laboratory conditions, resulting in high-quality recordings. Although the immersion grade of the subjects was confirmed by interviews and questionnaires, they might have behaved differently in an "in the wild" environment, undisturbed by the recording conditions and the surroundings. In contrast, the latter data set was recorded "in the wild", with a high amount of noise and telephone-induced artefacts – this posed a challenge for the annotation, resulting in low inter-rater agreement, as well as for further processing.

When comparing these data sets, we should briefly return to the general requirements introduced in Section 3.1. The first important point was to provide a fair amount of data with a sufficiently diverse distribution of the participating subjects' characteristics such as age and sex as well as cultural and educational background. All considered data sets contain data collected from dozens of subjects with various backgrounds. Especially for LMCv1, both sex and age distribution were accounted for by subject selection. Furthermore, the differences in personality traits and backgrounds were observed in the collected questionnaires. In the case of LMCv2, the participants were mostly young (under 33 years old) students – in this case, the variations between the subjects is indeed not as high as in the general population. For the DC, we can assume that the subjects cover a broad range of sex, age and other characteristics,

since the subjects are randomly chosen from a large power supplier's client base. Regarding the data generation process and the related inducing of the desired phenomena as well as their annotation, the LMC is very different from the DC. Both versions of the LMC were recorded specifically to investigate interaction-related behaviour in a pre-defined scenario. The DC contains unconstrained real-life interactions, the only common component being the call centre setting. Therefore, in order to obtain a reliable data base containing the desired information, it was necessary to perform a time-consuming annotation process with the help of trained annotators.

In the following three chapters, we will analyse the three interlocutor states of *trouble*, *satisfaction* and *cooperativeness* based on the introduced data. We will start with the state of *trouble* in the next chapter.

Chapter 4

Detecting Trouble in Interaction

Contents

As technical systems are applied to more areas of everyday life, it becomes increasingly important to attend to certain human behaviour and experience patterns occurring in natural, everyday interaction. In this chapter, we turn our attention to challenging interaction in order to detect those parts of the interaction that can impair the overall experience by "causing trouble". As already defined in Section 2.1.1, challenging interaction segments "challenge" the human counterpart of an interaction by not satisfying the previously raised expectations – in other words, these segments create a mismatch between expectation and reality, and induce a complex affective state similar to dissatisfaction, irritation and frustration that we will refer to as *trouble* in this chapter.

This chapter introduces methods to detect this interlocutor state using three different modalities: acoustic data, biophysiological data and 3D data. First

of all, we need to take a closer look at these modalities and their relevance for the recognition of internal states – such as *trouble* – as well as existing works in this field in Section 4.1. After that, we will learn how *trouble* can be induced using the example of the Last Minute Corpus in Section 4.2. Finally, we will examine the suitability of the three mentioned modalities for the task of *trouble* recognition one by one in Sections 4.3 to 4.5 and compare their advantages and disadvantages in Section 4.6. Section 4.7 summarises the chapter.

This chapter is mainly based on work published by my colleagues and myself [Egorow & Wendemuth 2016; Böck et al. 2017b; Böck et al. 2018].

4.1 Existing Works on Multimodal Trouble Recognition

As we already know from Section 1.1, emotions are part of the human nature and a highly studied aspect in psychology [Plutchik 2001]. Although they are often associated especially with the face and the voice, emotional or affective states are bodily states that influence physiological parameters of the experiencing person, e.g. the heart rate, sweat production, etc. In fact, this phenomenon is well reflected in our language: We experience "butterflies in the stomach" when we are in love, we blush when we are ashamed, and our heart races and the palms get sweaty when we are anxious. These relationships have been known for a long time [Ekman et al. 1983; Levenson et al. 1990; Cacioppo et al. 2000]. In Section 1.2, we have briefly touched the topic of affective state expression and recognition in different modalities. We will now analyse this topic in more detail, before turning to the task of *trouble* recognition itself.

4.1.1 Conveying Internal States with Different Modalities

In Section 1.1, we have already discussed the field of automatic emotion recognition in general, and especially for the problems of "in the wild" scenarios, which have been gaining interest recently. Using audio-only recognition and classifiers such as SVM and Hidden Markov Models, natural emotion recognition could achieve only low results of around 30% UAR for the four-dimensional valence-arousal space [Schuller et al. 2009]. In the seven-class spontaneous data set EmotiW14, results of around 32% UAR for audio-only recognition are reported [Ringeval et al. 2014].

On the one hand, the classification performance could be boosted by further fine-tuning and introducing advanced techniques such as deep learning – but

these techniques have yet to prove their usefulness. So far, accuracies of around 50% for three-class problems to up to around 64% for four-class problems have been reported [Chang & Scherer 2017; Heracleous et al. 2019]. On the other hand, we should not forget that internal states are not only conveyed by speech. As already mentioned, emotional states are experienced by the whole body. Therefore, further modalities need to be considered. For example, adding the dimension of co-speech gestures improves the results by 2% absolute compared to speech alone [Böck et al. 2014]. Besides gestures and facial expressions, the bodily dimension also includes physiological data, such as heart rate, skin conductivity and other physiological signals. One advantage of this kind of data is that it can be obtained even if the interacting person is not actively using speech or gestures. Furthermore, physiological reactions are very difficult to conceal, since most humans cannot actively control their heart rate and other physiological signals. Therefore, automatic recognition of internal states from such signals promises better results than from other signals, which is supported by various investigations. For categorial emotions, the recognition rate of approaches based on physiological signals ranges from 61% for four-class recognition to 83% for two-class recognition of induced emotions [Kim et al. 2004; Rainville et al. 2006]. For the multidimensional valence-arousal scale, recognition rates of over 90% for low and high valence and arousal are reported [Haag et al. 2004].

Even more interestingly, physiological data can be combined with acoustic data for the recognition of emotional states [Kim 2007]. In this particular setting, the interaction revolved around a WoZ quiz-show consisting of four stages, each of them corresponding to high / low valence, and high / low arousal, respectively. Six physiological parameters as well as acoustic feedback of the participating subjects were recorded and used for emotion recognition. The classification results were rather mediocre achieving around 55% accuracy, but combining acoustic and physiological data lead to improvements of up to 5% absolute in all cases [Kim 2007].

Unfortunately, using physiological signals is not an easy task, since obtaining such data is a difficult process. The first problem is that it often requires physical contact. There are solutions enabling us to obtain heart rate data in a contactless way, as we will see below, but for other physiological signals, the appropriate methods are still to be found. We have already discussed two challenges related to this in Section 3.1.2 – data security as well as technical difficulties to synchronise different data streams. But even more severe, the sensors can cause disturbance and affect the naturalness of the interaction. Even if we neglect the difficulties caused by intrusive electrodes, not every user is willing to provide such sensitive data. Therefore, physiological signals, although certainly useful, are not always accessible. Regarding multimodal

processing in general, it should be noted that, although promising, the combination of acoustic and physiological data is also difficult, since, as already mentioned, collecting and processing synchronous multimodal data poses certain challenges [Freitas et al. 2014]. Another open question in this regard is the search for suitable features that we have already discussed in detail in Section 1.2.2.

But physiological signals are not the only useful modality besides speech when it comes to the recognition of emotional or interaction-related states. As already mentioned, internal states such as emotions are experienced by the whole body, which allows considering the human body as an additional modality for the recognition of affect. In fact, in some cases body postures are even better suited to this task than other modalities: They are obtainable from a distance and mostly not consciously censored, acting as a "leaky" source of information on the internal state [De Gelder 2009]. Furthermore, emotions like anger and fear are clearly expressed only by the whole body [De Gelder 2009]. This is connected to the annotation process that we discussed in Chapter 3, since using additional modalities might facilitate it. Interestingly, observers are able to recognise emotions from the whole body [De Meijer 1989] or hand and arm movements [Wallbott 1998], with recognition results comparable to voice [Coulson 2004]. Furthermore, some bodily movements seem to be universal: There is evidence that sighted, blind and congenitally blind individuals display the same pride and shame behaviours [Tracy & Matsumoto 2008].

There are several interesting approaches that use bodily movements and postures for the recognition of emotions and interaction-related affective states. For example, using a combination of facial expression analysis and affective gesture analysis, surprise can be recognised with an accuracy of 85% [Balomenos et al. 2005]. Gestural features for emotion recognition were also investigated as a single modality for discriminating emotions in the pleasure-arousal-dominance space, finding eight significant relationships such as between pleasure and handedness [Kipp & Martin 2009]. Even the same gesture is performed differently by participants if they express different emotional conditions: Expressive motion cues can be used to discriminate in both, the arousal as well as the valence dimension [Castellano et al. 2007]. Further investigations on this topic are covered in extensive surveys [Zacharatos et al. 2014; Noroozi et al. 2018]. There are also some approaches for automatic discrimination of more natural affective states based on the Laban Movement Analysis, a gesture and movement description system – for instance recognition of concentration and excitement in game scenarios [Zacharatos et al. 2013], of various emotional states during musical performances [Truong et al. 2016], and theatre performances [Senecal et al. 2016].

The presented works show that the recognition of internal interlocutor states has been studied from multiple perspectives. Different modalities pose specific challenges in terms of signal availability, robustness, etc., but each modalitiy also presents a unique angle to accessing the interlocutor. After having reviewed the available approaches, we will now evaluate how these different modalities were used for *trouble* recognition in previous research.

4.1.2 Indicators of Trouble in Interaction

The recognition of *trouble* during interaction is an important dialogue aspect – for HCI and most user-centred applications, the automatic recognition of the user experiencing difficulties is a valuable input. Nonetheless, challenging parts of an interaction have not received much attention so far.

In HHI, the mismatch between the expected and real course of an interaction is normally resolved by using prosodic features such as intonational variation and pitch contours [Hirschberg 2002]. Pitch, amplitude and timing can be used as signals for misrecognised or misheard utterances [Hirschberg et al. 1999]. Prosodic peculiarities, such as increased pitch and loudness, are used as a cue to distinguish between conventional and surprised questions, signalling "a problem of expectation which requires special treatment" [Selting 1996]. Such signals should be seen as requests for clarification [Gehle et al. 2014].

The behaviour during challenging interaction with computer systems also has been in the focus of research, for example in a WoZ scenario [Batliner et al. 2003] similar to the LMC scenario described in Section 3.2. The aim of this investigation was to capture features that might be used as signals for trouble in communication. In one of the investigated scenarios, the users of a WoZ-system were confronted with several malfunctions of the system. Using acoustic data and detailed annotations containing part-of-speech tagging, dialogue acts, repetitions and syntactic boundaries, the authors developed a system for automatic "trouble recognition" with around 73% average recall for subject-independent classification. Furthermore, they argued for combining several information sources in order to achieve acceptable results for real-life applications.

Another interesting feature that has been in the focus of trouble detection is silence after certain speech acts [Sacks et al. 1974]. A duration of approximately one second seems to signal problematic moments [Jefferson 1989]. The duration of inter-turn silences affects the perception of the interactional tone – human raters judge the willingness and agreement of the interlocutors as declining when the duration of inter-turn silence increases [Roberts et al. 2006]. Another finding in this domain was reported for "no" responses – pauses, duration, intonation contour and pitch range of the response are used as signals

for problems in the preceding bit of communication [Krahmer et al. 2002]. In dialogues, prosodic and temporal features (delays, rising and falling boundary tones, pitch, etc.) are used by a speaker to signal whether she is able to follow the other speaker [Shimojima et al. 2002].

However, speech prosody and temporal cues are not the only features used for trouble signalling. In fact, interlocutors use multiple channels of behaviour when experiencing *trouble*, such as speech, gaze and posture directly after trouble occurs [Gehle et al. 2014]. When the interaction participants do not want to talk about the difficulties, irritation and frustration experienced during the interaction with a technical system openly, involuntary reactions such as physiological signals gain more relevance [Picard et al. 2001].

In the next section, we will closely examine the data used throughout this chapter, namely both versions of the LMC already described in Section 3.2, and analyse in detail how the interlocutor state of *trouble* can be induced in close-to-real-life conditions.

4.2 Inducing Trouble in the Last Minute Corpus

In the following investigations, both versions of the LMC (LMCv1 and LMCv2) described in Section 3.2 were used. This data set is suitable for the analysis of *trouble*, since it contains different interaction stages with increasingly and unexpectedly difficult tasks – the experiments were designed especially to induce the challenging interaction that we are interested in.

For the investigation presented here, the most important aspect of the LMC is the division of the data into the *baseline* class and the *trouble* class. The former denotes normal interaction without any problems. The latter denotes the problematic interaction with unexpected challenges that the subjects are facing during the second half of the interaction experiments.

As we already know from Section 3.2, the subjects had to interact with a WoZ system and prepare a suitcase for a trip to an unknown location. The interaction experiments consisted of five consecutive stages with different tasks that had to be solved. The "Warm-up" and "Listing" stages represent non-challenging or *baseline* class: Here the subjects had to introduce themselves and pick some items from several categories. The "Challenge" and "Waiuku" stages, where the subjects were confronted with the weight limit and an unexpected travelling direction respectively, represent the *trouble* class.

To illustrate the difference between the two classes, several excerpts from the LMCv1 are presented below[1].

In the first example, the wizard (W) introduces a new item category and the subject (S) selects an item from the list.

> W: You may now select from the category "Tops". If you have finished the selection from this category, please say that you want to move to the next category.
>
> S: T-shirt ((pause)) five
>
> W: Five t-shirts were added, you can continue.

This is a paramount example of a normal, untroubled communication between the two interaction partners.

In the second example, a subject – an elderly woman – is for the first time informed by the wizard that her suitcase has reached the weight limit:

> W: A swimsuit or bikini cannot be added, otherwise the maximum weight limit of the suitcase prescribed by the airline would be exceeded. Before further items can be selected, you must provide enough space in the suitcase. For this purpose, already packed items can be unpacked again. Upon request you will receive a list of already selected items.
>
> S: Yes uh ((pause)) I would like ((pause)) take out a pair of shoes.
>
> W: Your statement cannot be processed.

The pauses now occurring in the subject's utterances are a first hint that she is surprised by the wizard's prompt – the weight limit was not mentioned previously, therefore it is an unexpected obstacle. At the same time, the wizard also provides a solution to the problem by introducing the possibility of unpacking items. But the subject fails to do so, since she does not use the appropriate wording. This is done by design to induce challenging interaction – as the wizard does not react to the subject's wrongly worded request, it causes more confusion. In the following interaction segment, we see that the subject is becoming increasingly dissatisfied with the course of the interaction and is also experiencing more *trouble*.

> W: Before further items can be selected, you must provide enough space in the suitcase. For this purpose, already packed items can be unpacked again. Upon request you will receive a list of already selected items.

[1]The German transcripts were kindly provided by Prof. D. Rösner and his team, and translated by myself.

> S: Hm ((moaning)) yes ((pause)) then I want to hear the chosen items again please ((pause)) I told you I want to unpack shoes.

But it also should be stated that the problems occurring during the experiments do not impact all subjects in the same way. The next example illustrates the same part of the interaction between the wizard and another subject – this time a young woman, who seems to be less affected.

> W: A swimsuit or bikini cannot be added, otherwise the maximum weight limit of the suitcase prescribed by the airline would be exceeded. Before further items can be selected, you must provide enough space in the suitcase. For this purpose, already packed items can be unpacked again. Upon request you will receive a list of already selected items.
>
> S: Remove inflatable boat.
>
> W: One inflatable boat was removed, you can continue.
>
> S: ((smacks)) three bikinis ((swallows))

Here, we see that the subject managed to cope with the occurring problem and to continue fulfilling the task. Since there are no perception tests available, it cannot be ensured that all subjects experienced the same problems during the challenging interaction stages. Nevertheless, the experiments were designed to induce the described state of the subjects. Therefore, most of the subjects were experiencing the interaction as challenging – however, the example above demonstrates that not all of them expressed it in their behaviour to the same extent.

Since we obtained a better understanding of the data, we can proceed to analyse how the state of *trouble* can be detected using the three previously mentioned modalities. We will deal with the subjects' speech, physiological and 3D data one by one.

4.3 Detecting Trouble with Speech Data

As this thesis focuses on spoken communication, it is only natural to investigate the speech modality for *trouble* recognition. First, we take a look at the statistical differences in the acoustic features that can be found between the spoken content of non-challenging and challenging interaction in Section 4.3.1. Second, we analyse whether these differences can be used for the classification of *baseline* and *trouble* in Section 4.3.2.

4.3.1 Analysing the Statistical Differences

The investigation described below was first published by my colleagues and myself in [Böck et al. 2017b].

This investigation addressed the question whether there are differences in the acoustic features between different interaction stages of the LMC. It was conducted on a subset of the LMCv1 – this subset contains 18 subjects that also participated in further experiments [Rösner et al. 2016].

The investigated features are based on the well-known *emobase* feature set described in Section 2.2.4. The features were computed by the openSMILE feature extraction toolkit [Eyben et al. 2010]. As previously described, the full *emobase* feature set consists of 988 features derived from 26 LLDs such as MFCCs, Linear Spectral Pairs (LSPs), intensity and loudness. In the investigation presented here, only the mean values of the 26 LLDs and their first order derivatives (52 features in total), were used, since we focussed on general trends in the underlying features themselves instead of their derivatives. The features were extracted on utterance level and averaged over all utterances of the challenging and the non-challenging stages, resulting in one value per stage. In order to be able to compare these average values, a comparative measure D was introduced, which can be calculated for each feature mean value f_i for the two classes *baseline* (denoted as f_{i_b}) and *trouble* (denoted as f_{i_t}) in the following way:

$$D_{f_i} = f_{i_b} - f_{i_t} \tag{4.1}$$

This measure delivered an interpretable result and enabled us to directly assess the changes in the feature values.

Using the measure D, we calculated Person's correlation coefficients described in Section 2.2.2 for the two stages *baseline* and *trouble*. We found that 16 of the 52 features show significant differences between the *baseline* class and the *trouble* class – 7 of these features are MFCC-related, which corresponds to their relevance for the classification of affect already established for a long time [Kwon et al. 2003; Sato & Obuchi 2007], even in current deep learning architectures [Heracleous et al. 2019].

In order to expand on these results, the statistical analysis was conducted again using Pearson's correlation coefficients on all available LMCv1 data by my colleagues and myself [Böck et al. 2018]. Here, we extended the data subset to all 89 LMCv1 subjects and analysed the differences in the feature values between the two stages in the same manner as in the previous investigation. We identified 29 features where the number of changes across subjects was higher than the standard variation σ (this feature set is further referred to as

F_{29}). For 14 features, this number was higher than $2.5 \cdot \sigma$ (this feature set is further referred to as F_{14}).

All features from F_{29} are presented in Table 4.1. The features from F_{14} are highlighted. The most interesting features again appear to be MFCCs – they represent over one third of the F_{14} set. Besides confirming our previous findings reported in [Böck et al. 2017b], this corresponds to the relevance of these features known from the literature, as already mentioned above. For the F_{29} set, LSPs constitute almost half of the set. This also supports the relevance of spectral features for acoustic discrimination tasks.

Table 4.1: Features with remarkable differences between the two stages. All listed features constitute the F_{29} set, the highlighted features are highly remarkable and constitute the F_{14} set. The table is adapted from [Böck et al. 2018].

Energy-related	**MFCCs**	**LSPs**	**Other**
Intensity	MFCC1	LSP0	F0
ΔIntensity	MFCC2	LSP1	
	MFCC3	LSP2	
Loudness	MFCC4	LSP3	
ΔLoudness	MFCC6	LSP4	
	MFCC12	LSP5	
Zero-crossing-rate		LSP6	
	ΔMFCC1		
	ΔMFCC2	ΔLSP1	
	ΔMFCC9	ΔLSP2	
	ΔMFCC12	ΔLSP3	
		ΔLSP4	
		ΔLSP6	
		ΔLSP7	

4.3.2 From Statistical Differences to Classification

In the previous subsection, we have seen that features derived from the *emo-base* feature set show significant differences between baseline and challenging interaction stages. The next step is to investigate whether these differences are sufficient for the classification of the *trouble*.

The investigation described in this subsection was reported in [Egorow & Wendemuth 2016]. Again, a subset of the LMCv1 was used, consisting of 19 subjects. These recordings were selected, since they contained not only acoustic data but also physiological data, which will be covered in Section 4.4.

For the task of *trouble* recognition, the same two classes are introduced as discussed above. The *baseline* class contains the "Warm-up" stage, the

"Listing" stage and the "Conclusion" stage, and corresponds to normal, non-challenging interaction. The *trouble* class containing the "Challenge" stage and the "Waiuku" stage corresponds to the challenging interaction stages.

For the classification task, again the previously mentioned *emobase* feature set was employed – this time, the complete set with all 988 acoustic features, as described in Section 2.2.4. As specified in the *emobase* documentation [Eyben et al. 2010], the LLDs were extracted on a 25 ms window with a 10 ms shift, the derived statistical features were calculated on utterance level. As a last pre-processing step, the feature values were standardised ($\mu = 0$, $\sigma = 1$) using the procedure described in Section 2.2.4.

For the subsequent classification, a RF classifier as described in Section 2.2.6 was employed, in the implementation provided by Weka [Frank et al. 2016]. The implemented version was chosen due to its high processing speed as well as low complexity, including only two hyperparameters: the number of features used in each node (further referred to as n_f) and the number of trees (further referred to as n_t).

The evaluation was performed in a TDT setting as introduced in Section 2.2.7. The subjects were divided into three disjoint groups: a training set containing eleven subjects, and development and test sets containing four subjects each. The sex and age distribution of the subjects over these sets is illustrated in Table 4.2.

Table 4.2: Distribution of sex and age of the subjects over the training, development and test sets. The table is adapted from [Egorow & Wendemuth 2016].

Characteristic	Training	Development	Test	Overall
Sex				
– Female	5	2	2	9
– Male	6	2	2	10
Age				
– Young (< 30)	7	2	2	11
– Elder (> 60)	4	2	2	8

The training was performed on the training set, whereas the development set was used to derive the optimal hyperparameters in a grid search procedure. The test set was used for the final evaluation with the hyperparameters determined on the development set. During the hyperparameter optimisation procedure, the best hyperparameters were found to be $n_f = 6$ and $n_t = 30$.

The classification results are presented in Table 4.3 in terms of UAR, UAP and UAF. We see that the results differ between the development set and the test set – especially regarding the recall of the *trouble* class and the precision of the *baseline* class. However, the UAF achieved on the development set is

only 3% absolute higher than the UAF achieved on the test set, indicating that the classifier is able to generalise. Nevertheless, the classification outperforms chance level by only around 14%, which is rather dissatisfying. These results lead to the conclusion that the classification using acoustic data alone is not applicable for real-life systems.

Table 4.3: Classification performance on acoustic data for the two classes *trouble* and *baseline* and their unweighted average (UA). The results are shown for the development and the test set (highlighted) in terms of recall, precision and f-measure. The table is adapted from [Egorow & Wendemuth 2016].

Test Condition	Recall	Precision	F-Measure
Development			
– Trouble	76.64	51.25	61.42
– Baseline	62.68	83.97	71.78
– UA	69.66	67.61	66.60
Test			
– Trouble	66.43	57.76	61.79
– Baseline	62.01	70.25	65.88
– UA	64.22	64.01	63.83

Similar results are reported by my colleagues and myself, where we have shown that with five different feature sets and two classifiers, an UAR between 56.4% and 65.6% can be achieved for the same classification task on all 89 subjects of the LMCv1 – now using LOSO evaluation [Böck et al. 2018]. In this extended investigation, we focussed on the comparison of different feature sets. In total, we used five feature sets: the complete *emobase* feature set as described above, the reduced feature sets F_{14} and F_{29} introduced in Section 4.3.1 containing only 14 and 29 remarkable features, respectively, and two new feature sets F_{409} and F_{700}. F_{409} contains 409 features derived from F_{29} by calculating the functionals (e.g. the minimum and maximum value, skewness and kurtosis, etc.) applied in the well-known *GeMAPS* feature set [Eyben et al. 2015]. This procedure allowed us to implement a compromise between the low number of remarkable features and the high number of the *emobase* features. Furthermore, including the functionals introduces additional information on the temporal changes in the data. F_{700} contains 700 features and is originally developed for the task of addressee detection [Siegert et al. 2018].

The results of the comparison of the feature sets are presented in Table 4.4. For each of the five feature sets and each of the two classifiers, we see the average and standard deviation of each of the employed performance measures UAR, UAP and UAF over all subjects, indicated as $\overline{\text{UAR}}$, $\overline{\text{UAP}}$ and $\overline{\text{UAF}}$. The last two rows show the average performance for a subject: We see the

average and standard deviation over the results obtained on all features by a classifier, again averaged over all subjects.

