OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INSTITUT FÜR INFORMATIONS- UND
KOMMUNIKATIONSTECHNIK (IIKT)

Multimodal Automatic User Disposition Recognition in Human-Machine Interaction

DISSERTATION

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

von

Dipl.-Inf. Ronald BÖCK

geb. am 29.09.1982 in Sangerhausen

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

Gutachter: Prof. Dr. rer. nat. Andreas WENDEMUTH
Prof. Dr.-Ing. Michael WEBER
Prof. Dr.-Ing. Christian DIEDRICH

Promotionskolloquium am 23.09.2013

Für Zarah und meine Mutter

Danksagung

Nur durch die vielseitige Unterstützung meiner Kollegen und Freunde wurde es mir möglich, meine Dissertation vorzulegen.

Ich möchte mich zuerst bei meinem Doktorvater Prof. Dr. Andreas Wendemuth bedanken, der mir in all den Jahren ein freundschaftlicher Mentor und Ansprechpartner war. Neben der fachlichen Unterstützung danke ich auch für sein Vertrauen, welches mir für meine Forschung und Tätigkeit am Lehrstuhl das rechte Umfeld offerierte. Darin sind auch die vielen hilfreichen Tipps, die für meine wissenschaftliche Weiterentwicklung von fundamentaler Bedeutung waren, eingeschlossen. Ich danke Prof. Wendemuth zudem für die Unterstützung im Zusammenhang mit meinem Gastwissenschaftleraufenthalt am Trinity College Dublin.

Weiterhin danke ich Herrn Prof. Dr. Michael Weber (Universität Ulm) und Herrn Prof. Dr. Christian Diedrich (Otto-von-Guericke-Universität Magdeburg) für die Bereitschaft zur Begutachtung der vorgelegten Dissertation. Mir ist bewusst, dass Sie beruflich stark eingebunden sind, weshalb ich ihre Bereitschaft um so mehr schätze.
Darüber hinaus gilt mein Dank den weiteren Mitgliedern der Promotionskommision, Herr Prof. Dr. Bertram Schmidt und Herr Prof. Dr. Marco Leone.

Mir ist es auch ein Anliegen, allen Mitarbeitern – ja Freunden – am Lehrstuhl Kognitive Systeme zu danken. Unsere fachlichen Diskussionen und Gespräche haben mir immer viel bedeutet. Stefan Glüge danke ich für die Jahre, in denen wir uns ein Büro geteilt und uns auf unsere uns eigene Art wissenschaftlich beflügelt haben. Ingo Siegert danke ich besonders für die tolle Zeit, während wir *ikannotate* programmiert haben. Damit schließe ich aber keinen Kollegen aus! Für mich war es eine großartige Zeit mit euch.

Ich erwähne auch zusammenfassend alle Kollegen, mit denen ich die Ehre hatte und habe gemeinsam die Workshops MA3 2012, T2CT 2013 und ERM4HCI 2013 zu veranstalten. Die Diskussionen und die Organisation haben mir sehr viel Freude bereitet. Ich hoffe, es werden noch viele Workshops in den jeweiligen Reihen folgen.

Dem Land Sachsen-Anhalt sei für die Bereitstellung meiner Stelle gedankt, durch die ich auch viel Erfahrung in der Betreuung von Studenten und Lehrveranstaltungen sammeln konnte.

Außerdem möchte ich dem Sonderforschungsbereich SFB/TRR 62 „Eine Companion Technologie für Kognitive Technische Systeme“ gefördert durch die Deutsche Forschungsgemeinschaft danken. Im Rahmen dieses Projektes war es mir möglich, einen Gastwissenschaftleraufenthalt am Trinity College Dublin wahrzunehmen. Darüber hinaus bot dieser mir eine Plattform, um meine Forschung im Kontext der Dispositionserkennung aus gesprochener Sprache im Bereich von natürlicher Mensch-Maschine-Interaktion zu vertiefen.