Table 4.4: Classification performance of all 89 LMCv1 subjects with SVM and RF using the five different feature sets. Furthermore, the average performance of a classifier over all feature sets is shown. The worst and the best results in terms of $\overline{\text{UAF}}$ are highlighted. The table is adapted from [Böck et al. 2018].

Condition	$\overline{\text{UAR}}$	$\overline{\text{UAP}}$	$\overline{\text{UAF}}$
F_{14}			
– SVM	60.38 ± 12.31	59.09 ± 10.90	56.75 ± 11.78
– RF	56.31 ± 10.49	56.16 ± 10.04	54.64 ± 10.25
F_{29}			
– SVM	59.70 ± 11.26	59.49 ± 11.07	57.98 ± 11.37
– RF	56.80 ± 11.45	56.40 ± 10.94	54.74 ± 11.12
F_{409}			
– SVM	56.81 ± 11.00	57.18 ± 11.62	55.77 ± 11.05
– RF	64.53 ± 12.50	63.96 ± 11.77	61.56 ± 13.21
F_{700}			
– SVM	57.28 ± 11.82	57.38 ± 12.12	55.93 ± 11.81
– RF	64.91 ± 13.10	63.64 ± 11.91	61.04 ± 13.53
emobase			
– SVM	56.11 ± 12.03	56.34 ± 12.28	55.06 ± 12.16
– RF	63.37 ± 12.39	61.93 ± 11.80	59.05 ± 13.69
Average			
– SVM	58.07 ± 7.39	58.37 ± 7.58	56.43 ± 7.50
– RF	59.53 ± 7.89	60.35 ± 8.24	57.37 ± 8.64

Interestingly, the classification experiments show that there is no clear relationship between the feature sets and the classification performance – for neither of the two classifiers. Furthermore, the results indicate that the differences between subjects have a greater influence on the classification than the employed feature set and classifier, since the standard deviation between the different subjects of 10% to 14% shown in the upper part of Table 4.4 is higher than the standard deviation between the different classification procedures of around 7% to 8% shown in the last two rows of Table 4.4.

Overall, we can state that the findings presented in [Böck et al. 2018] correspond to the results obtained on the LMCv1 subset with 19 subjects described above, where the classification achieved around 64% UAF – this also confirms that the investigated subset represents the whole LMCv1 in an appropriate way.

Although these classification results are rather disappointing, they bear an important message: It is not possible to recognise certain user states on acous-

tic features alone in a robust way. Building on this, we turn to other modalities in the two following sections.

4.4 Detecting Trouble with Physiological Data

As we have seen in the previous section, acoustic features alone do not deliver satisfying results for the discrimination of non-challenging and challenging interaction. In this section, we want to consider a different modality, namely biophysiological data, also referred to as physiological signals or biosignals. One advantage of this kind of data compared to data obtained from speech is that it cannot be influenced by the human interaction partner and therefore corresponds to the "ground truth" [Kim et al. 2004; Andreassi 2010]. At the same time, this kind of data is also more difficult to obtain, since it requires intrusive sensors, whereas acoustic data can be collected from a distance.

This section is mainly based on the investigation reported in [Egorow & Wendemuth 2016]. In this investigation, we performed the classification experiments on the same 19 subjects of the LMCv1 as in Section 4.3, with the same training, development and test subsets as shown in Table 4.2.

4.4.1 From Physiological Signals to Physiological Features

The features used for this investigation are based on three original physiological signals: Electromyogram(EMG), Respiration(RSP) and Skin Conductivity(SC). From the different signals provided by the used *NeXus-32* system, these three were chosen since they could be obtained in consistently good quality throughout the interaction experiments. The features are calculated using the Augsburg Biosignal toolbox[2].

The original signals were pre-processed by a lowpass filter and then normalised, resulting in a total number of 104 features, such as first and second order derivatives and statistical measures (mean, standard deviation, etc.) of the original signal. The features were calculated at a sampling rate of 32 Hz. The employed feature set was originally developed for the recognition of affective states such as four emotional states (joy, anger, sadness, pleasure) and mental stress [Wagner et al. 2005; Wagner 2009].

[2] `http://www.informatik.uni-augsburg.de/lehrstuehle/hcm/projects/tools/aubt/`, retrieved December 15, 2019.

4.4.2 Trouble Classification on Physiological Features

For the classification, we used the same setup as for the acoustic features. Again, a RF classifier was employed. In order to maintain comparability, we performed the training on the training set containing the data from the same subjects as described above, the hyperparameters of the classifier were selected using the development set, the final evaluation is performed on the test set. For this setup, the best hyperparameters were found to be $k = 3$ and $n = 10$.

The results of the classification are shown in Table 4.5. We can see that the physiological features provide a remarkable improvement compared to the acoustic features: The achieved UAF is 91% for the development data and 90% for the test data. It should also be noted that the difference in performance in terms of UAF between the two sets is only 1% absolute, which confirms the generalisation ability of the classifier. In fact, this performance can be interpreted as a hint that the physiological signals convey the ground truth in an almost optimal way: Taking into account that not all participants might have experienced *trouble* to the same extent, the correct recognition in around 90% of the cases might indeed correspond to the ground truth.

Table 4.5: Classification performance of *trouble* and *baseline* as well as their unweighted average (UA) using physiological features. The results are shown for the development set and the test set (highlighted) in terms of recall, precision and f-measure. The table is adapted from [Egorow & Wendemuth 2016].

Test Condition	Recall	Precision	F-Measure
Development			
– Trouble	100	80.00	88.89
– Baseline	87.50	100	93.33
– UA	93.75	90.00	91.11
Test			
– Trouble	75.00	100	85.71
– Baseline	100	88.89	94.12
– UA	87.50	94.44	89.92

Although the classification results using physiological signals look promising, we should not forget that this kind of data also has one important disadvantage – it is difficult to obtain. Especially regarding the usability of systems based on such data, the intrusiveness of physiological sensors might be a considerable obstacle. Therefore, it is important to further investigate contactless methods for collecting physiological data.

One direction to look at is the estimation of heart rate from video data, which allows for around 90% IEC accuracy [Rapczynski et al. 2019]. The IEC accuracy is defined by the International Electrotechnical Commission standard

60601 – a value of $n\%$ means that $n\%$ of the estimated values deviate from the ground truth less than 10% or 5 beats per minute [Rapczynski et al. 2019].

Another direction for further development is relying on audio data – the estimation of heart rate ranges in terms of high and low heart rate is possible using speech. As shown my colleagues and myself, high and low heart rate can be distinguished using speech with an UAR of up to 80% [Egorow et al. 2019]. If the accuracy of such approaches is further improved, we can assume that the heart rate is obtainable in a contactless, non-intrusive way.

But until these problems are solved, it is rather impractical to rely on physiological signals for the detection of *trouble* in HCI, despite the high recognition performance. Therefore, in the next section, we will take a look at another modality that might help find the balance between the obtainability of acoustic signals and performance of physiological signals.

4.5 Detecting Trouble with 3D Data

So far, we have observed the acoustic and physiological data during an interaction – both modalities have their advantages and disadvantages. Speech is easily obtainable in a non-intrusive way, but the classification results are in need of improvement. Physiological signals deliver satisfying results, but are difficult to include in everyday systems. In this section, we analyse a modality that might combine the advantages of both modalities and focus on upper-body posture patterns obtained from 3D data.

In order to analyse 3D data, we should take a look at the LMCv2 described in detail in Section 3.2.2. For our purposes, the main difference in comparison to the first version is the availability of 3D movement data obtained by a Kinect 2 sensor as an additional modality. Unfortunately, there are no transcripts or segmentation available for the LMCv2, only time stamps of the triggering events for the stages. Due to the missing segmentation into utterance of the two interaction partners, it is not feasible to use the acoustic data of the human interlocutors, therefore it is not possible to use this data for comparison. However, since the experimental designs of both data sets are almost identical, we can expect the results to be similar to those obtained on the LMCv1. Nevertheless, since the triggering events of the stages are annotated, the movements and postures of the subjects during different interaction stages can be labelled.

In the investigation[3] reported in this section, data from 53 of the LMCv2's subjects were used, since for the other subjects, the audio and Kinect recordings could not be synchronised due to technical issues during the interaction

[3]This investigation describes original research performed by myself and is not yet published.

experiments. Since the timestamps of the triggering events are aligned to the audio recordings, the post-hoc synchronisation of the triggering events with the 3D data and consequently the annotation of the stages was not possible.

4.5.1 Calculating Upper-Body Postures from Kinect Data

The features used in this section are calculated directly from the original Kinect data. In the first step, the data were downsampled from 30 Hz to 3 Hz in order to reduce the amount of highly redundant information. For this, a moving average filter was applied over 10 frames to smooth the movements. After this, the data were downsampled to 10% of the original frames by removing nine out of ten of the averaged data points. The most plausible points to use as features are the coordinates of the head (H), left shoulder (LS) and right shoulder (RS) – the subjects were seated during the interaction experiments, and these points were visible throughout the sessions, in contrast to points from other body parts. Figure 4.1 shows one sample of the recordings captured by the various sensors[4]. The shoulders and the head of the subject are indeed visible and recognised correctly, whereas most of the other points are not accessible.

Figure 4.1: A scene from the LMCv2. From left to right: infrared image, depth image, recognised body points (H, LS and RS are encircled).

Each of the three body points provided three coordinates (x, y, z), resulting in 3×3 features. From these nine features, additional features were calculated: the average values of the (x, y, z) coordinates of H, LS and RS to include general information on the spatial position and the extension of the body, the Euclidean distance between the point of origin and the (x, y, z) coordinates of H, LS and RS to include information on the general direction in relation to the system, and their deltas to include information on temporal changes. As we have already seen in Section 1.2.2, both the velocity and the spatial extent

[4]The images were generated by myself based on the videos of the LMCv2 kindly provided by Prof. J. Frommer and Prof. D. Rösner and their teams.

of the body should be included in a feature set, since they perform best for the task of emotion classification [Ahmed & Gavrilova 2019].

This resulted in a feature vector of 27 3D features with a framerate of 3 Hz. The features are presented in Table 4.6. The feature values were finally standardised ($\mu = 0$, $\sigma = 1$) in the last processing step.

Table 4.6: List of the 3D features obtained from the coordinates of the head (H), left shoulder (LS) and right shoulder (RS).

Body part	Coordinates	Mean	Vector Length
H	x_H	$\frac{x_H+y_H+z_H}{3}$	$\sqrt{x_H^2+y_H^2+z_H^2}$
	y_H		$\Delta\sqrt{x_H^2+y_H^2+z_H^2}$
	z_H		
	Δx_H		
	Δy_H		
	Δz_H		
LS	x_{LS}	$\frac{x_{LS}+y_{LS}+z_{LS}}{3}$	$\sqrt{x_{LS}^2+y_{LS}^2+z_{LS}^2}$
	y_{LS}		$\Delta\sqrt{x_{LS}^2+y_{LS}^2+z_{LS}^2}$
	z_{LS}		
	Δx_{LS}		
	Δy_{LS}		
	Δz_{LS}		
RS	x_{RS}	$\frac{x_{RS}+y_{RS}+z_{RS}}{3}$	$\sqrt{x_{RS}^2+y_{RS}^2+z_{RS}^2}$
	y_{RS}		$\Delta\sqrt{x_{RS}^2+y_{RS}^2+z_{RS}^2}$
	z_{RS}		
	Δx_{RS}		
	Δy_{RS}		
	Δz_{RS}		

4.5.2 Trouble Classification on 3D Features

For the classification, a SVM classifier with a RBF kernel as provided by the Python Scikit-learn library [Pedregosa et al. 2011] was employed. The frame-wise classification was conducted on the individual feature vectors. The classification results were calculated over a total of around 72,000 samples.

For the evaluation, the two different settings described in Section 2.2.7 were employed, TDT and LOSO.

As we remember from Section 2.2.7, the TDT setting enables statements about the classification, not about the classified data, whereas the LOSO setting allows for statements about the classified data, and not so much about the classification procedure. To benefit from both settings, we can combine them and perform a hyperparameter optimisation using the TDT setting and

then evaluate the single-subject-performance using the LOSO setting with the previously chosen hyperparameters.

For the TDT setting, the 56 subjects of the LMCv2 are divided in a training subset with 37 subjects (roughly 70%) and a development and test subset with 8 subjects (roughly 15%) each. For the LOSO setting, there are 37 classification procedures to perform, one for each of the subjects from the TDT's training subset, using 36 subjects for training and the remaining subject for test. It is plausible to use only the training set's subjects in order to maintain comparable training set sizes – since the training data size influences the classification performance [Figueroa et al. 2012; Beleites et al. 2013]. The exact nature of the influence is reported rather rarely: For example in the domain of body part classification from computer tomography images, the accuracy increases from around 50% to around 77%, when the amount of training images increases from 20 to 50, further increasing to around 90% for 100 images [Cho et al. 2015]. Comparing the results obtained on the same amount of data appears as a plausible choice.

First, we want to look at the findings from the TDT setting. The classification results for this setting are presented in Table 4.7. The upper part of Table 4.7 reports the performance in terms of UAR, UAP and UAF on the development set. We see that the performance varies depending on the chosen hyperparameters from around 56% in the worst case to around 74% in the best case, with an average of 67% ± 5.1%. For the hyperparameter optimisation, a total of 200 trials was performed using a Random Search procedure [Bergstra & Bengio 2012].

The last line in Table 4.7 shows that using the best hyperparameters, namely $\gamma = 2 \cdot 10^{-8}$ and $C = 51574$, the final evaluation on the test set achieves results very similar to the performance on the development set.

Table 4.7: Classification performance for the TDT setting. The results include the average and the range of the results achieved in the 200 trials on the development set (Dev av. and Dev range, respectively), as well as the final test result (highlighted), in terms of UAR, UAP and UAF, in %.

Condition	UAR	UAP	UAF
Dev av.	67.00 ± 5.1	66.30 ± 5.2	66.60 ± 4.9
Dev range	56.21 ... 74.54	55.67 ... 73.73	55.93 ... 74.14
Test	74.14	74.21	74.18

The TDT setting allows us to analyse the classification in even more detail. Figure 4.2 illustrates the results achieved during the hyperparameter optimisation procedure. We can spot an interesting relation: Apparently, the γ parameter seems to have a much greater influence on the classification compared to C. The performance remains relatively stable over different C values

for the same γ (horizontal direction) but varies over different γ values for the same C value (vertical direction). This can be seen in Figure 4.2, since there is more variation in the vertical axis of the colour-coded UAF than in the horizontal axis.

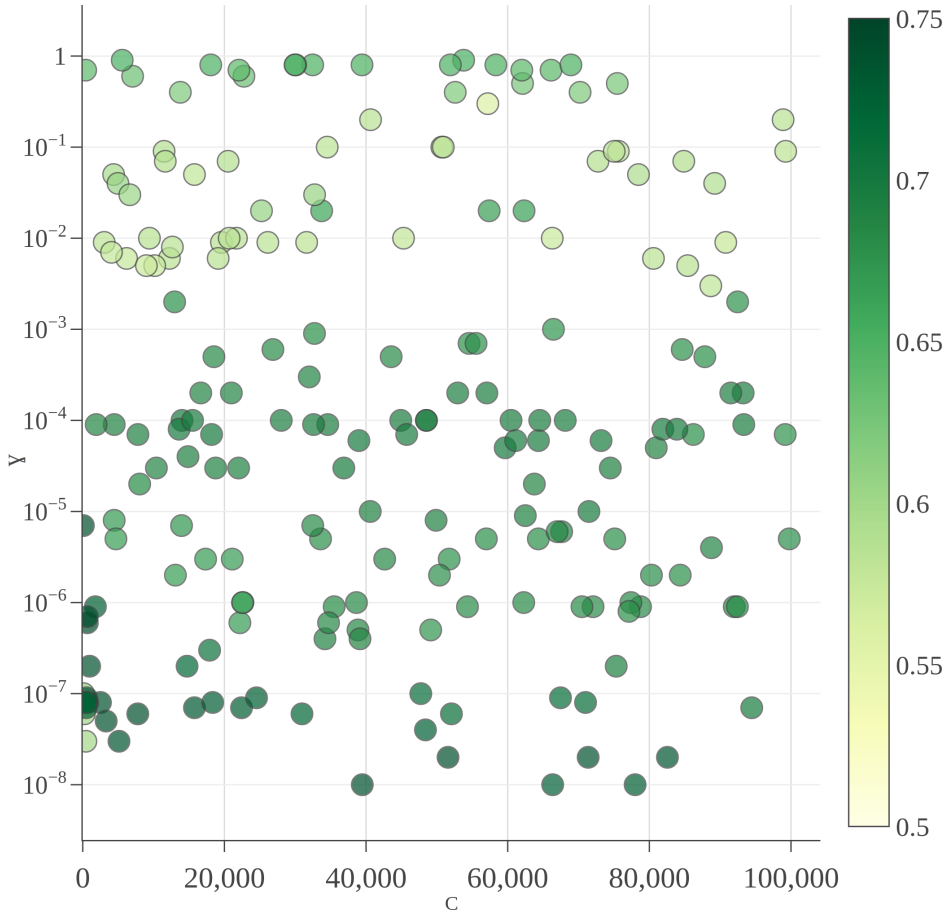

Figure 4.2: The influence of the hyperparameters C and γ on the classification result. Each dot corresponds to a tested tuple (C, γ), darker colours correspond to higher UAF.

Now we can turn to the findings obtained during the LOSO setting. The results are presented in Table 4.8, again in terms of UAR, UAP and UAF. The average achieved over all 37 folds is 73.75% ± 21.53% UAF – this is slightly lower than 74.18% UAF achieved in the TDT setting. However, this value is not significantly lower, taking into account the high standard deviation. Indeed, the results between single subjects vary remarkably, from around 26% UAF in the worst case to 100% UAF in the best case – with 5 subjects achieving almost perfect classification with over 99% UAF. This emphasises the ability of the LOSO setting to gain information on the inter-individual differences between the subjects, since the TDT setting cannot provide such details.

Table 4.8: Classification performance for the LOSO setting. The results include the average over the 37 folds (highlighted), the range between the worst and best performing subject, and the number of subjects with nearly perfect classification (Average, Range and # Perf. subj's, respectively), in terms of UAR, UAP and UAF, in %.

Condition	UAR	UAP	UAF
Average	75.57 ± 18.86	77.46 ± 20.26	73.75 ± 21.53
Range	42.58 ... 100	18.10 ... 100	25.60 ... 100
# Perf. subj's	4	5	5

But we should not only look at the average classification results – the LOSO setting allows us to take a look at inter-subject differences. Figure 4.3 presents an overview of all 37 considered subjects. For eleven of them, the results are rather dissatisfying, with seven subjects leading to results below the chance level of 50%. At the same time, for 21 of the subjects, the classification's UAF is at least 75%, which is higher than the average value of around 74%. As already mentioned, for five of the subjects, a value of over 99% UAF could be achieved.

As a result of both evaluation settings, it can be stated that the employed 3D features combine the advantages of acoustic and physiological data. On the one hand, such features are almost as easy to obtain as acoustic data using already available sensors such as the Kinect sensor. On the other hand, the classification using these features remarkably outperforms the classification on the acoustic data – although the achieved results are not as high as those of physiological data. We will return to the trade-off between the obtainability of the features versus the classification performance in Section 4.6.

One interesting aspect would be the fusion of acoustic features and 3D features. However, neither LMCv1 nor LMCv2 allow such fusion: LMCv1 lacks the necessary 3D data, and LMCv2 does not have segmented acoustic data.

4.6 Concluding Remarks on Trouble Recognition

The investigations conducted using acoustic, physiological and 3D signals of the interlocutor enable us to compare and discuss different modalities as input for the recognition of the *trouble* state.

The acoustic signals have several advantages – first and foremost, they are the most natural way of human communication and their usage seems not to rise any concerns in time of voice assistants such as Siri and Cortana. Additionally, acoustic signals are also relatively easy to obtain in different ap-

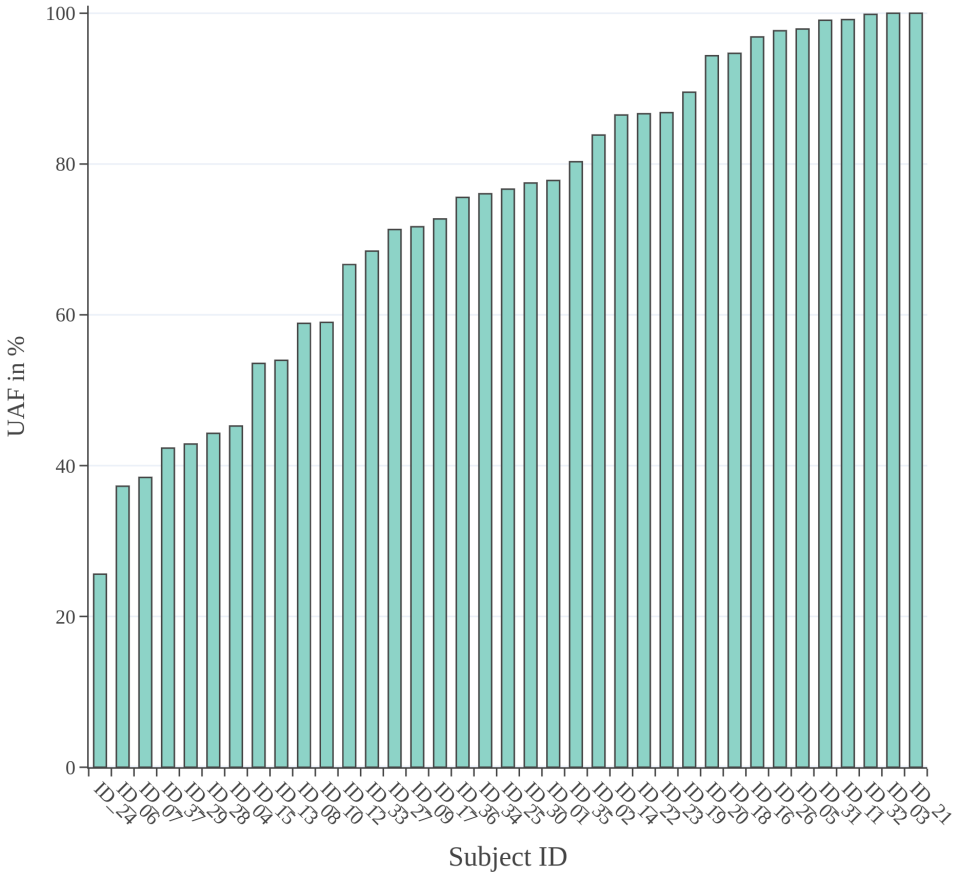

Figure 4.3: The distribution of the results on the individual subjects in terms of UAF in %. Each bar corresponds to a subject, the subjects are sorted according to the achieved UAF.

plication areas, especially when other modalities are not available, e.g. via telephone. Unfortunately, acoustic signals also deliver the least useful results when it comes to the task of *trouble* recognition. The experiments have shown that, although many acoustic features differ significantly between challenging and non-challenging interaction, they do not perform well for classification, resulting in around 64% UAF. Nevertheless, even with this dissatisfying performance, acoustic signals can be of use in specific applications – especially in combination with other signals.

Physiological signals such as EMG, RSP and SC seem to be just the opposite of acoustic signals: They are relatively difficult to access and therefore controversial when it comes to usability. But at the same time, they also correspond to the "ground truth", since humans cannot influence them deliberately. The classification of *trouble* using features based on physiological signals delivers impressive results of around 90% UAF – this corresponds to an improvement of around 26% absolute compared to acoustic features with the same classi-

fication and evaluation setup. Furthermore, we can even assume that this is the "gold standard", i.e. the best results we can obtain without a detailed self-assessment of the subjects.

The third modality investigated in this chapter, the 3D data, achieves a good compromise between obtainability and performance. We have seen that using upper-body posture features obtained from a Kinect sensor, a relatively high UAF of around 74% can be achieved. Especially in the context of personal assistants employed at home, this seems to be a promising research direction. The necessary sensors are readily available and are easy to install. The classification needs a low number of features that can be computed on low-resource devices and can be performed without cloud-based computing power. Furthermore, a possible combination of these features with gestural features might be interesting, since there are distinct gestures known from everyday life associated with *trouble*, such as scratching the head etc. Additionally, the fusion with acoustic features should be considered, as both modalities are acceptable for certain applications. Unfortunately, the two data sets used in the investigations do not allow a multimodal fusion, since for the LMCv1, no 3D data were recorded, whereas for LMCv2, no transcripts are available and therefore the acoustic data cannot be used. A possible direction for further work is to prepare the transcripts for LMCv2 and to re-evaluate the classification using the acoustic data.

Another interesting aspect worth looking into is the indirect gathering of physiological data, since, as already mentioned above, there are approaches to calculate some physiological signals from other modalities, e.g. heart rate from video and audio data [Rapczynski et al. 2019; Egorow et al. 2019]. Furthermore, there are wearable devices that allow to collect such data, starting from consumer smartwatches measuring the heart rate such as the *Fitbit Flex*[5] and the *Xiaomi Mi Band*[6] up to high-level devices used by professional athletes such as the *BioHarness*[7]. This development might change our possibilities for as well as opinions on using such sensitive data in the future.

Besides investigating different models for *trouble* recognition, we also compared two different evaluation settings, the TDT setting and the LOSO setting. The former illustrates the influence of the classifier's hyperparameters. Here, we have found that the classification performance is mostly influenced by the γ parameter of the SVM compared to the C parameter. The latter evaluation setting illustrates the inter-subject differences and allows us to get a deeper understanding of the investigated data. In the conducted investigation, the

[5] `http://www.fitbit.com/`, retrieved December 15, 2019.

[6] `http://www.mi.com/global/miband/`, retrieved December 15, 2019.

[7] `http://www.biopac.com/product-category/research/telemetry-and-data-logging/bioharness/`, retrieved December 15, 2019.

LOSO setting showed that for some of the subjects, a nearly perfect classification of *trouble* was possible, whereas for others, the classification performance was below chance level. These inter-subject differences should be further investigated in order to find the underlying causes – this might help to design truly personalised and anticipative systems.

4.7 Summary of the Chapter

In this chapter, we dedicated our attention to the topic of detecting challenging interaction stages, or simply, *trouble* in interaction. For this, we have employed three different modalities – acoustic, physiological and 3D signals – and looked into statistical differences as well as classification performances. The results confirm that the automatic classification of *trouble* is possible, however with varying success even on the same data, depending on the used modality. We have also analysed the influence of the evaluation setting on the classification performance by comparing LOSO and TDT evaluation.

We have seen that all of the three considered modalities bring their own positive and negative aspects, and discussed their trade-offs by developing ideas on what signals are suitable for specific applications.

The next chapter is dedicated to *satisfaction*, another HCI-relevant interlocutor state.

Chapter 5

Assessing the State of Satisfaction

Contents

IF an HCI system is to be used, it should provide a satisfactory user experience – otherwise, this system will lose its users. According to ISO 9241-11:2018, usability is characterised as an interplay between efficiency, effectiveness and satisfaction [ISO9241-11:2018 2018]. That is why the topic of user satisfaction as one of three constituents of usability is a most important one, with usability becoming the central issue of (software) design, highlighted by events such as the World Usability Day[1].

In this chapter, we look at the internal state of *satisfaction* during HCI using the Last Minute Corpus data that we have introduced in Section 3.2 and encountered again in Chapter 4. In Section 2.1.1, we have already developed an understanding of the state of *satisfaction*. Now, we will elaborate on it in more detail and review available work in this field in Section 5.1. Then, we will analyse how the LMC can be used for the task of *satisfaction* classification in Section 5.2, before turning to the features that are useful for this task in Section 5.3. The process of classification itself will be described in Section 5.4.

[1] `https://worldusabilityday.org/`, retrieved December 15, 2019.

We will conclude and summarise the conducted investigation in Section 5.5 and Section 5.6.

This chapter is based on the investigation published by my colleagues and myself [Egorow et al. 2017].

5.1 Existing Works on Satisfaction and Its Recognition

Satisfaction is important for most systems developed for end-users: It was already investigated for example in search applications [Hassan & White 2013] or instant mobile messaging [Ogara et al. 2014], but also data warehouses [Chen et al. 2000] and information security practices [Montesdioca & Macada 2015], etc. However, satisfaction has been known to be an ambiguous concept, even an "evasive beast" [Lindgaard & Dudek 2003] – it can be defined in various ways depending on the application domain. In Section 2.1.1, we discussed a working definition for the interlocutor state of *satisfaction*. Another possible definition is a "complex construct comprising several affective components as well as a concern for usability" [Lindgaard & Dudek 2003]. More broadly, satisfaction can also be defined as "fulfilment of a specified desire or goal" [Kelly 2009]. These reflections directly lead to a challenge – satisfaction seems to be a deeply subjective matter [Kiseleva et al. 2016a], since objective properties are accompanied by affective reactions that cannot be voluntarily controlled [Zajonc 1980]. If users are asked to rate their satisfaction regarding a technical system, these ratings seem to be shaped by prior experience and also to reflect the immediate impression [Lindgaard 1999]. One way to evaluate satisfaction is the WAMMI measure[2] of global satisfaction based on five dimensions of user experience (attractiveness, control, efficiency, helpfulness, learnability) [Kirakowski et al. 1998].

We have already discussed the question of obtaining the "ground truth" in Section 3.1.1. For the topic of this chapter, this question is not less important: How do we get reliable information on the state of *satisfaction* of an interaction participant?

The most obvious answer is to ask the participant directly, for example using existing questionnaires such as the WAMMI measure mentioned before, the Questionnaire for User Interface Satisfaction [Chin et al. 1988], or questionnaires developed for the specific task at hand [Shriberg et al. 1992; Kiseleva et al. 2016b]. But this is not always applicable, since it can distort the interaction course and influence the interaction behaviour. Furthermore, the aim of naturalistic interaction should be to anticipate the approaching dissatisfaction

[2]Website Analysis and MeasureMent Inventory

Table 5.1: An overview of current methods for automatic satisfaction classification and related tasks. The employed features, available information on data, methods and evaluation procedures (if given by the original authors) are shown. The performance is given in terms of accuracy (Acc.) or f-measure (F-1).

Features	Data	Method	Performance
Task: Search satisfaction prediction			
[Fox et al. 2005] Searching behaviour (clickthrough, dwell time, exit type, etc.)	Search engine, 146 subjects	Bayesian network, hold-out test set, 3 classes	Acc. 56%
[Kim et al. 2014] Searching behaviour (click dwell time-related)	Search engine, 3204 queries	Regression, 10-fold CV, 2 classes	F-1 80%
[Mehrotra et al. 2017] Action sequences (viewport, cursor, keyboard)	Search engine, 14670 sessions	Deep sequential models, random TDT	F-1 80%
[Feild et al. 2010] Behaviour (mouse & chair pressure, query log, scrolling)	Search engine, 30 subjects	Logistic regression	F-1 49-80%
Task: Interaction satisfaction prediction			
[Hara et al. 2010] Dialogue act sequences	Music retrieval, 518 subjects	N-gram model, LOSO, 6 classes	Acc. 35%
[Chowdhury et al. 2016] Acoustic (pitch, loudness, etc.), lexical, turn-taking	Call centre (HHI), 86h	SVM, TDT, 3 classes	F-1 40-61%
[Kiseleva et al. 2016a] Acoustic (phonetic similarity), click, touch	Mobile voice assistant, 30h	Decision Trees, 10-fold CV, 2 classes	F-1 61-79%

in order to be able to react, or, in the best case, to avoid it. A system causing dissatisfaction has low usability and might lead to its rejection – or at least it will not be used in the most efficient and effective way.

Besides direct questions on satisfaction, it is also possible to search for signs of other related internal states, such as frustration and empathy, in order to deduce the necessary information [Chowdhury et al. 2016]. Another possible solution is to refrain from trying to access internal states and to rely on objective measures, such as the success of users' tasks [Fox et al. 2005], scrolling and gazing behaviour (e.g. the visible portion *viewport* of the browser system) [Lagun et al. 2014] or clicking behaviour [Joachims et al. 2005]. Besides mere measuring the satisfaction level, these metrics can be used as features for satisfaction prediction. Table 5.1 presents a selection of current satisfaction prediction approaches. Especially in the domain of search satisfaction, action- or behaviour-related features such as clicking and scrolling are used. For satisfaction prediction during interactions other than searching, acoustic features have also been in the focus. One such example is using acoustic, lexical and turn-taking features for predicting satisfaction levels obtained from implicit metrics [Chowdhury et al. 2016], where the authors achieve an f-measure

of 40% (with only prosodic features) to 61% (with turn-taking features) for the classification of three satisfaction levels. In another example, acoustic, click and touch features are used in a self-assessment scenario to achieve an f-measure of 61% (with queries- and click-based features) to 79% (adding touch features) in a two-class-classification [Kiseleva et al. 2016a].

In this chapter, we focus on the state of *satisfaction* during HCI. As we already know from Section 3.2, the LMC experiments were concluded with a questionnaire module, where the participating subjects were questioned regarding their *satisfaction* with their own performance. This allows us to map the user's utterances directly to her *satisfaction* level. We will use acoustic features in order to distinguish between low and high *satisfaction*. But first, we need to look at the LMC in detail in order to understand how this data can be used for the task at hand.

5.2 Satisfaction in the Last Minute Corpus

In the current chapter, we once more use the data of the LMC introduced in Sections 3.2 and 3.2.1. As already described, in the interactions recorded for this corpus, the participating subjects had to pack a suitcase for a trip to an unknown location. For the presented investigation, we should take a look at the final part of these recordings, in which the subjects had to answer direct questions on their overall experience during the experiment. The question that is of most interest in the context of this chapter is: "How satisfied are you with the content of your suitcase?" [Frommer et al. 2012a]. This question implies that the subjects were asked about their *satisfaction* regarding their own performance (i.e. the selected items in the suitcase) and not regarding the system. This is an important point, since this direct question allows us to assume a one-to-one-mapping between the *satisfaction* state and the answer. Furthermore, we should also notice that the subjects were encouraged to talk freely and naturally in their own words, since no answer possibilities were offered.

The answers of the subjects covered a broad range of descriptions – from answers such as "extremely satisfied" to "not satisfied at all". In order to categorise the subjects' answers, a Likert-type scheme containing five *satisfaction* levels "very satisfied" (S+), "satisfied" (S), "moderately satisfied" (M), "dissatisfied" (D), "very dissatisfied" (D+) was developed prior to annotation [Egorow et al. 2017]. An overview of the exemplary answers for each of the levels can be seen in Table 5.2.

For the experiments described in this chapter, a subset of 89 of the 133 LMC subjects was used. They were chosen due to the audio data quality – these are the same subjects as in Chapter 4. The answers of these 89 subjects

Table 5.2: Examples for each of the five *satisfaction* levels. The table is adapted from [Egorow et al. 2017].

Satisfaction level	Examples
S+	"very / extremely / absolutely satisfied"
S	"quite okay", "(fairly) satisfied"
M	"moderately satisfied", "so-so", "reasonable"
D	"unsatisfied", "not very / less satisfied"
D+	"very unsatisfied", "absolutely not satisfied", "not at all"

were then annotated according to the previously described scheme by two annotators, excluding all subjects with unclear or missing statements on their *satisfaction*. This procedure resulted in a set of a total of 79 statements in the five previously mentioned levels. Especially the first and the last levels contained only few instances – therefore the five levels were condensed into two nearly balanced classes, with a positive class P containing all instances from "very satisfied" to "moderately satisfied", corresponding to the state of high *satisfaction*, and a negative class N containing the instances with the labels "dissatisfied" and "very dissatisfied", corresponding to the state of low *satisfaction*. The distribution of the instances over the original five *satisfaction* levels and the condensed two classes is shown in Table 5.3.

Table 5.3: Overview of the *satisfaction* level and class distribution over the 79 processed subject statements. The table is adapted from [Egorow et al. 2017].

Satisfaction level	# Samples	Class	# Samples
S+	4		
S	19	P	37
M	14		
D	37	N	42
D+	5		

The distribution of the classes with respect to the age and sex of the subjects is shown in Table 5.4. One point here deserves particular attention: For female as well as the younger subjects, there is a nearly equal distribution of the two classes, but there are around 50% more dissatisfied than satisfied male and elderly subjects. We can see that there are 23 female subjects in the class P and 22 in the class N (first row of Table 5.4), but only 14 male subjects in the class P versus 21 in the class N (second row of Table 5.4). At the same time, there are 22 young subjects in the class P and 20 in the class N (third row of Table 5.4), but 15 versus 22 elderly subjects in the P and N class, respectively (last row of Table 5.4).

Table 5.4: Overview of the sex and age distribution of the 79 subjects. The table is adapted from [Egorow et al. 2017].

Characteristic	P	N	Overall
Sex			
– female	23	21	44
– male	14	21	35
Age			
– young	22	20	42
– elderly	15	22	37

5.3 Acoustic Features for Satisfaction Recognition

Before turning our attention to the task of *satisfaction* classification, we should first take a look at the feature extraction process and subsequent post-processing. Since our data set offers two special challenges – the natural language used in the experiments as well as the low number of instances to be used for classification – we also introduce two procedures that might help to tackle these challenges: Voice Activity Detection (VAD) and feature selection.

5.3.1 Extracting Acoustic Features

The features were extracted from the entire satisfaction statement of the subject, including noise, pauses, etc. As a feature set, again the well-known *emobase* feature set, introduced in detail in Section 2.2.4, was chosen – we have already encountered it previously in Section 4.3. It contains 988 features based on 19 functionals of 26 LLDs and their first order derivatives. The features were extracted on utterance level in the same way as described in the *emobase* documentation [Eyben et al. 2010]. Since each of the subjects contributed exactly one statement to the data set, this resulted in one sample per subject, and therefore one feature vector. One beneficial side effect of this setting is an easily implemented subject-independent evaluation (we will come back to this point later in Section 5.4.1). As a final pre-processing step, the features were standardised ($\mu = 0$, $\sigma = 1$) using the procedure described in Section 2.2.4.

5.3.2 Applying Voice Activity Detection

As already mentioned, the subjects spoke in a natural, unrestricted way. Therefore, their statements also contained filled and silent pauses as well as other sounds, such as breathing, etc. One possible way to avoid the negative influence of such artefacts on the classification process is the implementation of a VAD routine. A tool that can be used for this task is the VAD routine

integrated in the previously mentioned openSMILE toolkit [Eyben et al. 2010]. The routine is based on a LSTM-RNN model provided by the authors of the toolkit, being a preliminary version of a commercial VAD system [Eyben et al. 2013]. The toolkit does not provide any possibility to select the parameters of the routine, therefore it was used with the default parameters. Figure 5.1 illustrates its effect applied to our data.

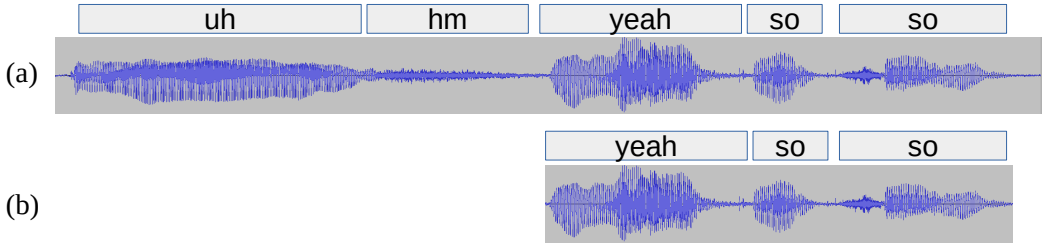

Figure 5.1: The effect of the VAD routine on an exemplary statement. Subfigure (a) shows the original speech signal, subfigure (b) shows the speech signal after the application of the VAD routine. The figure is taken from [Egorow et al. 2017].

In the depicted case, the original statement was "uh hm yeah so-so" – the statement begins with an unvoiced part consisting of two filled pauses. The processed statement contains only the last part, "yeah so-so". Applied to our data, the VAD routine detected and removed unvoiced parts in 23% of the 79 statements.

5.3.3 Reducing the Number of Features

As explained above, the feature set employed in this investigation consists of 988 features. At the same time, our data set contains only 79 statements, with each statement corresponding to a sample. This results in a sparsely populated high-dimensional feature space with only few instances. One possible solution for this issue is to reduce the number of features by implementing a feature selection routine. We will analyse how this can be done in our case by applying a common feature selection routine based on Random Forest (RF).

As we already know from Section 2.2.6, RF is a classifier consisting of a certain number of decision trees. The decisions in the trees are done using a random subset of the features and choosing the feature best suitable for the separation. The performance of this feature is measured based on the impurity measure, such as the Gini index that we have already seen in Section 2.2.6.

This mechanism can be exploited to implement a feature selection routine that we want to attend to in detail. The method is based on the available literature [Chen & Lin 2006; Silipo et al. 2014]. Besides using it in the investigation described in this chapter, it was also applied to improve the performance of

emotion recognition by my colleagues and myself [Egorow et al. 2018]. The main objective of this method is to rank the features according to their relevance for the classification task. For that, we calculate two scores. The first score is the training score S_t. In order to obtain it, a RF with a high number of trees n and a fixed low number of levels k – since the most relevant features are close to the root, k can be low – is trained on the training set. Then two statistical values for each feature f are calculated: the number of trees t_i that use f as split feature on tree level i, and the number of occurrences o_i of f in the feature sample for level i. S_t for each f is the fraction of these two values for f summed up over all levels:

$$S_t = \sum_{i=0}^{k} \frac{t_i}{o_i} \tag{5.1}$$

The second score S_r is necessary to reflect the undesired bias that might be contained in the structure of the data. It is calculated in the same way as S_t, but now the labels of the data are randomly shuffled before training the RF. In order to account for the randomness, both values are now calculated several times (for example ten times as in our case) and then averaged. The final score S_f is the difference between S_t and S_r:

$$S_f = S_t - S_r \tag{5.2}$$

The features can now be sorted according to their scores S_f that indicate their relevance for the classification task. For the feature selection, we select the top features from the ranking. The effectiveness of this implementation for the field of affective state recognition was shown by applying it to three benchmark data sets. Using only the top 40 to 60% of the features, an increase in performance could be achieved on all three data sets [Egorow et al. 2018].

Applied to our current setting, the top-most 10% of the ranked features were selected as an additional reduced feature set. The selecting RF was trained on the training subset of the data, the amount of the selected top features was chosen on the development subset of the data (the same training and development subsets that are later used for the classification as described in Section 5.4.1). These 99 features are derived from 33 out of the originally 54 LLDs constituting the complete *emobase* feature set. Most of the selected features are functionals of certain MFCCs and LSP frequencies. The selected features are presented as a word cloud in Figure 5.2. Larger font corresponds to features based on the respective features appearing more frequently in the selected set. Interestingly, both the MFCCs and LSPs were also relevant for the discrimination of challenging and non-challenging interaction stages as described in Section 4.3.1.

Figure 5.2: Word cloud of LLDs most frequently occurring in the reduced feature set. The numbers in brackets indicate which MFCCs or LSPs were selected. The figure is taken from [Egorow et al. 2017] by courtesy of I. Siegert.

5.4 Classification of Satisfaction Levels

In this section, we will analyse how the classification of the states of high and low *satisfaction* in the LMC can be performed. For this, we first establish the evaluation setup, before continuing with the description of the classification setup and presenting the results.

5.4.1 Evaluation Setup

As already mentioned in Section 5.3.1, the structure of the employed data allows for an easy implementation of subject-independent evaluation. The Train-Dev-Test (TDT) setup that we will encounter here was already introduced in Section 2.2.7 and used in Chapter 4. For the investigation described here, we again define subject-independent TDT sets for training, development and test with an approximately equal distribution of the relevant subject characteristics. The distribution of these characteristics can be seen in Table 5.5.

During the classification process, the classifier is trained on the training set and repeatedly evaluated on the development set to find the best hyperparameters in a grid search procedure. After finding the best-performing hyperparameters, the classifier is evaluated on the test set. We will look at the procedure in detail below.

5.4.2 Classification Setup

As a classifier, an SVM using the libSVM implementation [Chang & Lin 2011] was employed. This same classification setup was implemented for all three setups mentioned earlier in Section 5.3: Classification without VAD, classification with VAD, and classification using pre-selected features. Therefore, we will refer to these setups as S_0, S_{vad} and S_{sel}, respectively.

Table 5.5: Distribution of subject characteristics and class over the training, development and test sets as well as the overall distribution. The table is adapted from [Egorow et al. 2017].

Characteristic	Training	Development	Test	Overall
Sex				
– female	35	4	4	44
– male	28	4	4	35
Age				
– young	33	4	4	42
– elderly	30	4	4	37
Class				
– P	30	4	4	37
– N	33	4	4	42

The general classification setup S_0 is shown in Figure 5.3. The data were divided in a training, development and test subset, as described above. The training and fine-tuning were performed on the former two data subsets, the final evaluation on the latter one.

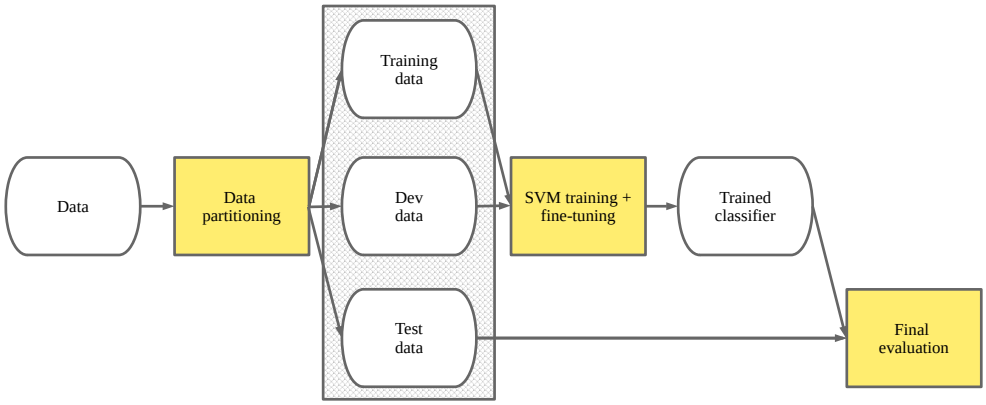

Figure 5.3: A general overview of the *satisfaction* classification setup S_0. The part highlighted by the striped box is replaced by the feature selection or voice activity detection routine in setups S_{vad} and S_{sel}.

Both the feature selection routine as well as the voice activity detection routine are applied to the data prior to the classifier training, replacing the part of the general pipeline highlighted by the striped box in Figure 5.3. The feature selection routine is schematically shown in Figure 5.4: The training and development data are used for the calculation of the feature scores and the ranking, with the final selection applied to the data, resulting in the same three data subsets containing only selected features.

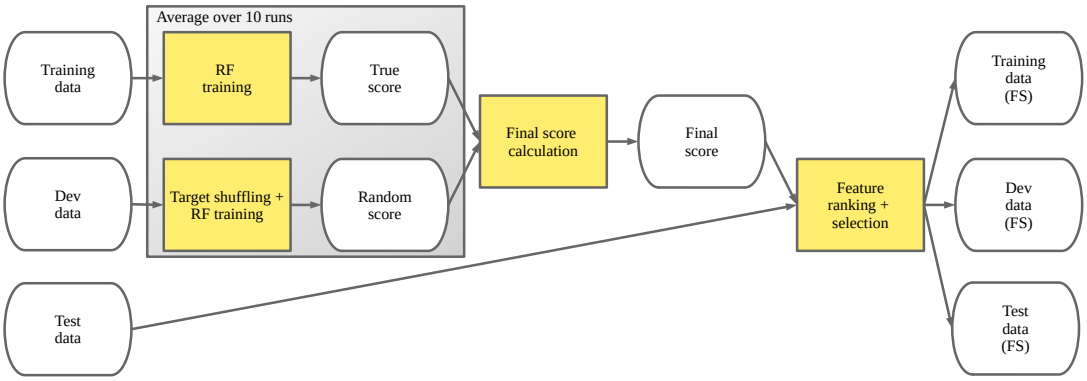

Figure 5.4: An overview of the feature selection routine for the *satisfaction* classification setup S_{sel}, resulting in data with only selected features (indicated as FS).

The voice activity detection routine is depicted in Figure 5.5: It is applied directly to the three data subsets, resulting in the same subsets containing only the detected voiced parts of the utterances.

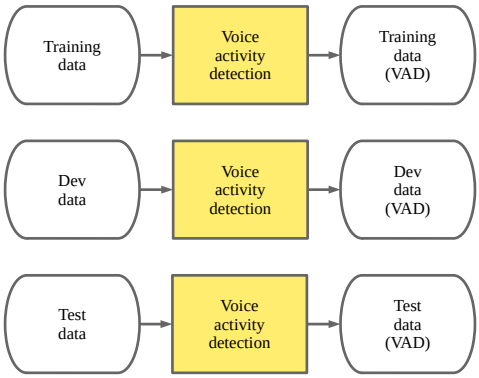

Figure 5.5: An overview of the voice activity detection routine for the *satisfaction* classification setup S_{vad}, resulting in data of only voiced parts of the utterances (indicated as VAD).

Prior to classification, a fine-tuning procedure was performed on the development set for all three setups. Linear, polynomial and RBF kernels as well as different hyperparameters of the kernels were tested in a grid search procedure. For S_0 and S_{vad}, the best performing model was a linear SVM with $C = 10$, for S_{sel}, an RBF SVM with $C = 10$ and $\gamma = 0.001$ performed best.

5.4.3 Classification Results

Now that we got familiar with the experimental setup, we want to take a look at the classification results for the three setups. First, we analyse the results obtained using S_0. These results are presented in Table 5.6. The classification

performance is the same for both, the development and the test set – this supports the assumption that the two subsets contain similar data and that the classifier is able to generalise. The achieved results are fairly high, with UAR, UAP and UAF values of almost 90%. This is indeed remarkable, since the employed features are relatively easy to obtain compared to the sophisticated feature sets employed in the literature described in Section 5.1.

Table 5.6: Classification results with F_0, for the two classes P and N and their unweighted average (UA). The results are shown for development and test set (highlighted) in terms of recall, precision and f-measure, in %. The table is adapted from [Egorow et al. 2017].

Test Condition	Recall	Precision	F-Measure
Development			
– P	75.0	100.0	85.7
– N	100.0	80.0	88.9
– UA	87.5	90.0	87.3
Test			
– P	75.0	100.0	85.7
– N	100.0	80.0	88.9
– UA	87.5	90.0	87.3

As already mentioned, the hypothesis for introducing the VAD routine described in Section 5.3.2 was that the occurring filled and unfilled pauses distort the acoustic features. Deleting these artefacts would then lead to an improved classification performance. The results using S_{vad} are presented in Table 5.7. Here, again, the results obtained on the development and test sets are the same. But the classification performance in terms of all three metrics UAR, UAP and UAF is lower than in the previous setting. This shows that our initial hypothesis was wrong – obviously, the parts removed by the VAD routine do contribute to the classification performance. On the one hand, it can be seen as an indicator that unvoiced parts deliver information on the state of *satisfaction* that should not be ignored. However, another possible explanation could be that the better performance without the VAD routine is a consequence of more data being available for training, as the classification performance depends on the amount of training data [Figueroa et al. 2012; Beleites et al. 2013]

Finally, the results obtained with F_{sel} are shown in Table 5.8. As a reminder: The classifier was trained on a subset of 99 features selected using the RF feature selection routine described in Section 5.3.3. Here, we can observe a very peculiar classification behaviour: On the development set, the classification results replicate those achieved with F_0, but on the test set, there is a remarkable performance drop of around 25% absolute. This performance drop is caused by the low recognition of the N class. The recall of the N class on

Table 5.7: Classification results with F_{vad}, for the two classes P and N and their unweighted average (UA). The results are shown for development and test set (highlighted) in terms of recall, precision and f-measure, in %. The table is adapted from [Egorow et al. 2017].

Test Condition	Recall	Precision	F-Measure
Development			
– P	50.0	100.0	66.7
– N	100.0	66.7	80.0
– UA	75.0	83.3	73.3
Test			
– P	50.0	100.0	66.7
– N	100.0	66.7	80.0
– UA	75.0	83.3	73.3

the development set is 100% and only 50% on the test set. One possible explanation is that the feature selection routine possibly lead to overfitting of the classifier, meaning that the selected features were not able to generalise. Another explanation is that even features with low information gain substantially contribute to the classification. These findings are in line with the current search for the most suitable acoustic features addressed in the literature [Liu et al. 2018; Song & Zheng 2018].

Table 5.8: Classification results with F_{sel}, for the two classes P and N and their unweighted average (UA). The results are shown for development and test set (highlighted) in terms of recall, precision and f-measure, in %. The table is adapted from [Egorow et al. 2017].

Test Condition	Recall	Precision	F-Measure
Development			
– P	75.0	100.0	85.7
– N	100.0	80.0	89.0
– UA	87.5	90.0	87.3
Test			
– P	75.0	60.0	66.7
– N	50.0	66.7	57.1
– UA	62.5	63.3	61.9

5.5 Concluding Remarks on Satisfaction Recognition

The presented experimental results confirm that it is possible to distinguish between the states of high and low *satisfaction* in interaction using only the

acoustic content of the interlocutor's utterances. The performance of the classification in the described investigation is relatively high, achieving around 87% UAF. Counter-intuitively, using whole utterances including breathing and pauses provides better classification results than implementing VAD to remove such artefacts. The performance drop of around 12% absolute shows that even the unvoiced parts contain information that contributes to the classification. Another question addressed in the described investigation was the reduction of features. In the standard setting, we used a relatively high amount of almost a thousand features of the *emobase* feature set to classify the statements of only 79 subjects. Our hypothesis was that selecting the top 10% features with the highest information gain would not impair the classification results or even lead to higher performance, since the implemented feature routine was able to slightly improve the emotion recognition performance by up to 5% absolute in a similar setting [Egorow et al. 2018]. However, we have seen that reducing the number of features leads to a substantial performance drop of around 25% absolute. This can be regarded as evidence that our understanding of the relevance of features for a certain task and the transferability of feature sets between tasks is still limited and should be further investigated.

The classification performance achieved in the described experiments is in the range of the state of the art presented in Section 5.1, although it is always difficult to compare results obtained on different data sets including different interaction designs. However, in the investigation presented here, only automatically extracted acoustic information was used, whereas the approaches in the literature rely on annotations of complex social emotions and a sophisticated mix of lexical and turn-taking features [Chowdhury et al. 2016] or behavioural (click and touch) features [Kiseleva et al. 2016a]. Unfortunately, since there is no benchmark data set available, we cannot compare the approaches directly. However, it can be hypothesised that fusing acoustic, interaction-related and behavioural information could outperform all three approaches.

The combination of different modalities might be the most interesting direction for further research. In particular, the analysis of the different modalities and their contribution to the classification would be of interest. Having identified the most informative channels, a feature selection could further improve the recognition results. This is also relevant with regard to relatively low amounts of available data and the problems associated with obtaining the ground truth on this data.

Another question relevant in the context of "in the wild scenarios" is the temporal aspect, related to *trouble* in interaction that we analysed in Chapter 4. In real-world applications, it is desirable to recognise the upcoming dissatisfaction early on, before the user of a system consciously experiences trouble.

Therefore, the temporal evolution of the internal state of an interlocutor should also be in the focus of future research.

5.6 Summary of the Chapter

In this chapter, we examined how the interlocutor's *satisfaction* state can be assessed. We have seen that high and low *satisfaction* can be recognised from the acoustic content of the interlocutor's utterances. For this task, we analysed the last part of the LMC. This part contains the participating subjects' statements on the satisfaction with their own performance. Using this explicit data, we were able to circumvent the problem of obtaining the ground truth on the internal state of the subjects. Besides the classification setup and recognition performance, we also implemented and discussed two pre-processing methods, namely VAD and RF-based feature selection, and their effects on the *satisfaction* recognition task.

Our best result showed that the *satisfaction* level can be predicted with around 87% UAF, which makes our method applicable in HCI settings, providing information on the *satisfaction* state of the user to steer the dialogue management towards avoiding dissatisfaction, or reacting to it.

In the next chapter, we will turn our attention to the interlocutor state of *cooperativeness*, distinguishing between cooperative and competitive speech behaviour.

Chapter 6

Cooperative and Competitive Speech

Contents

SPEECH is a very effective means of communication – especially when we look at it in a holistic way. In the previous chapters, we have seen how speech can be used to detect certain human experience patterns such as *trouble* and *satisfaction* in interaction. In this chapter, we will investigate speech overlaps as a specific speech phenomenon and a marker for cooperative and competitive behaviour.

As a first step, addressed in Section 6.1, we need a definition of different overlap types. We shall also take a look at the data used throughout this chapter, the Davero Corpus (DC) – this corpus was introduced in Section 3.3 and will be described in more detail in Section 6.2. Next, in Section 6.3, we will turn our attention to emotional changes happening before and after overlaps as discriminating features for cooperative and competitive overlaps. We will

see what the differences in these features are in Section 6.4 before using them for the classification of the overlap type in Section 6.5. In Section 6.6, we will discuss the findings, before summarising the conducted investigation in Section 6.7.

This chapter is mainly based on the work published by my colleagues and myself [Egorow & Wendemuth 2017; Egorow & Wendemuth 2019].

6.1 Existing Works on Speech Overlaps

Overlaps are parts of a conversation where two (or more) interaction partners speak simultaneously. In the case of dyadic interactions that we are interested in here, an overlap occurs when one of the partners (the overlapper) starts speaking while the other partner (the overlappee) has not yet finished her turn. In general, there are two opposing kinds of overlaps. On the one hand, overlaps can be used to "rudely" interrupt the current speaker and take over or "steal" the turn. On the other hand, there are also so-called "non-problematic" overlaps [Schegloff 2000]:

- "terminal overlaps" at the end of an utterance with almost immediate self-stopping of the overlapper,
- "continuers" or feedback signals such as *uh-huh*,
- "conditional access to the turn", e.g. when the speaker searches for a word and the overlapper makes a suggestion,
- "choral" talk, i.e. collective greetings or laughter.

In the context of this chapter, we focus on distinguishing between cooperative and competitive overlaps. Therefore, we need a decisive description of these overlap types. The following definitions comprise the main aspects of different definitions available in the literature and were used in the same form in [Egorow & Wendemuth 2017].

Cooperative overlaps are defined by the fact that the overlapper wants to support the current speaker rather than to interrupt her [Murata 1994]. They are used to express supportive agreement or to complete an anticipated point [Yang 2001]. In case of a cooperative overlap, the overlappee should not be offended [Chowdhury et al. 2015b]. The overlapper wants to maintain the conversation and has no intention to grab the floor by taking the turn [Li 2001]. There is also no disruption of the conversational flow and the intention of the overlapper is to keep attention on the main speaker's point. The previously mentioned four "non-problematic" overlap types all belong to this class.

Competitive overlaps are defined by the fact that the overlapper competes for speech time or topic, and wants to attract attention away from the current

speaker [Goldberg 1990] or to express disagreement [Yang 2001; Li 2001]. A competitive overlap is an attempt to steal the floor, and breaks the flow of the conversation [Murata 1994]. The overlappee could perceive this overlap as problematic and offending [Chowdhury et al. 2015b].

Based on these definitions, we can see that overlaps are related to the interlocutor's internal state and can be used as a proxy to access this state. Especially when taking into account the distinction between cooperative and competitive overlaps, we can expect that there is a relation between overlaps and conflicts. This relation was already investigated, finding the overlap ratio to be a "more reliable predictor of conflict than a large and diverse set of acoustic-prosodic features" [Grezes et al. 2013]. Another investigation comes to the conclusion that overlapping speech is a key feature for aggression level prediction and outperforms conventional acoustic feature sets for this task [Lefter & Jonker 2017]. This was also already investigated in call centre interactions, finding that the duration of competitive overlaps increases in anger segments, and the amount of non-competitive overlaps increases in conversations with empathic agents [Alam et al. 2016]. But there is also a relation between overlaps and other emotional states: In the IEMOCAP data base, not only angry but also excited parts of the data set contain remarkably more overlapping speech than the sad parts [Busso et al. 2008]. This also seems to hold for HCI, as overlapping speech occurs significantly more frequently at points where the affective state of the human interaction partner is changing [Siegert et al. 2015b]. Furthermore, the nature of turn-taking in HCI settings seems to be correlated with the user's satisfaction: Features based on overlapping speech can be used for user satisfaction prediction [Chowdhury et al. 2016].

The first step towards using overlaps to enrich and improve HCI is their recognition and classification. The most obvious starting point for this task is using acoustic characteristics of the speech. In prior investigations, intensity, loudness and pitch have been shown relevant. Higher pitch and increased loudness are known to be the most important markers of a competitive overlap [French & Local 1983]. The overlap position in an utterance is also an important feature, as competitive overlaps are located "before the last major accent of the turn in progress" [Wells & Macfarlane 1998].

An overview of different existing approaches for overlap detection and classification is given in Table 6.1. Most of the authors use acoustic features such as fundamental frequency, intensity and energy. Besides that, there are also lexical features such as part-of-speech-tags, behavioural features based on movements, and others. Nevertheless, we can observe that the performance of the existing approaches is relatively low – none of the approaches presented in Table 6.1 achieves more than 75% recognition performance for the binary classification of cooperative and competitive overlaps, with most of them being

Table 6.1: An overview of current methods for automatic overlap detection and classification (cooperative versus competitive). The employed features, available information on data (including duration dur.), methods and evaluation procedures (if given by the original authors) are shown. The performance is given in terms of accuracy (Acc.), f-measure (F-1) or equal error rate (EER). This table is adapted from [Egorow & Wendemuth 2019].

Features	**Data**	**Method**	**Performance**
Task: Overlaps start / occurrence prediction			
[Shriberg et al. 2001]			
Acoustic (pauses, duration, F0, speech rate energy), lexical, context	ICSI Corpus: spontaneous, meetings, dur.: 12.5h	Decision Trees	Acc. 65%
[Lee & Narayanan 2010]			
Acoustic overlappee (energy, pitch), movement overlapper (face, head)	IEMOCAP: spontaneous, subset dur.: 6h	Logistic regression + HCRF, 6-fold CV	F-1 54%, Acc. 71%
Task: Overlaps type classification			
[Lee et al. 2008]			
Acoustic (intensity-based), hand movements, speech disfluencies	IEMOCAP: s. above	Discriminant analysis	Acc. 71%
[Oertel et al. 2012]			
Acoustic (F0, intensity), body and face movements	D64 Corpus: spontaneous, dur.: 1h	linear & RBF SVM, 3-fold CV	Acc. 63%
[Kurtić et al. 2013]			
Acoustic (F0, intensity, speech rate, pausing), placement (duration, onset, recycling)	ICSI Corpus: s. above, subset dur.: 5.5h	Decision trees, 10-fold CV	Acc. 74%
[Truong 2013]			
Acoustic (F0, intensity, energy, distribution), head movements, gaze	AMIC Corpus: spontaneous, meetings, dur.: 3.5h	RBF SVM, 10-fold CV	EER 32%
[Chowdhury et al. 2015a]			
Acoustic (prosodic, spectral, voice quality, energy), lexical, part-of-speech,psycholinguistic	ICC corpus: spontaneous, call centre, dur.: 62h	linear SVM, TDT	F-1 66%
[Chowdhury & Riccardi 2017]			
Acoustic (prosodic, spectral, voice quality, energy), lexical (bag of trigrams)	ICC corpus: s. above	ReLU DNN, TDT	F-1 68%

in the range between 60% and 70%. Even with sophisticated features and deep learning architectures trained on high amounts of data, less than 70% of the overlaps can be classified correctly [Chowdhury & Riccardi 2017]. This shows that overlap classification is not an easy task and that the search for optimal features as well as the optimal classification procedure is still an open question.

In the following sections, we will pursue a different direction for the classification of overlaps. As already mentioned, there is a connection between the overlap behaviour of an interlocutor and her emotional state, with over-

laps occurring significantly more frequently in the proximity of affective state changes [Siegert et al. 2015b]. Therefore, we want to analyse the emotional changes occurring in the utterances surrounding the overlaps. For that, we will first take a look at the data that we want to investigate, the DC.

6.2 Speech Overlaps in the DAVERO Corpus

The data set used for the investigation described below is the DC introduced in Section 3.3. The subset used here is the annotated part of the corpus, with the labels containing increases and decreases in the emotional dimensions of control and valence, as described in Section 3.3. It contains 255 overlap instances with an overall duration of 14.1 minutes (with an average overlap being 3 ± 2 seconds long). These overlaps were labelled by three annotators as competitive or cooperative. The final labels were generated via a majority vote, resulting in 192 cooperative and 63 competitive instances with high inter-rater reliability $\alpha = 0.82$.

Besides the annotation of the overlap class, additional information was annotated. The annotations consist of the role of the overlapper (call centre agent vs. call centre client), the sex of both overlapper and overlappee, and whether the overlapper continues speaking after the overlap.

The distribution of these characteristics as well as the average and total duration of the overlaps in the data is given in Table 6.2.

6.3 Emotions as Features for Overlaps

The emotional features used in this chapter are derived from the changes in the emotional dimensions of valence and control in the utterances surrounding the overlaps – two utterances before and two utterances after the overlap. Based on everyday experience, we can expect that the emotional changes of the overlapper and the overlappee are not necessarily similar and therefore we need to distinguish between the overlapper's and the overlappee's turns.

One exemplary excerpt from the annotation of the DC used in this Chapter is depicted in Figure 6.1. Here we can see two overlaps, a cooperative overlap in the first utterance (marked as *U1*), and a competitive one in the third utterance (marked as *U3*) – these two overlap instances will be further referred to as *O1* and *O2*. U2 is the "after overlap" utterance for O1, it's also the overlapper's utterance. At the same time, it is the "before overlap" utterance for O2 – and it's the overlappee's utterance in this case. U4 and U5 are the "after overlap" utterances for O2, where U4 is the overlapper's utterance, U5 is the overlappee's utterance.

Table 6.2: Overview of the overlap characteristics in the DC. The values are given for cooperative (Coop) and competitive (Comp) instances as well as all instances (Total). The table is adapted from [Egorow & Wendemuth 2019].

Characteristic	**Coop**	**Comp**	**Total**
Duration, in sec			
– average	3.0 ± 1.9	3.0 ± 2.2	3.0 ± 2.0
– total	568.1	278.7	846.8
Overlapper role			
– agent	64	35	99
– client	99	57	156
Sex Overlapper			
– female	84	51	135
– male	79	41	120
Sex Overlappee			
– female	73	39	112
– male	90	53	143
Overlapper continues			
– yes	72	51	123
– no	91	41	132

In total, we define eight emotional features: They are based on the levels of control and valence in the utterances before the overlap of both, overlapper and overlappee, as well as their utterances after the overlap. The values are discrete: -1 indicates a decrease, $+1$ indicates an increase, 0 indicates no change. Furthermore, ? indicates a missing value. The features are listed in Table 6.3.

One important point regarding these features is that not all utterances have an emotional annotation, as already mentioned in Section 3.3. This leads to a certain amount of missing values and therefore sparse features. This has two causes. First, the annotation of the emotional changes itself is sparse, since it is based on a majority vote that not always resulted in a label for an utterance: The inter-rater reliability between the five emotional raters is rather low ($\alpha = 0.3$), as described earlier in Section 3.3. Second, the DC also contains parts where several overlaps occur one after another (as in the example depicted in Figure 6.1), so there is not always a "last utterance" before an overlap or "first utterance" after an overlap. Out of the 255 investigated overlaps, 124 contain at least one missing value, and there are 344 missing values in total.

As mentioned in Section 6.2, additional socio-cultural (sex and roles of the interaction participants) as well as interaction-related (continuing after overlap) information was annotated. This was done to account for cultural as well

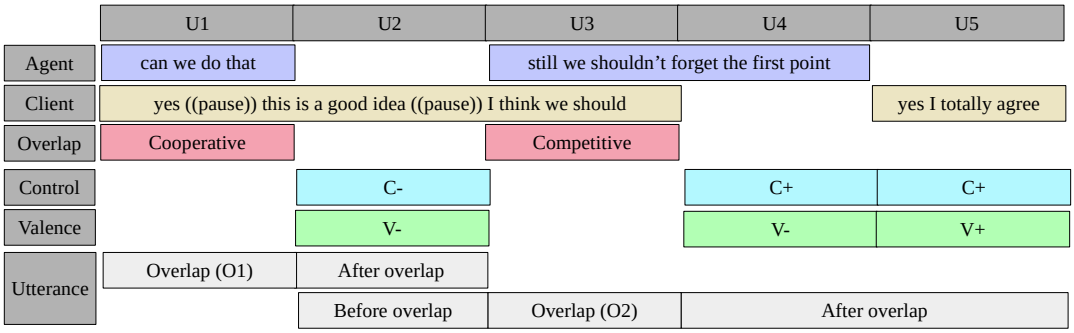

Figure 6.1: An annotation example for five consecutive utterances U1–U5. The annotated tiers are (from top to bottom): Agent's turns, Client's turns, Overlap class, labels for Control and Valence ("-" indicates a level decrease and "+" indicates a level increase), Utterance function. The figure is adapted from [Egorow & Wendemuth 2019].

Table 6.3: Overview of the eight emotional features.

Feature	Emotion	Interlocutor	Turn	Remark
C_O_B	**C**ontrol	**O**verlapper	**B**efore overlap	Always penultimate turn before the overlap
V_O_B	**V**alence			
C_O_A	**C**ontrol		**A**fter overlap	Either directly succeeding the overlap or the next one
V_O_A	**V**alence			
C_P_B	**C**ontrol	Overla**P**pee	**B**efore overlap	Always directly preceding the overlap
V_P_B	**V**alence			
C_P_A	**C**ontrol		**A**fter overlap	Either directly succeeding the overlap or the next one
V_P_A	**V**alence			

as gender differences in the usage of overlaps known from literature [Murata 1994; Makri-Tsilipakou 1994]. This resulted in four additional features that were already described in Section 6.2 and are shown in Table 6.4.

For the competitive overlap O2 from Figure 6.1, assuming the agent to be male and the client to be female, the following features can be derived:

- C_O_B = ?, V_O_B = ?, C_P_B = −1, V_P_B = −1
- C_O_A = +1, V_O_A = −1, C_P_A = +1, V_P_A = +1
- O_R = agent
- O_C = yes
- O_S = male, P_S = female

Table 6.4: Overview of the four socio-cultural and interaction-related features.

Feature	Description	Possible values
O_R	**O**verlapper **R**ole	Agent, Client
O_C	**O**verlapper **C**ontinues	Yes, No
O_S	**O**verlapper **S**ex	Female, Male
P_S	Overla**P**pee **S**ex	Female, Male

- Class = competitive

6.4 Analysing the Statistical Differences

The starting point of this investigation is the hypothesis that overlaps and emotions are correlated, which might correspond to the common "gut feeling" that competitive overlaps are related to control, whereas cooperative overlaps are related to valence. An overlapper who successfully "steals the turn" might experience an increasing feeling of control and valence, whereas the overlappee's feelings of control and valence might decrease – this does not have to occur necessarily in all cases (especially not in those resulting in a "fight" for the next turn), but one can imagine that it applies in the general case that we are interested in here. The DC data with its emotional annotation provides an insight into the relationship between the overlaps type and the emotional changes in the interlocutors' utterances surrounding the overlap.

In other words, the research question here was whether the emotional levels change significantly in the proximity of the overlap with respect to the nature of the overlap, e.g. whether the level of control of an overlapper rises in significantly more cases after a competitive overlap compared to after a cooperative overlap. To answer this question, we can turn to statistical testing and use *Fisher's exact test* explained in detail in Section 2.2.2.

In order to illustrate the application of Fisher's exact test to the current research question, we can take a look at the previously introduced example, where we asked whether the level of control of an overlapper rises in significantly more cases after a competitive overlap than after a cooperative overlap.

In total, the investigated data set contains 255 overlaps, 92 of which are competitive and 163 cooperative – this constitutes the prior class distribution. There is a control increase in the overlapper's utterance after 55 of the competitive overlaps and 51 of the cooperative overlaps. According to the class distribution, a control increase would be expected only after 38 competitive overlaps and 68 cooperative overlaps. Fisher's exact test can now be used to prove whether the difference between the actual class distribution and the expected class distribution is significant. This would mean that a control in-

crease occurs more often after a competitive overlap than after a cooperative overlap.

6.4.1 Emotions of the Overlapper

First, we analyse the results of the Fisher's exact test on the overlapper's emotional features. For each feature, we want to see whether there is a correlation for increase versus no increase and the class of the overlap as well as decrease versus no decrease and the class of the overlap.

The null hypothesis to be tested is that there is no difference in the number of increases or decreases of an emotional level between a competitive and a cooperative overlap. This would mean that the distribution of the increases or decreases does not differ significantly from the prior class probability. This hypothesis is rejected by the Fisher's exact test for all four emotional features of the overlapper – the results of the test show that there are significant correlations for all of them.

An overview of the distribution of the feature values as well as the test results is shown in Table 6.5. In order to understand the results of the test, we can return to the example we have discussed in the beginning of this section. This example maps to the fifth row of Table 6.5 – the increase in the overlapper's level of control after an overlap (C_O_A = +1). Here, we can see that the overlapper's level of control rises in 51 cases after a cooperative overlap (second column) and in 55 cases after a competitive overlap (third column) – it is significantly lower than expected for the class Coop with a p value of < 0.01 (fifth column).

In the same manner, the relations between the other features and the nature of the overlap can be deducted from Table 6.5.

Table 6.5: Results of Fisher's exact test for the emotional features of the overlapper. For each feature, the number of instances in both overlap classes and the test results on the difference of the distribution compared to the prior class distribution are shown. The table is adapted from [Egorow & Wendemuth 2019].

Emotion	**Coop**	**Comp**	**Total**	**Fischer's test**	**Significance**
C_O_B = +1	60	40	100	not significant	–
C_O_B = −1	32	7	39	higher for Coop	$p < 0.05$
V_O_B = +1	56	10	66	higher for Coop	$p < 0.01$
V_O_B = −1	33	31	64	lower for Coop	$p < 0.05$
C_O_A = +1	51	55	106	lower for Coop	$p < 0.01$
C_O_A = −1	37	9	46	higher for Coop	$p < 0.05$
V_O_A = +1	47	11	58	higher for Coop	$p < 0.01$
V_O_A = −1	34	51	85	lower for Coop	$p < 0.01$

We can summarise the results in the following way:

- for C_O_B, the number of increases does not differ significantly depending on the overlap class, the number of decreases is significantly higher for cooperative and lower for competitive overlaps ($p = 0.0109$),
- for V_O_B, the number of increases is higher for cooperative overlaps ($p < 0.0001$), the number of decreases is higher for competitive overlaps ($p = 0.0237$),
- for C_O_A, the number of increases is higher for competitive overlaps ($p = 0.0107$), the number of decreases is higher for cooperative overlaps ($p < 0.0001$),
- for V_O_A, the relation is the same as for V_O_B, the number of increases is higher for cooperative overlaps ($p < 0.0001$), the number of decreases is higher for competitive overlaps ($p < 0.0001$).

An overview of these results is presented in Figure 6.2.

These results provide us with two major findings. First, regarding the overlapper's level of valence, we can state that it increases in significantly ($p < 0.0001$) more cases before and after a cooperative overlap than before and after a competitive overlap. It also decreases in significantly ($p < 0.05$) more cases before and after a competitive overlap than before and after a cooperative overlap. Second, regarding the overlapper's level of control, we can state that it increases significantly ($p < 0.05$) more often after a competitive overlap than after a cooperative overlap. It also decreases significantly ($p < 0.05$) more often before and after a competitive overlap than before and after a cooperative overlap.

Based on this, we can highlight the following finding: If a person overlaps competitively, this person experiences an increasing level of control, and if a person produces a cooperative overlap, this person experiences an increasing level of valence.

6.4.2 Emotions of the Overlappee

Previously, we have seen how the emotions of the overlapper are related to the overlap class. Now we focus on the emotions of the overlappee and on the results of the Fisher's exact test on the corresponding features. This will be done in the same way as for the overlapper.

Again, we can take a look at the results of the Fisher's exact test presented in Table 6.6. The results can be summarised in the following way:

- for C_P_B, the number of increases does not differ significantly depending on the overlap class, the number of decreases is significantly higher for cooperative and lower for competitive overlaps ($p = 0.0024$),

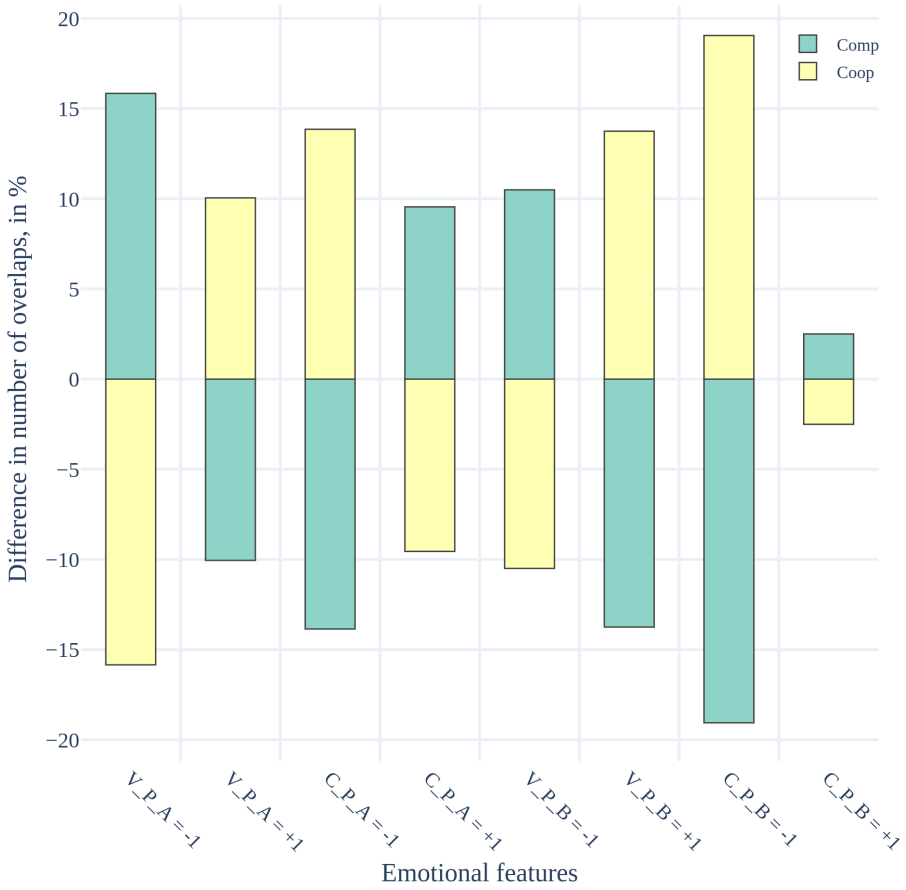

Figure 6.2: Difference between the expected and the observed number of emotional changes in the overlapper. The difference is shown for an increase (feature value = +1) or a decrease (feature value = −1) in the emotions of the overlapper, depending on the overlaps class, in % (0% difference corresponds to an equal distribution over the two classes). The figure is adapted from [Egorow & Wendemuth 2019].

- for V_P_B, the number of increases is higher for cooperative overlaps (p = 0.0001), the number of decreases is higher for competitive overlaps (p = 0.0310),
- for C_P_A, the number of increases is higher for competitive overlaps (p = 0.0115)), the number of decreases does not differ significantly,
- for V_P_A, the relation is the same as for V_P_B, the number of increases is higher for cooperative overlaps (p = 0.0171), the number of decreases is higher for competitive overlaps (p = 0.0431).

An overview of these results is presented in Figure 6.3.

The results allow us to conclude that the overlappee's level of control often decreases before cooperative and after competitive overlaps. Furthermore, the

Table 6.6: Results of Fisher's exact test for the emotional features of the overlappee. For each feature, the number of instances in both overlap classes and the test results on the difference of the distribution compared to the prior class distribution are shown. The table is adapted from [Egorow & Wendemuth 2019].

Emotion	Coop	Comp	Total	Fisher's test	Significance
C_P_B = +1	78	49	127	not significant	–
C_P_B = −1	39	8	47	higher for Coop	$p < 0.01$
V_P_B = +1	80	23	103	higher for Coop	$p < 0.01$
V_P_B = −1	39	34	73	lower for Coop	$p < 0.05$
C_P_A = +1	56	47	103	lower for Coop	$p < 0.01$
C_P_A = −1	28	8	36	not significant	–
V_P_A = +1	54	19	73	higher for Coop	$p < 0.05$
V_P_A = −1	25	27	52	lower for Coop	$p < 0.01$

overlappee's level of valence increases before and after cooperative overlaps and decreases before and after competitive overlaps.

This leads us to an important finding and an interesting conclusion: The results suggest that if a person is interrupted competitively, this person has a decreasing feeling of control, and in the case of a cooperative overlap, the person has an increasing feeling of valence.

This relation between the dialogue course and the interlocutor's inner state can help to develop novel design ideas for HCI systems. One possible idea is to incorporate cooperatively overlapping feedback into HCI to increase the user's valence and to enhance the interaction experience.

6.4.3 Other Features

As already stated, there are also other features that should be taken into consideration when investigating overlaps, namely the socio-cultural and interaction-related features. Overall, the Fisher's exact test applied to these features shows that there are no significant correlations.

For the feature O_R, the overlapper's interaction role, it can be stated that whether the overlapper is a client or an agent does not correlate with the overlap class. For the feature O_C, it can be stated that the number of cooperative overlaps where the overlapper continues speaking does not differ significantly from the number of competitive overlaps where the overlapper continues speaking. For the features O_S and P_S, it can be stated that the sex of both the overlapper and the overlappee does not differ significantly between cooperative and competitive overlaps. Furthermore, the data also do not show significant differences between these two features, so it can be

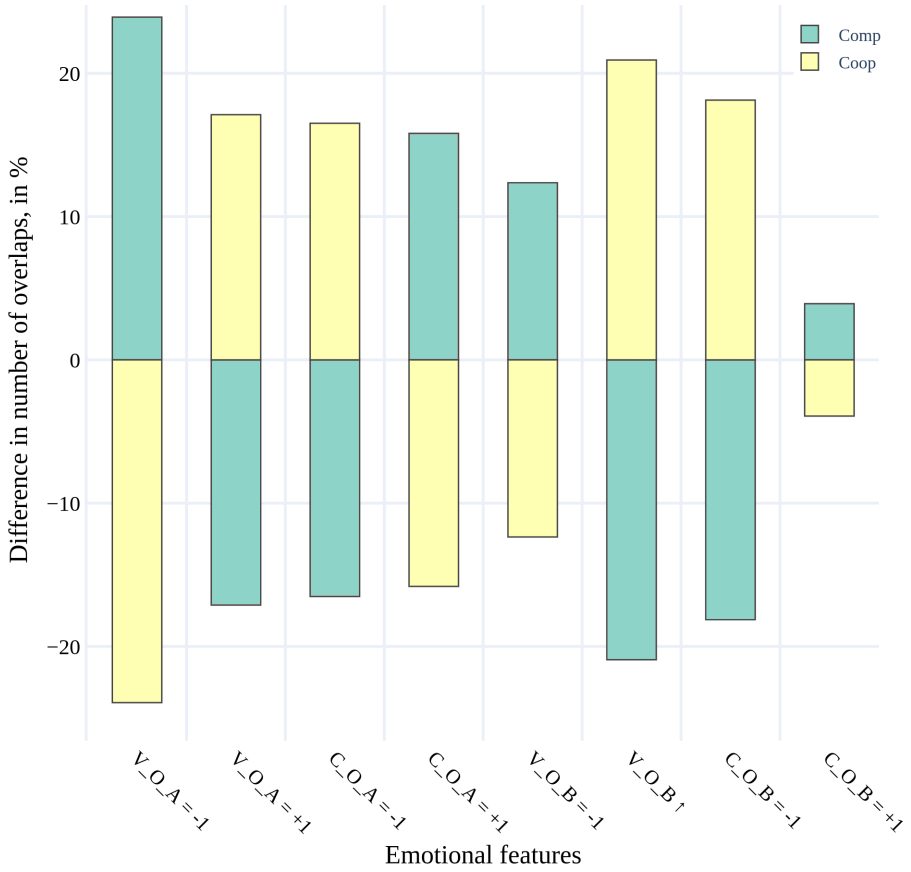

Figure 6.3: Difference between the expected and the observed number of emotional changes in the overlappee. The difference is shown for an increase (feature value = +1) or a decrease (feature value = −1) in the emotions of the overlappee, depending on the overlaps class, in % (0% difference corresponds to an equal distribution over the two classes). The figure is adapted from [Egorow & Wendemuth 2019].

concluded for both overlap types that men do not interrupt women more often or vice versa.

This finding means the following: In the case of the DC data, the sociocultural and interaction-related features do not differ significantly depending on the overlap class. Therefore, we can conclude that they do not contribute to the discrimination of cooperative and competitive overlaps.

6.5 Overlap Classification using Emotional Features

In the previous section, we have already seen that there is a significant relationship between the overlap class and the emotional changes in the surrounding

turns. In this section, we want to prove whether these changes are useful for a classification of the overlaps using the emotional changes as features.

6.5.1 Emotional and Acoustic Features

In order to evaluate the performance of the previously introduced emotional features, we should not only look at their discriminating power but also compare them to "conventional" acoustic features as used in the literature.

In total, we will compare five different feature sets, three of them based on the emotional changes described above and two on acoustic features. Feature set *Emo1* comprises only the overlapper's emotional features. One reason to use only these features is because they were associated with the largest differences in the statistical tests, the other reason is to ensure a subject-independent evaluation, which will be further explained in the next section. The second feature set, further referred to as *Emo2*, comprises the emotional features of both, the overlapper as well as the overlappee. The third feature set, feature set *Emo3*, contains all emotional and interaction-related features of both interaction participants. The two acoustic feature sets are derived from the *emobase* set. Both of them contain a subset of the *emobase* features, namely only the 38 features derived from intensity and fundamental frequency, since these two LLDs have been proven to deliver the best classification performance in the literature as well as on the DC data [Egorow & Wendemuth 2017]. The first of the two acoustic feature sets, further referred to as *Aco1*, contains the features calculated from the utterances before and after an overlap that are concatenated to a feature vector of 76 features. The other acoustic feature set, *Aco2*, contains the features calculated over the utterance before the overlap, the overlap itself, and the utterance after the overlap as one audio sequence – this procedure results in 38 features.

6.5.2 Implementing a Leaving-one-out Evaluation

In order to prove that the emotional features are suitable for the classification of overlaps, we will use them in a classification setup and compare them to the "conventional" approach based on acoustic features.

First of all, we need to ensure the generalisation ability of the classification by implementing an appropriate evaluation routine, as already discussed in detail in Section 2.2.7. In order to analyse the classification performance on different interaction participants, we will implement Leave-One-Subject-Out (LOSO) evaluation which is introduced in Section 2.2.7.

To implement a LOSO evaluation, we need n folds for the n subjects, with each fold containing $n - 1$ subjects for training and the remaining subject for

the test. In the case of the DC data, this would correspond to 48 pairs of training and test sets, one for each of the dialogues. But there is still one problem: The 48 dialogues share four agents. Implementing the evaluation in the way described above would lead to a "leaving-one-client-out" rather than "leaving-one-subject-out". This problem can be solved by implementing a slightly different approach. We need to ensure that we use only the data of the clients, and therefore choose only those overlaps where the overlapper is the client. Furthermore, we use only the overlapper's (i.e. the client's) emotional features. In this setting, the classification is performed only on the features obtained from subjects occurring in one dialogue each, which makes the evaluation truly subject-independent.

Unfortunately, five of the dialogues of the DC contain only overlaps where the agent acts as the overlapper – these dialogues had to be excluded, resulting in a total of 43 dialogues and therefore folds. This decreases the amount of overlap samples to 156 overlaps, 99 cooperative and 57 competitive, in 43 dialogues. This setting is further referred to as "leaving-one-overlapper-out" (LOOO) evaluation.

Nevertheless, there is also a beneficial side effect achieved by this evaluation. In general, we can assume that in HCI applications, it would more likely be the human interaction partner interrupting the system. This means that the overlapper would be the human in most of the cases. As we turn to emotional changes as features for the classification, it is only logical to use the data of the overlapper even in HHI, since in HCI, there would be no emotional data of the overlappee (i.e. the computer system) available.

There is also another problem regarding the evaluation procedure – the evaluation metrics, that we have already introduced in general in Section 2.2.7. In a LOSO setting, one would normally calculate the recall R , precision P and f-measure F_1 for each of the classes in each of the folds, and obtain their unweighted averages UAR, UAP, UAF for each of the folds. The following equations illustrate this process for R for the evaluation of k classes (TP_j and FP_j correspond to true positives and false positives of the class j, respectively):

$$R_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j} \tag{6.1}$$

$$\mathrm{UAR} = \frac{\sum_{j=1}^{k} R_j}{k} = \frac{\sum_{j=1}^{k} \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FP}_j}}{k} \tag{6.2}$$

In the next step, one would average these values over the n folds, obtaining the average value $\overline{UAR}$ over the folds:

$$\overline{\text{UAR}} = \frac{\sum_{i=1}^{n} \text{UAR}_i}{n} = \frac{\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j}}{k}}{n} \tag{6.3}$$

But in the case of the DC data, some of the folds contain only instances of one class. In this case, the number of true positives for the missing class is zero, resulting in $R = 0$ for this class and subsequently UAR = 0, regardless of the performance of the classification.

The distribution of the classes over the 43 folds is illustrated in Figure 6.4. On average, every dialogue contains 2.3 cooperative and 1.3 competitive overlaps – but the median for the competitive class is zero, due to the distribution skewness. In Figure 6.4, we can see that 27 of the dialogues (and, respectively, folds) have no instances of the competitive class. In the same way, there are four dialogues without any instances of the cooperative class.

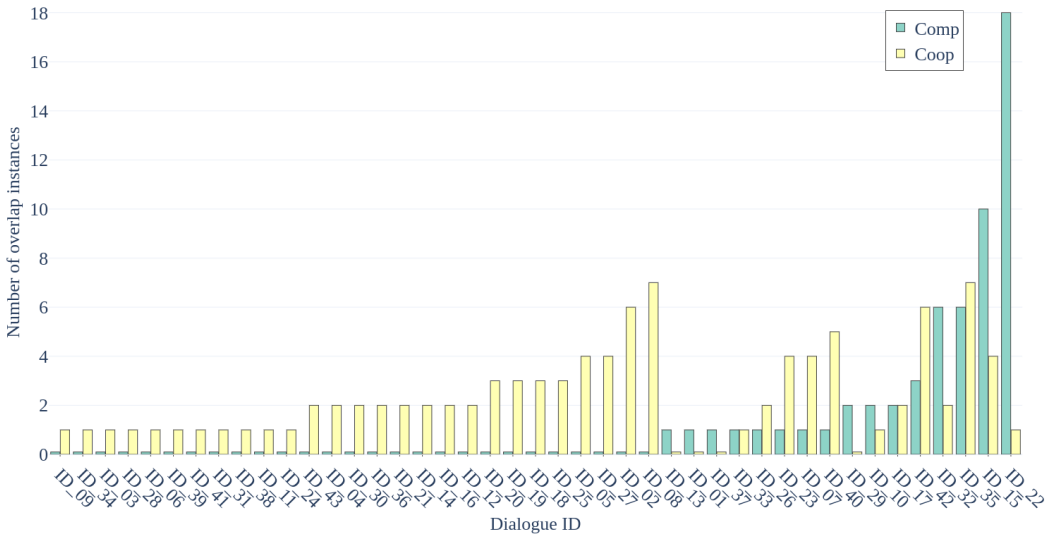

Figure 6.4: Distribution of overlap classes over the 43 dialogues. The figure is adapted from [Egorow & Wendemuth 2019].

To solve this problem, the recall is calculated over all m instances of all classes of all folds. For this, the classification results are accumulated over all folds: The total number of true positives, false positives, true negatives and false negatives are calculated as the sum of all folds. These values are then averaged over the classes to obtain the unweighted average over classes. The recall R_j for class j is still calculated in the same way as above, but for all folds, leading to the following equation for the average recall $\overline{R_j}$:

$$\overline{R_j} = \frac{\sum_{i=1}^{n} \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}}{n} \tag{6.4}$$

In the next step, the obtained $\overline{R_j}$ values are averaged over all k classes to obtain the average UAR value:

$$\overline{\mathrm{UAR}} = \frac{\sum_{j=1}^{k} \overline{R_j}}{k} = \frac{\sum_{j=1}^{k} \frac{\sum_{i=1}^{n} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}}{n}}{k} \tag{6.5}$$

Analogous procedures are implemented to calculate the values for UAP and UAF.

These procedures allow us to calculate the average values over all folds. Nevertheless, it is not possible to obtain the standard deviation over the folds, since, as explained above, we cannot calculate the individual values of the folds. Therefore, only the average values can be reported.

6.5.3 Classification Setup for Missing Values

Before turning to the classification itself, we have to observe one further restriction, namely missing values for some of the features. The DC data contains at least one missing feature value in 76 of the 156 instances used for the classification. One way to solve this problem is to employ feature replacement methods [Liu et al. 1997]. But in the case of emotional features, replacing or otherwise processing the missing values (such as averaging the features or presuming constant values) would lead to changing the feature space. Another solution is to use a classifier that can work with missing feature values – such as the Naïve Bayes classifier described in detail in Section 2.2.6.

The Naïve Bayes classifier implementation used for the presented investigation was provided by KNIME [Berthold et al. 2007]. As described in the previous subsection, the classification was carried out in a LOOO manner with subject-independent training and test sets, leading to 43 distinct classification models. There was no fine-tuning process of the classifier's parameters to maintain comparability, since in the worst case, the fine-tuning would have lead to 43 sets of hyperparameters. Therefore, the classifier was trained using default hyperparameter values as provided by the KNIME implementation.

The classification was performed in exactly the same manner for all five introduced feature sets.

6.5.4 Analysis of the Classification Results

A comprehensive overview of the classification results is presented in Table 6.7 – here, the results are given in terms of recall, precision and f-measure for each of the classes as well as the unweighted average, for all five feature sets.

Table 6.7: Classification results in terms of recall, precision and f-measure for the each of the classes and their unweighted average (Coop., Comp. and UA, respectively), in %. The average results are highlighted. The table is adapted from [Egorow & Wendemuth 2019].

Feature Set	Class	Recall	Precision	F-Measure
Emo1	Coop	90.91	75.00	82.19
	Comp	47.37	75.00	58.06
	UA	69.14	75.00	70.13
Emo2	Coop	89.90	71.77	79.82
	Comp	38.60	68.75	49.44
	UA	64.25	70.26	64.63
Emo3	Coop	92.93	70.23	80.00
	Comp	31.58	72.00	43.90
	UA	62.24	71.11	61.95
Aco1	Coop	77.91	64.57	70.61
	Comp	23.95	37.93	29.36
	UA	50.93	51.24	50.00
Aco2	Coop	88.96	68.06	77.13
	Comp	26.09	57.14	35.82
	UA	57.52	62.61	56.47

The first finding supports our initial hypothesis: The feature sets based on the emotional changes can be used for overlap classification and clearly outperform the acoustic ones. The best results are achieved using feature set *Emo1*, which contains only the emotional features of the overlapper, with a UAR of 69% and a UAF of 70%. Interestingly, adding either the features of the overlappee or interaction-related features impairs the results: Feature set *Emo2* and *Emo3* deliver a UAR of around 64% and 62%, respectively. Still, these results surpass those achieved using the acoustic features.

The feature set *Aco1* does not outperform chance level (UAR of roughly 51%), *Aco2* performs slightly better with a UAR of around 58%, which is rather disappointing. Interestingly, it is the competitive class that causes problems for the acoustic feature sets, since the recognition performance for the two classes differs remarkably. One possible explanation for this is the bias in the data, since the data contains more instances of the cooperative class compared to the competitive class. To resolve this, a downsampling of the overrepresented class can be implemented. Applied to our case, this strategy was not able to solve the problem. After downsampling the cooperative class to an approximately equal distribution, the recall for the competitive class indeed increased, but at the same time, the recall for the cooperative class decreased to a greater extent. As a result, the UAR for *Aco1* stayed at 50%, and the UAR for *Aco2* decreased to 54%. Another possible explanation for the disappointing

performance of the acoustic features is the low quality of the recordings, which contain a high amount of noise and other artefacts, since the material was recorded in a real-life telephone setting.

Coming back to the emotional features, we can also see another interesting finding: The best classification results are achieved using the overlapper's features only. It is not surprising that the interaction-related features do not have additional value and, moreover, distort the results, since the conducted Fisher's exact tests described in Section 6.4 showed no significant differences in these features with respect to the overlap class. But the fact that the feature set *Emo2* delivers a lower classification performance than *Emo1* is unexpected, since Fisher's tests showed a significant difference between cooperative and competitive overlaps. However, the difference was smaller than the one observed on the overlapper's features – together with the classification performance, this suggests that the information provided by the overlappee's features is already contained in the overlapper's features, leading to a performance drop due to redundancy.

6.6 Concluding Remarks on Emotional Features

The experimental results allow us to draw several interesting conclusions that will be discussed in detail below. First of all, we have seen that there is a relationship between the emotions in the interlocutor's utterances and her overlapping behaviour. This finding corresponds to the common "gut feeling" that being interrupted can decrease one's valence whereas interrupting someone can increase one's feeling of control. Furthermore, this finding also supports the hypothesis that angry segments of interactions require more coordination than interactions with empathic speakers [Alam et al. 2016].

A more interesting and novel finding is that this relationship can be used for overlap classification. So far, most of the approaches for the classification rely on acoustic, lexical and conversational features (such as intensity, fundamental frequency, but also turn taking, etc.). But in Section 6.5, we have seen that emotional features also deliver useful classification results – in fact, although the results are not directly comparable to those found in the literature due to different data sets and evaluation procedures, they are in the same range of around 70% recognition accuracy. These values are still arguably low. Furthermore, this approach cannot be used for online classification, taking into account that the classification relies on features obtained from utterances after the overlap.

Nevertheless, there is another application for this procedure: The classification results can be used to improve the overall interaction experience by

accumulating the results over the course of an interaction. One potential application is the use in personal conversational assistants. For example, such assistants could detect an increase in cooperative overlaps and decide whether it is the right moment to engage the user in some cooperative activities. Furthermore, the classification performance could be improved by fine-tuning the recognition model – especially in the case of personal assistants always working with the same person, the training of the classifier could be adapted to their user's specific behavioural patterns.

Since the presented approach uses only the overlapper's utterances, it can be applied in HCI settings, assuming that the overlapper will mostly be the human interaction partner. Nevertheless, it should be noted that the findings presented here are based on HHI data and we can only hypothesise that they are transferable to the domain of HCI. Having a robust classification system for cooperative and competitive overlaps can be of use in different application scenarios. One especially interesting thought in this regard is to incorporate the behaviour of the overlappee into system behaviour – a technical system frequently interrupted by the user could react in the "human" way by showing a decrease in valence, just as human interlocutors do in the presented data. At the same time, a system could provide slightly overlapping positive feedback in order to achieve an increasing valence in the human interaction partner. These ideas are in tune with other findings on the behaviour of technical systems, such as introducing cooperative and dominant virtual agents [Straßmann et al. 2016]. The research in this area has started recently, for example implementing socially cooperative assistance systems for the elderly care, and evaluating the persuasive abilities of dominant agents for both, younger and elderly subjects [Kopp et al. 2018; Rosenthal-von der Pütten et al. 2019].

However, in order to use emotional features for automatic overlap classification, these features need to be extracted in an automatic way. So far, it seems to be a difficult task. In the case of the DC data, there are several problems. The main problem is the low amount of data as well as the low audio quality. Furthermore, the data distribution and data annotation are also suboptimal. These issues lead to expectedly low recognition results: The emotion recognition system implemented for this data achieved less than 80% average accuracy for the valence dimension and could not outperform chance level in the control dimension [Siegert & Ohnemus 2015]. Even with the latest approaches such as deep and recurrent learning, emotion recognition results are not sufficiently high for a reliable automatic annotation. The results achieved using these methods on the benchmark IEMOCAP corpus are in the range of 60% to 68% for the four emotional classes [Lee & Tashev 2015; Kurpukdee et al. 2017; Etienne et al. 2018b].

Once the automatic emotion recognition delivers better results, obtaining emotion-related features in a fully automatic way will be feasible, enabling us to process more data. The results presented here are obtained considering only 255 overlap instances due to the demanding manual annotation of emotions. For the development of useful applications, more data could provide even more interesting results – for example, it could lead to discovering other relations between the emotional state of an interlocutor and her overlapping behaviour, and, subsequently, allow us to develop more powerful classification systems.

6.7 Summary of the Chapter

In this chapter, we turned our attention to speech overlaps as markers for *cooperativeness*. We have seen that the emotional changes in the two dimensions of valence and control in the utterances surrounding the overlap differ depending on the overlap class. In some cases, these differences are significant – especially regarding the emotions of the overlapper. We have also seen how these emotional changes can be used as powerful features for overlap classification, outperforming "classic" acoustic features.

Using the emotional features extracted from the overlapper's utterances, we were able to achieve around 70% UAF for the task overlap classification. In contrast, using acoustic features such as intensity and fundamental frequency, as proposed in the literature, only 50% to 58% UAF could be achieved on the investigated data. We have discussed the potential of these findings and analysed possible application scenarios, but also considered the shortcomings, since, so far, the features have to be obtained manually.

In the next chapter, we will recapitulate and summarise the overall work accomplished in this thesis in order to highlight the findings and discuss possible future developments.

Chapter 7

Conclusion

Contents

THIS thesis was dedicated to the investigation of three interaction-related internal interlocutor states of *trouble*, *satisfaction* and *cooperativeness* in naturalistic or real-life interaction. In Chapter 1, we introduced the field of affective computing. We motivated the choice of these specific interlocutor states in Section 1.1, in Section 1.2, we reviewed the research on interlocutor state recognition, in Section 1.3 we identified several challenges that need to be solved in order to achieve a better understanding of human interaction. Having established the research objectives in Section 1.4, we developed approaches to reach them. We reviewed the basics of the methodology for solving this task in Chapter 2, discussing the important theoretical concepts in Section 2.1 and the technical details of currently applied methods in Section 2.2. In Chapter 3, we examined the importance of data for the task of data-driven modelling: We derived general requirements for data in Section 3.1 and presented existing data sets providing the empirical base for the conducted research in Section 3.2 and Section 3.3. In the Chapters 4, 5 and 6, we investigated the interlocutor states of *trouble*, *satisfaction* and *cooperativeness*, respectively. We will return to the detailed results of these chapters in Section 7.1.

In the current and final chapter, we conclude the thesis, and therefore return to its overall aim in order to reflect the progress towards it, to summarise the realised work, evaluate the accomplished results and draw conclusions from the gained insight. In Section 7.1, we will recapitulate the achieved results regarding the interlocutor state recognition. In Section 7.2, we will analyse the results and evaluate whether we have reached the overall goal of the thesis. In Section 7.3, we will discuss the contribution of this thesis to the state of the art. In Section 7.4, we will conclude by proposing possibilities for future research.

7.1 Results on Interlocutor State Recognition

The major part of this thesis investigated interaction-related interlocutor states in different interlocutor signals, namely speech with its spectral, prosodic and emotional information, physiological signals in form of EMG, RSP and SC, and 3D movements and postures of the upper body.

In Chapter 4, we focussed on the *trouble* state. For the recognition of this state, we used all three mentioned modalities, namely acoustic data, physiological signals, and upper-body postures. Acoustic data – be it speech itself or paralinguistic information – are relatively easy to obtain and seem to gain importance in the days of omnipresent voice assistants such as Apple's Siri, Microsoft's Cortana and Amazon's Alexa. However, for our task of binary *trouble* classification, we found that the audio signal alone delivered unsatisfactory performance compared to other available approaches, with around 64% UAF, outperforming the chance level by only 14%, as described in Section 4.3. Physiological signals, on the other hand, are rather difficult to get hold of, often requiring physical contact and special equipment and sensors. At the same time, they seem to have the most direct link to the ground truth – the classification delivered impressive results of around 90% UAF, as described in Section 4.4. Bearing in mind that the labels serving as the basis for the classification are generated by external annotators who cannot directly access the internal state of the interlocutor and therefore are prone to errors, it can be assumed that the classification results represent the "gold standard". In the last part of the presented work on the *trouble* state, we used the interlocutor's upper-body postures. Using features based on the 3D coordinates of Kinect recordings, we achieved a classification performance of around 74% UAF, as described in Section 4.5. This seems to be the most promising direction for future research which combines the readily available consumer sensors with a relatively high performance and easy installation.

In Chapter 5, we investigated the recognition of the state of *satisfaction* implementing binary classification of satisfied and dissatisfied interlocutor utterances. Using only the acoustic content of a set of selected utterances without considering the linguistic content, we were able to achieve around 87% UAF for this task, as described in Section 5.3. This performance could neither be improved by focussing on the voiced part of the utterances nor by implementing a feature selection routine. The negative effect of the implemented voice activity routine shows that even unvoiced parts contribute to the classification of *satisfaction* and need our attention. The negative effect of the feature selection routine confirms that our understanding for the relevance of features is still limited: Selecting the top 10% of the features with the highest information gain led to a substantial performance drop, although the same procedure delivered a performance boost in a similar task of emotion recognition. This

is a hint that the quest for the best feature set for the recognition of affective states remains unsolved. However, the relatively high UAF of 87% allows the development of *satisfaction* assessment systems based on voice alone.

In Chapter 6, we dealt with speech overlaps as markers for the state of *cooperativeness*. We developed a classification approach for cooperative and competitive overlaps based on the emotional content of the surrounding utterances. After the comparison of different feature sets and setups, we were able to achieve around 70% UAF for this binary task in the best case, using the emotional changes in valence and control of the overlapper as features, as described in Section 6.5. Although the developed approach cannot be used for online classification of overlaps, it can be applied to monitor the entire course of interaction, for instance, to prevent the dialogue from becoming increasingly competitive. Furthermore, the statistical analysis of the data described in Section 6.4 confirmed that there exists a long suspected relation between overlapping behaviour and the levels of valence and control in both, the overlapper as well as the overlappee. This allows us to develop new approaches for affective reactions in HCI. One such idea is introducing slightly overlapping cooperative feedback signals to increase the valence level in the human interaction partner.

The achieved results show that it is possible to recognise the three interlocutor states in question in naturalistic data with relatively high performance depending on the used modality. For the states of *satisfaction* and *cooperativeness*, acoustic data of the interlocutor seems to be sufficient for real-life applications, whereas for the state of *trouble*, 3D data of the interlocutor offers the best trade-off between performance and usability. However, the development of classification models was not the only aim of this thesis. We will review the progress achieved in solving the methodological issues in the next section.

7.2 Results on Methodological Issues

The work presented in this thesis shows that a recognition of the three considered interlocutor states is possible, delivering a relatively high performance depending on the chosen state and modality, as we have seen in Section 7.1. However, in Section 1.3, we have identified three methodological questions that we addressed throughout the thesis.

First, we focussed on establishing the appropriate data. This was done in Chapter 3 by developing a set of general requirements for data that can be used in investigations aiming at understanding natural interaction, and applying these criteria to three existing data sets: the LMCv1 and the LMCv2

comprising naturalistic HCI, as well as the DC, a real-life HHI corpus. Working with naturalistic data poses certain challenges that we discussed in Section 3.1:

- the generation of data in natural surroundings struggling with low expressiveness,
- possible technical issues such as differences in the experimental conditions and noise as well as asynchronous data streams,
- the necessity to obtain the ground truth externally and the resource-consuming process of annotation,
- the storage of sensitive data while maintaining security and privacy protection, etc.

However, if we intend to develop systems to be applied to real-life scenarios, processing naturalistic data is the only possible way.

The second question that we addressed concerned the modalities and features to be used for the classification approaches. We discussed the modalities and their characteristics, especially their obtainability and acceptability as well as their relation to the ground truth, for the task of *trouble* recognition in Chapter 4. As already mentioned, the acoustic signal alone did not deliver satisfactory results for this task. However, for the task of *satisfaction* recognition described in Chapter 5, the acoustic signals seem to be the means of choice. The practical problems should also not be neglected. If the final objective is to develop a technical system recognising its interaction partner's internal state, it needs to capture all information necessary for the classification at runtime, i.e. to record and store the raw signals, to extract the features, to execute the classification, and – last but not least – react accordingly. Reliable sensors and fast feature extraction routines are essential for this task. Especially for embedded systems, feature sets with thousands of features might be unfeasible. This aspect must be kept in mind when choosing the modalities and features. Another issue is the usability of such a system and the trade-off between the problems linked to the use of the system and its added value. There is no doubt that wearing a gel-pad-based EMG sensor while interacting with an assistive system is hardly acceptable, even if this improves the performance of the system. But other factors also have an influence – a system that is acceptable for home use might not be as acceptable in the public space, whereas the attitudes towards cloud-based services might differ from on-premise systems. This means that there is no "one-to-rule-them-all" system, with each problem requiring an individual course of action.

The final subject we identified for our research was the correct evaluation of classification approaches. For this, we compared two different evaluation settings, TDT and LOSO, for the same task in Section 4.5. We found that both setups have advantages and disadvantages and allow for different statements

about the classification performance. Furthermore, we presented a way to implement these evaluation procedures even in adverse conditions in Section 6.5 for the task of overlap classification: the subject-independent evaluation of dialogues with overlapping interlocutors as well as the calculation of evaluation metrics for highly unbalanced class distributions.

In this way, the conducted research contributes to the current scientific discussion on these three methodological questions. In the next section, we will summarise this thesis's overall contributions to the state of the art in the field of affective computing.

7.3 Contribution to the Scientific Field

The thesis offers insight into the topic of interlocutor state recognition, presenting classification approaches based on three different kinds of interlocutor signals. For each of the considered classification problems, we could either achieve better results compared to available approaches or employ novel features, opening new possibilities for further development. Based on the presented research, we can state that it is possible to access the internal interlocutor state during HCI and HHI. The main contributions to the state of the art, which have been presented to the scientific community in seven conference papers and one journal paper in international, peer-reviewed media, are the following:

- For the task of *trouble recognition*, we found that the best classification results can be achieved using a set of around 100 features based on physiological signals, achieving around 90% UAF. Spatial features also provide a relatively high average performance of around 70% UAF, yet varying greatly between subjects.
- For the task of *satisfaction recognition*, we found that the spectral and prosodic features of the interlocutor's voice allow a satisfying classification performance with around 87% UAF.
- For the task of *overlap classification*, we found that the emotional changes in the overlapper's utterances enable a classification performance of around 70% UAF. We also found significant correlations between both, the overlapper's and the overlappee's levels of control and valence and the overlap class.
- For the evaluation of classification approaches, we showed how subject-independent evaluation can be performed in adverse conditions and implemented a way to calculate evaluation metrics for data with highly imbalanced class distribution.

The research presented in the previous chapters answered some of the current questions in the quickly developing field of affective computing, and also raised

new ones, offering possibilities for future work that will be discussed in the next section.

7.4 Future Work

Based on the findings of this thesis, we can identify several possible directions for future investigations. Although the speech signal does not deliver the best performance in accessing the interlocutor, it can be useful for certain applications – especially when it is the only available signal, e.g. in call centre or emergency centre scenarios. With the recent progress in end-to-end speech processing, we can expect that novel feature sets (or the abandonment of feature sets altogether in favour of processing the raw signal) will be able to improve upon the current performance. However, we can speculate that this may as well not happen, since no already available knowledge on the generating processes is included in the recognition models. Also, despite the success of deep-learning-based methods, so far they provide little insight on the modelled phenomena, lacking explanatory power and generalisation ability, as we have seen in Section 1.2.1. The physiological signals also deserve our attention, since they offer the unique possibility to "look directly" into the interlocutor. Especially taking into account the development of wearable devices obtaining at least some of the signals (such as heart rate measuring smart watches), employing physiological signals for personal and private systems might be a big benefit. The 3D signals of the interlocutor in form of spatial features deliver promising results and should be further investigated – especially the inter-individual differences need more research in order to develop a general recognition system. Also, introducing high-level features such as self-touch or specific gestures as behavioural cues for internal states might produce valuable insight. However, the interaction between humans is a multimodal process – therefore, we cannot hope to develop unimodal systems that offer the same experience that a true interaction between humans does. It seems natural to first focus on signals also processed by human interaction partners, such as speech, gestures and facial expressions. But technical systems can go beyond that, since they have access to different sensors and do not process information in the same way as humans do, for example, being able to detect the heart rate or the body temperature from video or infrared images. Therefore, it would be interesting to leave the human-centred viewpoint behind and to develop approaches relying on other kinds of information. This would enable completely different multimodal systems.

In order to develop comparable approaches and to measure the progress in the field of affective computing, the HCI researcher community needs to come to an agreement regarding the evaluation procedures and metrics, as

well as reliable benchmarks. In terms of metrics, the UAF seems to be a good choice, since it avoids the problems of using accuracy and also offers a good compromise between the recall and precision of the classification. In terms of procedures, the experimental setups should be implemented in a way that guarantees subject-independent evaluation.

The link between the emotional state of the interlocutor and her interaction behaviour should be further investigated. So far, the research focuses on recognising the internal state of the interlocutor to develop systems reacting to this state. However, the relation between the levels of valence and control and the speaking behaviour found for the domain of speech overlaps shows that there are additional factors to be discovered and considered for future development.

Connected to this is also the question of finding the ground truth: So far, we try to access the internal state of an interlocutor by evaluating her behaviour and employing external annotators to do this task manually, based on their training and experience. This is a valid approach. However, the expressed behaviour does not necessarily convey the real internal state of the interlocutor. Without reliable information on it, we cannot prove that the systems learn the truly significant aspects. Currently, we can only evaluate their performance by extensive testing in real-life scenarios. Our inability to break the barrier between the interlocutor's internal state and the external expression of this state remains one of the fundamental challenges in affective computing.

References

Afzal, S. & Robinson, P. (2011). 'Natural Affect Data: Collection and Annotation'. In: *New Perspectives on Affect and Learning Technologies*. Ed. by Calvo, R. A. & D'Mello, S. K. New York, NY: Springer New York, pp. 55–70.

Ahmed, F. & Gavrilova, M. L. (2019). 'Two-Layer Feature Selection Algorithm for Recognizing Human Emotions from 3D Motion Analysis'. In: *Proc. of Advances in Computer Graphics*. Cham: Springer International Publishing, pp. 53–67.

Alam, F.; Chowdhury, S. A.; Danieli, M. & Riccardi, G. (2016). 'How Interlocutors Coordinate with Each Other within Emotional Segments?' In: *Proc.s of COLING 2016, the 26th Int. Conf. on Computational Linguistics: Technical Papers*. Osaka: The COLING 2016 Organizing Committee, pp. 728–738.

Alarcão, S. M. & Fonseca, M. J. (2019). 'Emotions Recognition Using EEG Signals: A Survey'. *IEEE Transactions on Affective Computing* 10.3, pp. 374–393.

Alhagry, S.; Fahmy, A. A. & El-Khoribi, R. A. (2017). 'Emotion Recognition Based on EEG Using LSTM Recurrent Neural Network'. *Int. Journal of Advanced Computer Science and Applications (IJACSA)* 8.10, pp. 355–358.

Ambady, N. & Rosenthal, R. (1992). 'Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A meta-analysis.' *Psychological Bulletin* 111.2, pp. 256–274.

Andreassi, J. L. (2010). *Psychophysiology: Human behavior and physiological response*. New York: Psychology Press.

Andrés, A. M. & Tejedor, I. H. (1995). 'Is Fisher's Exact Test Very Conservative?' *Computational Statistics & Data Analysis* 19.5, pp. 579–591.

Anis, K.; Zakia, H.; Mohamed, D. & Jeffrey, C. (2018). 'Detecting Depression Severity by Interpretable Representations of Motion Dynamics'. In: *Proc. of the 13th IEEE Int. Conf. on Automatic Face Gesture Recognition (FG 2018)*, pp. 739–745.

Arel, I.; Rose, D. C. & Karnowski, T. P. (2010). 'Deep Machine Learning – a New Frontier in Artificial Intelligence Research'. *IEEE Computational Intelligence Magazine* 5.4, pp. 13–18.

Atkinson, A. P.; Dittrich, W. H.; Gemmell, A. J. & Young, A. W. (2004). 'Emotion Perception from Dynamic and Static Body Expressions in Point-Light and Full-Light Displays'. *Perception* 33.6, pp. 717–746.

Badshah, A. M.; Ahmad, J.; Rahim, N. & Baik, S. W. (2017). 'Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network'. In: *Proc. of the 2017 Int. Conf. on Platform Technology and Service (PlatCon)*, pp. 1–5.

Balomenos, T.; Raouzaiou, A.; Ioannou, S.; Drosopoulos, A.; Karpouzis, K. & Kollias, S. (2005). 'Emotion Analysis in Man-Machine Interaction Systems'. In: *Proc. of the Int. Workshop on Machine Learning for Multimodal Interaction*. Springer Berlin Heidelberg, pp. 318–328.

Bannach, D.; Amft, O. & Lukowicz, P. (2009). 'Automatic Event-Based Synchronization of Multimodal Data Streams from Wearable and Ambient Sensors'. In: *Proc. of Smart Sensing and Context*. Springer Berlin Heidelberg, pp. 135–148.

Barutchu, A.; Crewther, S. G.; Kiely, P.; Murphy, M. J. & Crewther, D. P. (2008). 'When /b/ill with /g/ill Becomes /d/ill: Evidence for a Lexical Effect in Audiovisual Speech Perception'. *European Journal of Cognitive Psychology* 20.1, pp. 1–11.

Batliner, A.; Fischer, K.; Huber, R.; Spilker, J. & Nöth, E. (2003). 'How to Find Trouble in Communication'. *Speech Communication* 40 (1-2), pp. 117–143.

Batliner, A. et al. (2011). 'Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-related User States in Speech'. *Computer Speech & Language* 25.1, pp. 4–28.

Beleites, C.; Neugebauer, U.; Bocklitz, T.; Krafft, C. & Popp, J. (2013). 'Sample Size Planning for Classification Models'. *Analytica Chimica Acta* 760, pp. 25–33.

Bergstra, J. & Bengio, Y. (2012). 'Random search for hyper-parameter optimization'. *Journal of Machine Learning Research* 13, pp. 281–305.

Bernard, S.; Heutte, L. & Adam, S. (2009). 'Influence of Hyperparameters on Random Forest Accuracy'. In: *Proc. of Multiple Classifier Systems (MCS 2009)*. Springer Berlin Heidelberg, pp. 171–180.

Bertero, D. & Fung, P. (2016). 'Deep Learning of Audio and Language Features for Humor Prediction'. In: *Proc. of the 10th Int. Conf. on Language Resources and Evaluation (LREC)*. European Language Resources Association, pp. 496–501.

Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K. & Wiswedel, B. (2007). 'KNIME: The Konstanz Information Miner'. In: *Data Analysis, Machine Learning and Applications. Proc. of the 31st Annu. Conf. of the Gesellschaft für Klassifikation e.V.* Springer Berlin Heidelberg, pp. 319–326.

Biau, G. & Scornet, E. (2016). 'A Random Forest Guided Tour'. *TEST* 25.2, pp. 197–227.

Bitouk, D.; Verma, R. & Nenkova, A. (2010). 'Class-level Spectral Features for Emotion Recognition'. *Speech Communication* 52.7, pp. 613–625.

Böck, R.; Bergmann, K. & Jaecks, P. (2014). 'Disposition Recognition from Spontaneous Speech towards a Combination with Co-speech Gestures'. In: *Int. Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*. Cham: Springer International Publishing, pp. 57–66.

Böck, R.; Egorow, O.; Siegert, I. & Wendemuth, A. (2017a). 'Comparative Study on Normalisation in Emotion Recognition from Speech'. In: *Intelligent Human Computer Interaction*. Cham: Springer International Publishing, pp. 189–201.

Böck, R.; Egorow, O. & Wendemuth, A. (2017b). 'Speaker-group Specific Acoustic Differences in Consecutive Stages of Spoken Interaction'. In: *Elektronische Sprachsignalverarbeitung 2017: Tagungsband der 28. Konferenz*, pp. 211–218.

— (2018). 'Acoustic Detection of Consecutive Stages of Spoken Interaction Based on Speaker-group Specific Features'. In: *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*, pp. 252–254.

Böck, R.; Egorow, O.; Höbel-Müller, J.; Requardt, A. F.; Siegert, I. & Wendemuth, A. (2019). 'Anticipating the User: Acoustic Disposition Recognition in Intelligent Interactions'. In: *Innovations in Big Data Mining and Em-*

bedded Knowledge. Ed. by Esposito, A.; Esposito, A. M. & Jain, L. C. Cham: Springer International Publishing, pp. 203–233.

Boser, B. E.; Guyon, I. M. & Vapnik, V. N. (1992). 'A Training Algorithm for Optimal Margin Classifiers'. In: *Proc. of the 5th Annu. Workshop on Computational Learning Theory (COLT '92)*. New York, NY, USA: ACM, pp. 144–152.

Breiman, L. (2001). 'Random Forests'. *Machine Learning* 45.1, pp. 5–32.

Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C. M.; Kazemzadeh, A.; Lee, S.; Neumann, U. & Narayanan, S. (2004). 'Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information'. In: *Proc. of the 6th Int. Conf. on Multimodal Interfaces (ICMI '04)*. New York, NY, USA: ACM, pp. 205–211.

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* 42.4, pp. 335–339.

Cacioppo, J. T.; Berntson, G. G.; Larsen, J. T.; Poehlmann, K. M. & Ito, T. A. (2000). 'The Psychophysiology of Emotion'. *Handbook of Emotions* 2, pp. 173–191.

Cai, Z.; Fan, X. & Du, J. (2017). 'Gender and Attitudes toward Technology Use: A Meta-analysis'. *Computers & Education* 105, pp. 1 –13.

Cambria, E.; Livingstone, A. & Hussain, A. (2012). 'The Hourglass of Emotions'. In: *Proc. of Cognitive Behavioural Systems*. Springer Berlin Heidelberg, pp. 144–157.

Canento, F.; Fred, A.; Silva, H.; Gamboa, H. & Lourenço, A. (2011). 'Multimodal Biosignal Sensor Data Handling for Emotion Recognition'. In: *Proc. of the IEEE SENSORS 2011*, pp. 647–650.

Castellano, G.; Villalba, S. D. & Camurri, A. (2007). 'Recognising Human Emotions from Body Movement and Gesture Dynamics'. In: *Proc. of the 2nd Int. Conf. on Affective Computing and Intelligent Interaction (ACII'07)*. Springer, pp. 71–82.

Chandaka, S.; Chatterjee, A. & Munshi, S. (2009). 'Support Vector Machines Employing Cross-correlation for Emotional Speech Recognition'. *Measurement* 42.4, pp. 611–618.

Chang, C.-C. & Lin, C.-J. (2011). 'LIBSVM: A Library for Support Vector Machines'. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2 (3), pp. 1–27.

Chang, J. & Scherer, S. (2017). 'Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks'. In: *Proc. of the 2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'17)*, pp. 2746–2750.

Chawla, N. V. (2010). 'Data Mining for Imbalanced Datasets: An Overview'. In: *Data Mining and Knowledge Discovery Handbook*. Ed. by Maimon, O. & Rokach, L. Boston, MA: Springer US, pp. 875–886.

Chen, L. S.-H. (2000). 'Joint Processing of Audio-visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction'. PhD thesis. University of Illinois at Urbana-Champaign.

Chen, L. da; Soliman, K. S.; Mao, E. & Frolick, M. N. (2000). 'Measuring User Satisfaction with Data Warehouses: an Exploratory Study'. *Information & Management* 37.3, pp. 103–110.

Chen, Y.-W. & Lin, C.-J. (2006). 'Combining SVMs with Various Feature Selection Strategies'. In: *Feature Extraction: Foundations and Applications*. Ed. by Guyon, I.; Nikravesh, M.; Gunn, S. & Zadeh, L. A. Springer Berlin Heidelberg, pp. 315–324.

Chin, J. P.; Diehl, V. A. & Norman, K. L. (1988). 'Development of an Instrument Measuring User Satisfaction of the Human-computer Interface'. In: *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI '88)*. New York, NY, USA: ACM, pp. 213–218.

Cho, J.; Lee, K.; Shin, E.; Choy, G. & Do, S. (2015). 'How Much Data is Needed to Train a Medical Image Deep Learning System to Achieve Necessary High Accuracy?' *arXiv preprint arXiv:1511.06348*.

Chowdhury, A; Danieli, M. & Riccardi, G. (2015a). 'The Role of Speakers and Context in Classifying Competition in Overlapping Speech'. In: *Proc. of the Interspeech-2015*. International Speech Communication Association, pp. 1844–1848.

Chowdhury, S. A. & Riccardi, G. (2017). 'A Deep Learning Approach to Modeling Competitiveness in Spoken Conversations'. In: *Proc. of the 2017 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'17)*, pp. 5680–5684.

Chowdhury, S. A.; Danieli, M. & Riccardi, G. (2015b). 'Annotating and Categorizing Competition in Overlap Speech'. In: *Proc. of the 2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'15)*, pp. 5316–5320.

Chowdhury, S. A.; Stepanov, E. A. & Riccardi, G. (2016). 'Predicting User Satisfaction from Turn-Taking in Spoken Conversations.' In: *Proc. of the Interspeech-2016*. International Speech Communication Association, pp. 2910–2914.

Christodoulides, G. (2017). 'A New Corpus of Collaborative Dialogue Produced Under Cognitive Load Using a Driving Simulator'. In: *Proc. of Text, Speech, and Dialogue*. Cham: Springer International Publishing, pp. 380–392.

Cohn, J. F. & Schmidt, K. (2013). 'The Timing of Facial Motion in Posed and Spontaneous Smiles'. In: *Active Media Technology*, pp. 57–69.

Corradini, A.; Mehta, M.; Bernsen, N. O.; Martin, J & Abrilian, S. (2005). 'Multimodal Input Fusion in Human-Computer Interaction'. In: *Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*. NATO Science Series Sub Series III Computer and Systems Sciences. IOS PRESS, pp. 223–234.

Cortes, C. & Vapnik, V. (1995). 'Support-Vector Networks'. *Machine learning* 20.3, pp. 273–297.

Coulson, M. (2004). 'Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence'. *Journal of Nonverbal Behavior* 28.2, pp. 117–139.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W. & Taylor, J. G. (2001). 'Emotion Recognition in Human-Computer Interaction'. *IEEE Signal Processing Magazine* 18.1, pp. 32–80.

Cowie, R.; Sussman, N. & Ben-Ze'ev, A. (2011). 'Emotion: Concepts and Definitions'. In: *Emotion-Oriented Systems: The Humaine Handbook*. Ed. by Cowie, R.; Pelachaud, C. & Petta, P. Springer Berlin Heidelberg, pp. 9–30.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Publishing.

De Gelder, B. (2009). 'Why Bodies? Twelve Reasons for Including Bodily Expressions in Affective Neuroscience'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1535, pp. 3475–3484.

De Meijer, M. (1989). 'The Contribution of General Features of Body Movement to the Attribution of Emotions'. *Journal of Nonverbal Behavior* 13.4, pp. 247–268.

Domingos, P. & Pazzani, M. (1997). 'On the Optimality of the Simple Bayesian Classifier under Zero-One Loss'. *Machine Learning* 29.2, pp. 103–130.

Douglas-Cowie, E.; Cowie, R.; Cox, C.; Amir, N. & Heylen, D. (2008). 'The Sensitive Artificial Listener: an Induction Technique for Generating Emotionally Coloured Conversation'. In: *Proc. of the LREC Workshop on Corpora for Research on Emotion and Affect*. European Language Resources Association, pp. 1–4.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York: Wiley-Interscience.

Egorow, O. & Wendemuth, A. (2016). 'Detection of Challenging Dialogue Stages Using Acoustic Signals and Biosignals'. In: *Proc. of the 24th Int. Conf. in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG'16)*, pp. 137–143.

— (2017). 'Emotional Features for Speech Overlaps Classification'. In: *Proc. of the Interspeech-2017*. International Speech Communication Association, pp. 2356–2360.

— (2019). 'On Emotions as Features for Speech Overlaps Classification'. *IEEE Transactions on Affective Computing*, s.p.

Egorow, O.; Siegert, I. & Wendemuth, A. (2017). 'Prediction of User Satisfaction in Naturalistic Human-computer Interaction'. *Kognitive Systeme* 2017.1, s.p.

— (2018). 'Improving Emotion Recognition Performance by Random-forest-based Feature Selection'. In: *Speech and Computer*. Cham: Springer International Publishing, pp. 134–144.

Egorow, O.; Mrech, T.; Weisskirchen, N. & Wendemuth, A. (2019). 'Employing Bottleneck and Convolutional Features for Speech-Based Physical Load Detection on Limited Data Amounts'. In: *Proc. of the Interspeech-2019*. International Speech Communication Association, pp. 1666–1670.

Ekman, P. (1970). 'Universal Facial Expressions of Emotion'. *California Mental Health Research Digest* 8 (4), pp. 151–158.

Ekman, P.; Levenson, R. W. & Friesen, W. V. (1983). 'Autonomic Nervous System Activity Distinguishes among Emotions'. *Science* 221.4616, pp. 1208–1210.

Etienne, C.; Fidanza, G.; Petrovskii, A.; Devillers, L. & Schmauch, B. (2018a). 'CNN+ LSTM Architecture for Speech Emotion Recognition with Data Augmentation'. *arXiv preprint arXiv:1802.05630*.

— (2018b). 'Speech Emotion Recognition with Data Augmentation and Layer-wise Learning Rate Adjustment'. *arXiv preprint arXiv:1802.05630*.

Eyben, F.; Weninger, F.; Squartini, S. & Schuller, B. (2013). 'Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies'. In: *Proc. of the 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'13)*, pp. 483–487.

Eyben, F. et al. (2015). 'The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing'. *IEEE Transactions on Affective Computing* 7.2, pp. 190–202.

Eyben, F.; Wöllmer, M. & Schuller, B. W. (2010). 'openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor'. In: *Proc. of the 18th ACM Int. Conf on Multimedia (MM '10)*, pp. 1459–1462.

Eyben, F.; Weninger, F.; Wöllmer, M. & Schuller, B. W. (2014). *Open-Source Media Interpretation by Large Feature-Space Extraction*. Tech. rep. Version 2.1. Audeering UG.

Fan, Y.; Lu, X.; Li, D. & Liu, Y. (2016). 'Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks'. In: *Proceedings of the 18th ACM Int. Conf. on Multimodal Interaction (ICMI'16)*, pp. 445–450.

Feild, H. A.; Allan, J. & Jones, R. (2010). 'Predicting Searcher Frustration'. In: *Proc. of the 33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'10)*, pp. 34–41.

Ferchow, S.; Haase, M.; Krüger, J.; Vogel, M.; Wahl, M. & Frommer, J. (2016). 'Speech Matters – Psychological Aspects of Artificial versus Anthropomorphic System Voices in User-Companion Interaction'. In: *Proc. of the 5th Int. Conf. of Design, User Experience, and Usability (DUXU 2016)*, pp. 319–327.

Figueroa, R. L.; Zeng-Treitler, Q.; Kandula, S. & Ngo, L. H. (2012). 'Predicting Sample Size Required for Classification Performance'. *BMC Medical Informatics and Decision Making* 12.1, p. 8.

Fisher, R. A. (1956). 'Mathematics of a Lady Tasting Tea'. In: *The World of Mathematics*. Vol. 3. Part VIII. Simon and Schuster New York, pp. 1514–1521.

Fontaine, J. R.; Scherer, K. R.; Roesch, E. B. & Ellsworth, P. C. (2007). 'The World of Emotions is not Two-Dimensional'. *Psychological Science* 18.12, pp. 1050–1057.

Fox, S.; Karnawat, K.; Mydland, M.; Dumais, S. & White, T. (2005). 'Evaluating Implicit Measures to Improve Web Search'. *ACM Transactions on Information Systems* 23.2, pp. 147–168.

Frank, E.; Hall, M. & Witten, I. (2016). *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Freitas, J.; Teixeira, A. & Dias, M. (2014). 'Multimodal Corpora for Silent Speech Interaction'. In: *Proc. of the 9th Int. Conf. on Language Resources and Evaluation (LREC'14)*. European Language Resources Association.

French, P. & Local, J. (1983). 'Turn-competitive Incomings'. *Journal of Pragmatics* 7.1, pp. 17–38.

Friedman, N.; Geiger, D. & Goldszmidt, M. (1997). 'Bayesian Network Classifiers'. *Machine Learning* 29.2-3, pp. 131–163.

Frommer, J.; Rösner, D.; Haase, M.; Lange, J.; Friesen, R. & Otto, M. (2012a). *Detection and Avoidance of Failures in Dialogues – Wizard of Oz Experiment Operator's Manual*. Pabst Science Publishers.

Frommer, J. et al. (2012b). 'Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus'. In: *Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC 2012)*. European Language Resources Association, pp. 3064–3069.

Gehle, R.; Pitsch, K. & Wrede, S. (2014). 'Signaling Trouble in Robot-to-group Interaction. Emerging Visitor Dynamics with a Museum Guide Robot'. In: *Proc. of the 2nd Int. Conf. on Human-Agent Interaction (HAI'14)*. New York, NY, USA: ACM, pp. 361–368.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M. & Kagal, L. (2018). 'Explaining Explanations: An Overview of Interpretability of Machine Learning'. In: *Proc. of the 2018 IEEE 5th Int. Conf. on Data Science and Advanced Analytics (DSAA)*, pp. 80–89.

Goldberg, J. A. (1990). 'Interrupting the Discourse on Interruptions: An Analysis in Terms of Relationally Neutral, Power- and Rapport-oriented Acts'. *Journal of Pragmatics* 14.6, pp. 883–903.

Grezes, F.; Richards, J. & Rosenberg, A. (2013). 'Let Me Finish: Automatic Conflict Detection Using Speaker Overlap'. In: *Proc. of the Interspeech-2013*. International Speech Communication Association, pp. 200–204.

Grimm, M.; Kroschel, K. & Narayanan, S. (2007). 'Support Vector Regression for Automatic Recognition of Spontaneous Emotions in Speech'. In: *Proc. of the 2007 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*, pp. IV–1085–IV–1088.

Gunes, H. & Piccardi, M. (2007). 'Bi-modal Emotion Recognition from Expressive Face and Body Gestures'. *Journal of Network and Computer Applications* 30.4. Special Issue on Information Technology, pp. 1334–1345.

Haag, A.; Goronzy, S.; Schaich, P. & Williams, J. (2004). 'Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System'. In: *Proc. of Tutorial and Research Workshop on Affective Dialogue Systems*. Springer Berlin Heidelberg, pp. 36–48.

Hager, J. C. & Ekman, P. (1985). 'The Asymmetry of Facial Actions is Inconsistent with Models of Hemispheric Specialization'. *Psychophysiology* 22.3, pp. 307–318.

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.

Hara, S.; Kitaoka, N. & Takeda, K. (2010). 'Estimation Method of User Satisfaction Using N-gram-based Dialog History Model for Spoken Dialog System'. In: *Proc. of the 7th Int. Conf. on Language Resources and Evaluation (LREC'10)*. European Language Resource Association, pp. 78–83.

Hassan, A. & White, R. W. (2013). 'Personalized Models of Search Satisfaction'. In: *Proc. of the 22nd ACM Int. Conf. on Information & Knowledge Management*, pp. 2009–2018.

Hayes, A. F. & Krippendorff, K. (2007). 'Answering the Call for a Standard Reliability Measure for Coding Data'. *Communication Methods and Measures* 1.1, pp. 77–89.

Healey, J.; Picard, R. & Vyzas, E. (2001). 'Toward Machine Emotional Intelligence: Analysis of Affective Physiological State'. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 23.10, pp. 1175–1191.

Heracleous, P.; Mohammad, Y. & Yoneyama, A. (2019). 'Deep Convolutional Neural Networks for Feature Extraction in Speech Emotion Recognition'. In: *Proc. of 2019 Int. Conf. on Human-Computer Interaction*. Springer, pp. 117–132.

Hirschberg, J. (2002). 'Communication and Prosody: Functional Aspects of Prosody'. *Speech Communication* 36.1, pp. 31–43.

Hirschberg, J.; Litman, D. & Swerts, M. (1999). 'Prosodic Cues to Recognition Errors'. In: *Proc. of the Automatic Speech Recognition & Understanding Workshop (ASRU'99)*, pp. 349–352.

Ho, T. K. (1995). 'Random Decision Forests'. In: *Proc. of 3rd Int. Conf. on Document Analysis and Recognition*, pp. 278–282.

Hossin, M. & Sulaiman, M. (2015). 'A Review on Evaluation Metrics for Data Classification Evaluations'. *Int. Journal of Data Mining & Knowledge Management Process* 5.2, pp. 1–11.

ISO9241-11:1998 (1998). *Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) — Part 11 Guidance on Usability*. Standard. International Organization for Standardization.

ISO9241-11:2015 (2015). *Ergonomics of Human-system Interaction — Part 11: Usability: Definitions and Concepts*. Standard. International Organization for Standardization.

ISO9241-11:2018 (2018). *Ergonomics of Human-system Interaction — Part 11: Usability: Definitions and Concepts*. Standard. International Organization for Standardization.

Ives, B.; Olson, M. & Baroudi, J. J. (1983). 'The Measurement of User Information Satisfaction'. *Communications of the ACM* 26.10, pp. 785–793.

Jaimes, A. & Sebe, N. (2007). 'Multimodal Human–Computer Interaction: A Survey'. *Computer Vision and Image Understanding* 108.1. Special Issue on Vision for Human-Computer Interaction, pp. 116–134.

Jatupaiboon, N.; Pan-Ngum, S. & Israsena, P. (2015). 'Subject-Dependent and Subject-Independent Emotion Classification Using Unimodal and Multimodal Physiological Signals'. *Journal of Medical Imaging and Health Informatics* 5.5, pp. 1020–1027.

Jefferson, G. (1989). 'Preliminary Notes on a Possible Metric Which Provides for a Standard Maximum Silence of Approximately One Second in Conversation'. *Conversation: An Interdisciplinary Perspective* 3, pp. 166–196.

Jing, S.; Mao, X. & Chen, L. (2018). 'Prominence Features: Effective Emotional Features for Speech Emotion Recognition'. *Digital Signal Processing* 72, pp. 216–231.

Joachims, T.; Granka, L.; Pan, B.; Hembrooke, H. & Gay, G. (2005). 'Accurately Interpreting Clickthrough Data As Implicit Feedback'. In: *Proc. of the 28th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'05)*, pp. 154–161.

Johnson, D. W. (1975). 'Cooperativeness and Social Perspective Taking'. *Journal of Personality and Social Psychology* 31.2, pp. 241–244.

Karg, M.; Kuhnlenz, K. & Buss, M. (2010). 'Recognition of Affect Based on Gait Patterns'. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 40.4, pp. 1050–1061.

Katsis, C. D.; Katertsidis, N.; Ganiatsas, G. & Fotiadis, D. I. (2008). 'Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach'. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38.3, pp. 502–512.

Kaza, K.; Psaltis, A.; Stefanidis, K.; Apostolakis, K. C.; Thermos, S.; Dimitropoulos, K. & Daras, P. (2016). 'Body Motion Analysis for Emotion Recognition in Serious Games'. In: *Proc. of the 2016 Int. Conf. on Universal Access in Human-Computer Interaction. Interaction Techniques and Environments*. Cham: Springer International Publishing, pp. 33–42.

Kelly, D. (2009). 'Methods for Evaluating Interactive Information Retrieval Systems with Users'. *Foundations and Trends in Information Retrieval* 3.1–2, pp. 1–224.

Këpuska, V. & Bohouta, G. (2017). 'Comparing speech recognition systems (Microsoft API, Google API and CMU Sphinx)'. *Int. Journal of Engineering Research and Application* 7.3, pp. 20–24.

Kim, J. (2007). 'Bimodal Emotion Recognition using Speech and Physiological Changes'. In: *Robust Speech Recognition and Understanding*. Ed. by Grimm, M. & Kroschel, K. Vienna, Austria: I-Tech Education and Publishing, pp. 265–280.

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 Interspeech-2017*. International Speech Communication Association, pp. 1113–1117.

Kim, K. H.; Bang, S. W. & Kim, S. R. (2004). 'Emotion Recognition System Using Short-term Monitoring of Physiological Signals'. *Medical and Biological Engineering and Computing* 42.3, pp. 419–427.

Kim, Y.; Lee, H. & Provost, E. M. (2013). 'Deep Learning for Robust Feature Generation in Audiovisual Emotion Recognition'. In: *Proc. of the 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'13)*, pp. 3687–3691.

Kim, Y.; Hassan, A.; White, R. W. & Zitouni, I. (2014). 'Modeling Dwell Time to Predict Click-level Satisfaction'. In: *Proc. of the 7th ACM Int. Conf. on Web Search and Data Mining (WSDM'14)*. ACM, pp. 193–202.

Kipp, M. & Martin, J.-C. (2009). 'Gesture and Emotion: Can Basic Gestural Form Features Discriminate Emotions?' In: *Proc. of the 3rd Int. Conf. on Affective Computing and Intelligent Interaction (ACII'09)*. IEEE, pp. 1–8.

Kirakowski, J.; Claridge, N. & Whitehand, R. (1998). 'Human Centered Measures of Success in Web Site Design'. In: *Proc. of the 4th Conf. on Human Factors & the Web*, s.p.

Kiseleva, J.; Williams, K.; Jiang, J.; Awadallah, A.; Zitouni, I.; Crook, A. & Anastasakos, T. (2016a). 'Predicting User Satisfaction with Intelligent Assistants'. In: *Proc. of the 39th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'16*, pp. 495–505.

Kiseleva, J.; Williams, K.; Jiang, J.; Hassan Awadallah, A.; Crook, A. C.; Zitouni, I. & Anastasakos, T. (2016b). 'Understanding User Satisfaction with Intelligent Assistants'. In: *Proc. of the 2016 ACM Conf. on Human Information Interaction and Retrieval (CHIIR'16)*. New York, NY, USA: ACM, pp. 121–130.

Kishore, K. V. K. & Satish, P. K. (2013). 'Emotion Recognition in Speech Using MFCC and Wavelet Features'. In: *Proc. of the 3rd IEEE Int. Advance Computing Conf. (IACC'13)*, pp. 842–847.

Kleinsmith, A. & Bianchi-Berthouze, N. (2013). 'Affective Body Expression Perception and Recognition: A Survey'. *IEEE Transactions on Affective Computing* 4.1, pp. 15–33.

Kopp, S. et al. (2018). 'Conversational Assistants for Elderly Users – The Importance of Socially Cooperative Dialogue'. In: *Proc. of the AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications*, pp. 10–17.

Kothe, C. A.; Makeig, S. & Onton, J. A. (2013). 'Emotion Recognition from EEG during Self-Paced Emotional Imagery'. In: *Proc. of the 5th Int. Conf. on Affective Computing and Intelligent Interaction (ACII'13)*, pp. 855–858.

Krahmer, E.; Swerts, M.; Theune, M. & Weegels, M. (2002). 'The Dual of Denial: Two Uses of Disconfirmations in Dialogue and Their Prosodic Correlates'. *Speech Communication* 36.1, pp. 133–145.

Kurpukdee, N.; Koriyama, T.; Kobayashi, T.; Kasuriya, S.; Wutiwiwatchai, C. & Lamsrichan, P. (2017). 'Speech Emotion Recognition Using Convolutional Long Short-term Memory Neural Network and Support Vector Machines'. In: *Proc. of the 2017 Asia-Pacific Signal and Information Processing Association Annu. Summit and Conf. (APSIPA ASC)*, pp. 1744–1749.

Kurtić, E.; Brown, G. J. & Wells, B. (2013). 'Resources for Turn Competition in Overlapping Talk'. *Speech Communication* 55.5, pp. 721–743.

Kwon, O.-W.; Chan, K.; Hao, J. & Lee, T.-W. (2003). 'Emotion Recognition by Speech Signals'. In: *Proc. of the 8th European Conf. on Speech Communication and Technology (EUROSPEECH-2003)*, pp. 125–128.

Lagun, D.; Hsieh, C.-H.; Webster, D. & Navalpakkam, V. (2014). 'Towards Better Measurement of Attention and Satisfaction in Mobile Search'. In:

Proc. of the 37th Int ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'14). ACM, pp. 113–122.

LeBreton, J. M. & Senter, J. L. (2008). 'Answers to 20 Questions About Interrater Reliability and Interrater Agreement'. *Organizational Research Methods* 11.4, pp. 815–852.

Lee, C.-C. & Narayanan, S. (2010). 'Predicting Interruptions in Dyadic Spoken Interactions'. In: *Proc. of the 2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'10)*, pp. 5250–5253.

Lee, C.-C.; Lee, S. & Narayanan, S. S. (2008). 'An Analysis of Multimodal Cues of Interruption in Dyadic Spoken Interactions'. In: *Proc. of the Interspeech-2008*. International Speech Communication Association, pp. 1678–1681.

Lee, J. & Tashev, I. (2015). 'High-level Feature Representation Using Recurrent Neural Network for Speech Emotion Recognition'. In: *Proc. of the Interspeech-2015*. International Speech Communication Association.

Lee Rodgers, J. & Nicewander, W. A. (1988). 'Thirteen Ways to Look at the Correlation Coefficient'. *The American Statistician* 42.1, pp. 59–66.

Lefter, I. & Jonker, C. M. (2017). 'Aggression Recognition Using Overlapping Speech'. In: *Proc. of the 7th Int. Conf. on Affective Computing and Intelligent Interaction (ACII'17)*, pp. 299–304.

Leonard, R. (1984). 'A Database for Speaker-independent Digit Recognition'. In: *Proc. of the 1984 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'84)*, pp. 328–331.

Levenson, R. W.; Ekman, P. & Friesen, W. V. (1990). 'Voluntary Facial Action Generates Emotion-Specific Autonomic Nervous System Activity'. *Psychophysiology* 27.4, pp. 363–384.

Li, H. Z. (2001). 'Cooperative and Intrusive Interruptions in Inter- and Intracultural Dyadic Discourse'. *Journal of Language and Social Psychology* 20.3, pp. 259–284.

Li, T.; Ogihara, M. & Li, Q. (2003). 'A Comparative Study on Content-Based Music Genre Classification'. In: *Proc. of the 26th Annu. Int. ACM SIGIR Conf. on Research and Development in Informaion Retrieval SIGIR '03*, 282—289.

Li, Y.; Chao, L.; Liu, Y.; Bao, W. & Tao, J. (2015). 'From Simulated Speech to Natural Speech, What Are the Robust Features for Emotion Recognition?' In: *Proc. of the 6th Int. Conf. on Affective Computing and Intelligent Interaction (ACII'15)*, pp. 368–373.

Lim, W.; Jang, D. & Lee, T. (2016). 'Speech Emotion Recognition Using Convolutional and Recurrent Neural Networks'. In: *Proc. of the 2016 Asia-Pacific Signal and Information Processing Association Annu. Summit and Conf. (APSIPA ASC)*, pp. 1–4.

Lindgaard, G. (1999). 'Does Emotional Appeal Determine Perceived Usability of Web Sites'. In: *Proc. of the 2nd Int. Cyberspace Conf. on Ergonomics (CybErg)*, pp. 202–211.

Lindgaard, G. & Dudek, C. (2003). 'What Is This Evasive Beast We Call User Satisfaction?' *Interacting with computers* 15.3, pp. 429–452.

Liu, W. Z.; White, A. P.; Thompson, S. G. & Bramer, M. A. (1997). 'Techniques for Dealing with Missing Values in Classification'. In: *Proc. of the Int. Symp. on Intelligent Data Analysis*. Springer Berlin Heidelberg, pp. 527–536.

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.

Logeswaran, N. & Bhattacharya, J. (2009). 'Crossmodal Transfer of Emotion by Music'. *Neuroscience Letters* 455.2, pp. 129 –133.

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

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: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2018*. TUDpress, Dresden, pp. 208–215.

Lozano-Monasor, E.; López, M. T.; Fernández-Caballero, A. & Vigo-Bustos, F. (2014). 'Facial Expression Recognition from Webcam Based on Active Shape Models and Support Vector Machines'. In: *Proc. of the Int. Workshop on Ambient Assisted Living (IWAAL'14)*. Ed. by Pecchia, L.; Chen,

L. L.; Nugent, C. & Bravo, J. Cham: Springer International Publishing, pp. 147–154.

Luengo, I.; Navas, E.; Hernaez, I. & Sanchez, J. (2005). 'Automatic Emotion Recognition Using Prosodic Parameters'. In: *Proc. of the Interspeech-2005*. International Speech Communication Association, pp. 493–496.

Makri-Tsilipakou, M. (1994). 'Interruption Revisited: Affiliative vs. Disaffiliative Intervention'. *Journal of Pragmatics* 21.4, pp. 401–426.

Mao, Q.; Dong, M.; Huang, Z. & Zhan, Y. (2014). 'Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks'. *IEEE Transactions on Multimedia* 16.8, pp. 2203–2213.

Marchi, E.; Tonelli, D.; Xu, X.; Ringeval, F.; Deng, J.; Squartini, S. & Schuller, B. (2016). 'Pairwise Decomposition with Deep Neural Networks and Multiscale Kernel Subspace Learning for Acoustic Scene Classification'. In: *Proc. of the 2016 Workshop on the Detection and Classification of Acoustic Scenes and Events (DCASE2016)*, pp. 543–547.

Marelli, L. & Testa, G. (2018). 'Scrutinizing the EU General Data Protection Regulation'. *Science* 360.6388, pp. 496–498.

McGurk, H. & MacDonald, J. (1976). 'Hearing Lips and Seeing Voices'. *Nature* 264.5588, 746—748.

Mehrabian, A. (1996). 'Pleasure-arousal-dominance: A General Framework for Describing and Measuring Individual Differences in Temperament'. *Current Psychology* 14.4, pp. 261–292.

Mehrotra, R.; Awadallah, A. H.; Shokouhi, M.; Yilmaz, E.; Zitouni, I.; El Kholy, A. & Khabsa, M. (2017). 'Deep Sequential Models for Task Satisfaction Prediction'. In: *Proc. of the 2017 ACM Conf. on Information and Knowledge Management (CIKM '17)*, pp. 737–746.

Metallinou, A. & Narayanan, S. (2013). 'Annotation and Processing of Continuous Emotional Attributes: Challenges and Opportunities'. In: *Proc. of the 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG 2013)*, pp. 1–8.

Michel, P. & El Kaliouby, R. (2003). 'Real Time Facial Expression Recognition in Video Using Support Vector Machines'. In: *Proc. of the 5th Int. Conf. on Multimodal Interfaces (ICMI'03)*, pp. 258–264.

Mitchell, T. (1997). *Machine Learning*. English. Frankfurt/Main: McGraw-Hill.

Montesdioca, G. P. Z. & Macada, A. C. G. (2015). 'Measuring User Satisfaction with Information Security Practices'. *Computers & Security* 48, pp. 267–280.

Muñoz, J. E.; Gouveia, E. R.; Cameirão, M. S. & Badia, S. B. i. (2018). 'PhysioLab – a Multivariate Physiological Computing Toolbox for ECG, EMG and EDA Signals: a Case of Study of Cardiorespiratory Fitness Assessment in the Elderly Population'. *Multimedia Tools and Applications* 77.9, pp. 11521–11546.

Murata, K. (1994). 'Intrusive or Co-operative? A Cross-cultural Study of Interruption'. *Journal of Pragmatics* 21.4, pp. 385–400.

Naji, M.; Firoozabadi, M. & Azadfallah, P. (2015). 'Emotion Classification during Music Listening from Forehead Biosignals'. *Signal, Image and Video Processing* 9.6, pp. 1365–1375.

Nass, C.; Steuer, J. & Tauber, E. R. (1994). 'Computers Are Social Actors'. In: *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI'94)*. ACM, pp. 72–78.

Noroozi, F.; Corneanu, C. A.; Kamińska, D.; Sapiński, T.; Escalera, S. & Anbarjafari, G. (2018). 'Survey on Emotional Body Gesture Recognition'. *arXiv preprint arXiv:1801.07481*.

Nowak, S. & Rüger, S. (2010). 'How Reliable Are Annotations via Crowdsourcing: A Study About Inter-annotator Agreement for Multi-label Image Annotation'. In: *Proc. of the Int. Conf. on Multimedia Information Retrieval (MIR '10)*, pp. 557–566.

Oertel, C.; Wlodarczak, M.; Tarasov, A.; Campbell, N. & Wagner, P. (2012). 'Context Cues for Classification of Competitive and Collaborative Overlaps'. In: *Proc. of Speech Prosody*, pp. 721–724.

Ogara, S. O.; Koh, C. E. & Prybutok, V. R. (2014). 'Investigating Factors Affecting Social Presence and User Satisfaction with Mobile Instant Messaging'. *Computers in Human Behavior* 36, pp. 453–459.

Oshrat, Y.; Bloch, A.; Lerner, A.; Cohen, A.; Avigal, M. & Zeilig, G. (2016). 'Speech Prosody as a Biosignal for Physical Pain Detection'. In: *Proc. of Speech Prosody*, pp. 420–424.

Park, S.; Shoemark, P. & Morency, L.-P. (2014). 'Toward Crowdsourcing Micro-level Behavior Annotations: The Challenges of Interface, Training, and Generalization'. In: *Proc. of the 19th Int. Conf. on Intelligent User Interfaces (IUI'14)*, pp. 37–46.

Pedregosa, F. et al. (2011). 'Scikit-learn: Machine Learning in Python'. *Journal of Machine Learning Research* 12, pp. 2825–2830.

Picard, R. W. (1997). *Affective Computing*. Tech. rep. Cambridge, Massachsusetts: M.I.T. Media Laboratory Perceptual Computing Section.

Picard, R. W.; Vyzas, E. & Healey, J. (2001). 'Toward Machine Emotional Intelligence: Analysis of Affective Physiological State'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.10, pp. 1175–1191.

Plutchik, R. (2001). 'The Nature of Emotions'. *American Scientist* 89 (4), pp. 344–350.

Poria, S.; Cambria, E.; Bajpai, R. & Hussain, A. (2017). 'A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion'. *Information Fusion* 37, pp. 98–125.

Prylipko, D.; Rösner, D.; Siegert, I.; Günther, S.; Friesen, R.; Haase, M.; Vlasenko, B. & Wendemuth, A. (2014). 'Analysis of Significant Dialog Events in Realistic Human-Computer Interaction'. *Journal on Multimodal User Interfaces* 8.1, pp. 75–86.

Rainville, P.; Bechara, A.; Naqvi, N. & Damasio, A. R. (2006). 'Basic Emotions are Associated with Distinct Patterns of Cardiorespiratory Activity'. *Int. Journal of Psychophysiology* 61.1, pp. 5–18.

Ramanarayanan, V.; Lange, P.; Evanini, K.; Molloy, H.; Tsuprun, E.; Qian, Y. & Suendermann-Oeft, D. (2017). 'Using Vision and Speech Features for Automated Prediction of Performance Metrics in Multimodal Dialogs'. *ETS Research Report Series* 2017.1, pp. 1–11.

Rapczynski, M.; Werner, P. & Al-Hamadi, A. (2019). 'Effects of Video Encoding on Camera Based Heart Rate Estimation'. *IEEE Transactions on Biomedical Engineering* 66.12, pp. 3360–3370.

Ren, S.; Cao, X.; Wei, Y. & Sun, J. (2015). 'Global Refinement of Random Forest'. In: *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 723–730.

Requardt, A. F.; Ihme, K.; Wilbrink, M. & Wendemuth, A. (2019). 'Towards Affect-Aware Vehicles for Increasing Safety and Comfort: Recognizing Driver Emotions from Audio Recordings in a Realistic Driving Study'. *IET Research Journals*. Submitted.

Ringeval, F.; Amiriparian, S.; Eyben, F.; Scherer, K. & Schuller, B. W. (2014). 'Emotion Recognition in the Wild: Incorporating Voice and Lip Activity in Multimodal Decision-Level Fusion'. In: *Proc. of the 16th Int. Conf. on Multimodal Interaction (ICMI'14)*. ACM, pp. 473–480.

Roberts, F.; Francis, A. L. & Morgan, M. (2006). 'The Interaction of Inter-turn Silence with Prosodic Cues in Listener Perceptions of "Trouble" in Conversation'. *Speech Communication* 48.9, pp. 1079–1093.

Rokach, L. & Maimon, O. (2005). 'Top-Down Induction of Decision Trees Classifiers – a Survey'. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35.4, pp. 476–487.

Rong, J.; Li, G. & Chen, Y.-P. P. (2009). 'Acoustic Feature Selection for Automatic Emotion Recognition from Speech'. *Information Processing & Management* 45.3, pp. 315–328.

Rosenthal-von der Pütten, A. M.; Straßmann, C.; Yaghoubzadeh, R.; Kopp, S. & Krämer, N. C. (2019). 'Dominant and Submissive Nonverbal Behavior of Virtual Agents and Its Effects on Evaluation and Negotiation Outcome in Different Age Groups'. *Computers in Human Behavior* 90, pp. 397–409.

Rösner, D.; Frommer, J.; Andrich, R.; Friesen, R.; Haase, M.; Kunze, M.; Lange, J. & Otto, M. (2012a). 'LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing'. In: *Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC'12)*. European Language Resources Association, pp. 82–89.

Rösner, D. F.; Frommer, J.; Friesen, R.; Haase, M.; Lange, J. & Otto, M. (2012b). 'LAST MINUTE: A Multimodal Corpus of Speech-based User-Companion Interactions'. In: *Proc. of the 8th Int. Conf. on Language Resources and Evaluation (LREC'12)*. European Language Resources Association, pp. 2559–2566.

Rösner, D. F.; Hazer-Rau, D.; Kohrs, C.; Bauer, T.; Günther, S.; Hoffmann, H.; Zhang, L. & Brechmann, A. (2016). 'Is There a Biological Basis for Success in Human Companion Interaction? – Results from a Transsituational Study'. In: *Proc. of the 18th Int. Conf. on Human-Computer Interaction (HCII'16)*. Springer, pp. 77–88.

Sacks, H.; Schegloff, E. A. & Jefferson, G. (1974). 'A Simplest Systematics for the Organization of Turn-taking for Conversation'. *Language*, pp. 696–735.

Saha, S.; Datta, S.; Konar, A. & Janarthanan, R. (2014). 'A Study on Emotion Recognition from Body Gestures Using Kinect Sensor'. In: *Proc. of the 2014 Int. Conf. on Communication and Signal Processing*, pp. 56–60.

Sammut, C. & Webb, G. I. (eds.). *Encyclopedia of Machine Learning*. Boston, MA: Springer US.

Sato, N. & Obuchi, Y. (2007). 'Emotion Recognition Using Mel-frequency Cepstral Coefficients'. *Information and Media Technologies* 2.3, pp. 835–848.

Savva, N.; Scarinzi, A. & Bianchi-Berthouze, N. (2012). 'Continuous Recognition of Player's Affective Body Expression as Dynamic Quality of Aesthetic Experience'. *IEEE Transactions on Computational Intelligence and AI in Games* 4.3, pp. 199–212.

Schegloff, E. A. (2000). 'Overlapping Talk and the Organization of Turn-taking for Conversation'. *Language in Society* 29.1, pp. 1–63.

Scherer, K. R. (2003). 'Vocal Communication of Emotion: A Review of Research Paradigms'. *Speech Communication* 40.1, pp. 227–256.

Scherer, K. R. (2005). 'What Are Emotions? And How Can They Be Measured?' *Social Science Information* 44.4, pp. 695–729.

Schmitt, M.; Ringeval, F. & Schuller, B. W. (2016). 'At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech.' In: *Proc. of the Interspeech-2016*. International Speech Communication Association, pp. 495–499.

Schuller, B.; Rigoll, G. & Lang, M. (2003). 'Hidden Markov Model-based Speech Emotion Recognition'. In: *Proc. of the 2003 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'03)*. Vol. 2, pp. II–1–II–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 Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'07)*. Vol. 4, pp. IV–941–IV–944.

Schuller, B. W.; 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'09)*. Merano, Italy, pp. 552–557.

Schulz von Thun, F. (2013). *Miteinander reden 1: Störungen und Klärungen: Allgemeine Psychologie der Kommunikation*. Rowohlt Verlag GmbH.

Selting, M. (1996). 'Prosody as an Activity-Type Distinctive Cue in Conversation: The case of So-Called 'Astonished' Questions in Repair Initiation'. *Prosody in Conversation: Interactional Studies* 12, pp. 231–270.

Senecal, S.; Cuel, L.; Aristidou, A. & Magnenat-Thalmann, N. (2016). 'Continuous Body Emotion Recognition System During Theater Performances'. *Computer Animation and Virtual Worlds* 27.3–4, pp. 311–320.

Shimojima, A.; Katagiri, Y.; Koiso, H. & Swerts, M. (2002). 'Informational and Dialogue-coordinating Functions of Prosodic Features of Japanese Echoic Responses'. *Speech Communication* 36.1, pp. 113–132.

Shin, D.; Shin, D. & Shin, D. (2017). 'Development of Emotion Recognition Interface Using Complex EEG/ECG Bio-signal for Interactive Contents'. *Multimedia Tools and Applications* 76.9, pp. 11449–11470.

Shriberg, E.; Wade, E. & Price, P. (1992). 'Human-machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction'. In: *Proc. of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, pp. 49–54.

Shriberg, E.; Stolcke, A. & Baron, D. (2001). 'Can Prosody Aid the Automatic Processing of Multi-Party Meetings? Evidence from Predicting Punctuation, Disfuencies, and Overlapping Speech'. In: *Proc. of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*. International Speech Communication Association.

Siegert, I. & Ohnemus, K. (2015). 'A New Dataset of Telephone-based Human-human Call-center Interaction with Emotional Evaluation'. In: *Proc. of the 1st Int. Symposium on Companion Technology (ISCT'15)*, pp. 143–148.

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* 8.1, pp. 17–28.

Siegert, I.; Philippou-Hübner, D.; Tornow, M.; Heinemann, R.; Wendemuth, A.; Ohnemus, K.; Fischer, S. & Schreiber, G. (2015a). 'Ein Datenset zur Untersuchung emotionaler Sprache in Kundenbindungsdialogen'. In: *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2015*. TUDpress, Dresden, pp. 180–187.

Siegert, I.; Böck, R.; Wendemuth, A.; Vlasenko, B. & Ohnemus, K. (2015b). 'Overlapping Speech, Utterance Duration and Fffective Content in HHI and HCI – An Comparison'. In: *Proc. of the 6th IEEE Int. Conf. on Cognitive Infocommunications (CogInfoCom 2015)*, pp. 83–88.

Siegert, I.; Tang, S. & Lotz, A. F. (2018). 'Acoustic Addressee-Detection — Analysing the Impact of Age, Gender and Technical Knowledge'. In: *Elektronische Sprachsignalverarbeitung 2018: Tagungsband der 29. Konferenz*. Ulm, Germany, pp. 113–120.

Silipo, R.; Adae, I.; Hart, A. & Berthold, M. (2014). *Seven Techniques for Dimensionality Reduction*. Tech. rep. KNIME.

Sivrikaya, F. & Yener, B. (2004). 'Time Synchronization in Sensor Networks: a Survey'. *IEEE Network* 18.4, pp. 45–50.

Slavich, G. M.; Taylor, S. & Picard, R. W. (2019). 'Stress Measurement Using Speech: Recent Advancements, Validation Issues, and Ethical and Privacy Considerations'. *Stress* 22.4, pp. 408–413.

Sloman, A. (1987). 'Motives, Mechanisms, and Emotions'. *Cognition and Emotion* 1.3, pp. 217–233.

Soleymani, M.; Villaro-Dixon, F.; Pun, T. & Chanel, G. (2017). 'Toolbox for Emotional feAture extraction from Physiological signals (TEAP)'. *Frontiers in ICT* 4, p. 1.

Song, P. & Zheng, W. (2018). 'Feature Selection Based Transfer Subspace Learning for Speech Emotion Recognition'. *IEEE Transactions on Affective Computing*, s.p.

Straßmann, C.; Rosenthal-von der Pütten, A. M.; Yaghoubzadeh, R.; Kaminski, R. & Krämer, N. (2016). 'The Effect of an Intelligent Virtual Agent's Nonverbal Behavior with Regard to Dominance and Cooperativity'. In: *Proc. of the Int. Conf. on Intelligent Virtual Agents*. Springer International Publishing, pp. 15–28.

Suchman, L. A. (1987). *Plans and Situated Actions: The Problem of Human-machine Communication.* Cambridge University Press.

Sundararaman, B.; Buy, U. & Kshemkalyani, A. D. (2005). 'Clock Synchronization for Wireless Sensor Networks: A Survey'. *Ad Hoc Networks* 3.3, pp. 281–323.

Takahashi, K. (2004). 'Remarks on Emotion Recognition from Bio-potential Signals'. In: *Proc. of the 2nd Int. Conf. on Autonomous Robots and Agents*, pp. 186–191.

Tao, J. & Tan, T. (2005). 'Affective Computing: A Review'. In: *Proc. of the 1st Int. Conf. on Affective Computing and Intelligent Interaction (ACII'05)*. Springer Berlin Heidelberg, pp. 981–995.

Thiam, P. et al. (2019). 'Multi-modal Pain Intensity Recognition based on the SenseEmotion Database'. *IEEE Transactions on Affective Computing*, s.p.

Ting-Toomey, S. & Dorjee, T. (2018). *Communicating across Cultures.* Guilford Publications.

Torralba, A. & Efros, A. A. (2011). 'Unbiased Look at Dataset Bias'. In: *Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'11)*, pp. 1521–1528.

Toyama, S.; Saito, D. & Minematsu, N. (2017). 'Use of Global and Acoustic Features Associated with Contextual Factors to Adapt Language Models for Spontaneous Speech Recognition'. In: *Proc. of the Interspeech'17*. International Speech Communication Association, pp. 543–547.

Trabelsi, I. & Bouhlel, M. S. (2015). 'Feature Selection for GUMI Kernel-based SVM in Speech Emotion Recognition'. *International Journal of Synthetic Emotions (IJSE)* 6.2, pp. 57–68.

Tracy, J. L. & Matsumoto, D. (2008). 'The Spontaneous Expression of Pride and Shame: Evidence for Biologically Innate Nonverbal Displays'. *Proc. of the National Academy of Sciences* 105.33, pp. 11655–11660.

Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M. A.; Schuller, B. & Zafeiriou, S. (2016). 'Adieu Features? End-to-end Speech Emotion Recognition Using a Deep Convolutional Recurrent Network'. In: *Proc. of the 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'16)*, pp. 5200–5204.

Truong, A.; Boujut, H. & Zaharia, T. (2016). 'Laban Descriptors for Gesture Recognition and Emotional Analysis'. *The Visual Computer* 32.1, pp. 83–98.

Truong, K. P. (2013). 'Classification of Cooperative and Competitive Overlaps in Speech Using Cues from the Context, Overlapper, and Overlappee'. In: *Proc. of the Interspeech-2013*. International Speech Communication Association, pp. 1404–1408.

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.

Valli, A. (2008). 'The Design of Natural Interaction'. *Multimedia Tools and Applications* 38.3, pp. 295–305.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Science & Business Media.

Velchev, Y.; Radeva, S.; Sokolov, S. & Radev, D. (2016). 'Automated Estimation of Human Emotion from EEG Using Statistical Features and SVM'. In: *Proc. of the 2016 Digital Media Industry Academic Forum (DMIAF)*, pp. 40–42.

Ververidis, D. & Kotropoulos, C. (2008). 'Fast and Accurate Sequential Floating forward Feature Selection with the Bayes Classifier Applied to Speech Emotion Recognition'. *Signal Processing* 88.12, pp. 2956–2970.

Vogt, T. & Andre, E. (2005). 'Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition'. In: *Proc. of the 2005 IEEE Int. Conf. on Multimedia and Expo*, pp. 474–477.

Wagner, J.; Jonghwa Kim & Andre, E. (2005). 'From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification'. In: *Proc. of the 2005 IEEE Int. Conf. on Multimedia and Expo*, pp. 940–943.

Wagner, J. (2009). *The Augsburg Biosignal Toolbox*. Tech. rep.

Wallbott, H. G. (1998). 'Bodily Expression of Emotion'. *European Journal of Social Psychology* 28.6, pp. 879–896.

Wang, W.; Enescu, V. & Sahli, H. (2013). 'Towards Real-Time Continuous Emotion Recognition from Body Movements'. In: *Proc. of the Int. Work-*

shop on Human Behavior Understanding. Ed. by Salah, A. A.; Hung, H.; Aran, O. & Gunes, H. Cham: Springer International Publishing, pp. 235–245.

Wang, X.-W.; Nie, D. & Lu, B.-L. (2011). 'EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines'. In: *Proc. of the Int. Conf. on Neural Information Processing*. Ed. by Lu, B.-L.; Zhang, L. & Kwok, J. Springer Berlin Heidelberg, pp. 734–743.

Watzlawick, P. (1964). *An Anthology of Human Communication*. Science and Behavior Books.

Watzlawick, P.; Bavelas, J. B. & Jackson, D. D. (2011). *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes*. New York: WW Norton & Company.

Wells, B. & Macfarlane, S. (1998). 'Prosody as an Interactional Resource: Turn-projection and Overlap'. *Language and Speech* 41.3-4, pp. 265–294.

Wendemuth, A. & Biundo, S. (2012). 'A Companion Technology for Cognitive Technical Systems'. In: *Proc. of Cognitive Behavioural Systems*. Springer Berlin Heidelberg, pp. 89–103.

Weninger, F.; Wöllmer, M. & Schuller, B. (2015). 'Emotion Recognition in Naturalistic Speech and Language – A Survey'. In: *Emotion Recognition*. John Wiley & Sons, Ltd, pp. 237–267.

Whitley, B. E. (1997). 'Gender Differences in Computer-related Attitudes and Behavior: A Meta-analysis'. *Computers in Human Behavior* 13.1, pp. 1–22.

Witkower, Z. & Tracy, J. L. (2018). 'Bodily Communication of Emotion: Evidence for Extrafacial Behavioral Expressions and Available Coding Ssystems'. *Emotion Review* 11.2, pp. 184–193.

Yang, L.-C. (2001). 'Visualizing Spoken Discourse: Prosodic Form and Discourse Functions of Interruptions'. In: *Proc. of the 2nd SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, pp. 1–10.

Yannakakis, G. N. & Hallam, J. (2008). 'Entertainment Modeling through Physiology in Physical Play'. *International Journal of Human-Computer Studies* 66.10, pp. 741–755.

Zacharatos, H.; Gatzoulis, C.; Chrysanthou, Y. & Aristidou, A. (2013). 'Emotion Recognition for Exergames Using Laban Movement Analysis'. In: *Proc. of the 6th Int. Conf. on Motion in Games (MIG'13)*. ACM, 39:61–39:66.

Zacharatos, H.; Gatzoulis, C. & Chrysanthou, Y. L. (2014). 'Automatic Emotion Recognition Based on Body Movement Analysis: A Survey'. *IEEE Computer Graphics and Applications* 34.6, pp. 35–45.

Zajonc, R. B. (1980). 'Feeling and Thinking: Preferences Need No Inferences'. *American Psychologist* 35.2, pp. 151–175.

Zepf, S.; Dittrich, M.; Hernandez, J. & Schmitt, A. (2019). 'Towards Empathetic Car Interfaces: Emotional Triggers While Driving'. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. New York, NY, USA: ACM, LBW0129:1–LBW0129:6.

Zhang, H. (2004). 'The Optimality of Naive Bayes'. In: *Proc. of the 17th Int. FLAIRS Conf. (FLAIRS2004)*, pp. 562–567.

Zhen, B.; Wu, X.; Liu, Z. & Chi, H. (2000). 'On the Importance of Components of the MFCC in Speech and Speaker Recognition'. In: *Proc. of the 6th Int. Conf. on Spoken Language Processing (ICSLP'00)*, pp. 487–490.

Declaration of Honour

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 non-monetary 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
- Plagiarised 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 a similar form in Germany or in any other country. It has not yet been published as a whole.

Magdeburg, 28-01-2020

_____

Signature