Ganz besonders danke ich meiner Familie, die mich in all der Zeit unterstützt hat. Ohne euch wäre dies alles nicht möglich gewesen und ich nicht der, der ich jetzt bin. Während der Zeit des Schreibens habt ihr oft auf mich Rücksicht nehmen müssen. Das vergesse ich euch nie.

Insbesondere richtet sich meine Hochachtung und mein Dank an meine Eltern und meine Verlobte Zarah. Besonderen Dank möchte ich meiner Mutter sagen, die nach dem Tod meines Vaters in allen Belangen für mich da war.

Weiterhin denke ich an alle meine Freunde, die dafür gesorgt haben, dass ich mich in Magdeburg zu Hause fühle.

Als Erkenntnis bleibt: Eine Dissertation verfasst nur einer – was aber nur gelingt, wenn ihm viele den Rücken stärken. Danke!

Zusammenfassung

DIE Erkennung von Dispositionen ist aktuell ein wichtiger Untersuchungsgegenstand in der Analyse von Mensch-Maschine-Interaktionen. Die vorliegende Arbeit untersucht den Aspekt der Dispositionserkennung und -klassifikation aus dem Blickwinkel der Sprachverarbeitung.

Zunächst wird der Begriff *Disposition* als solcher definiert. Hierbei ist eine klare Abgrenzung zur Verwendung des gleichen Begriffs in der Psychologie notwendig; vielmehr wird das Augenmerk auf eine Definition unter technischem Blickwinkel gelegt. Disposition in einem systemisch inspirierten Sinn ist eine universelle Beschreibung der Situiertheit und der Eigenschaften eines Nutzers. Darin sind weiterhin Informationen über die aktuelle Interaktion enthalten; diese gehen in der Beschreibung der Situiertheit auf. Ausgehend von der allgemeinen Definition der Disposition lassen sich weitere Begrifflichkeiten wie *Emotion*, *Stimmung*, *Intention*, *Situiertheit*, etc. in den Gesamtkontext einordnen. Neben der Einführung von Begrifflichkeiten werden auch Annotations- beziehungsweise Labellingverfahren, die im Zusammenhang mit Klassifikationsexperimenten Verwendung finden, diskutiert. Auf deren Grundlage werden Disposition, wie positiv oder negativ verlaufende Interaktionen, eingeführt.

Basierend auf den in der Arbeit vorgestellten Konzepten werden Klassifikationsexperimente zur Dispositionserkennung aus gesprochener Sprache durchgeführt. Hierfür werden zunächst Datensätze vorgestellt, die ich in dieser Arbeit nach gespielten und nicht-gespielten Daten unterscheide. Zur ersten Kategorie zählen unter anderem die Korpora EmoDB und eNTERFACE, auf deren Grundlage Parametersätze für Klassifikatoren und mögliche Experimentalgestaltungen ermittelt wurden. Ausgehend von diesen Betrachtungen konnten die gewonnenen Parameter auf natürliche Datensätze, das heißt nicht-gespieltes Material übertragen werden. Explorative Untersuchungen haben diesen Sachverhalt bestätigt. Insbesondere wurden zwei Datensätze, nämlich LAST MINUTE und EmoRec, untersucht, die beide im SFB/TRR 62 aufgenommen wurden. Für Untersuchungen wurde verstärkt der EmoRec-Datensatz herangezogen.

In den Experimenten zeigen besonders *Gaussian Mixture Models* signifikant gute Erkennungsleistungen für die Klassifikation von Dispositionen. Da diese in den erweiterten Kontext von *Hidden Markov Models* eingebettet sind, ist es möglich *Gaussian Mixture Models* mit dem Hidden Markov Toolkit der Universität Cam-

bridge zu trainieren und zu evaluieren.
Bei der audiobasierten Analyse der beiden natürlichen Datensätze wurden gute Erkennungsleistungen erzielt. Für Last Minute erreichten die Klassifikatoren eine gewichtete mittlere Genauigkeit von 32.0% bei vier Klassen (*baseline*, *challenge*, *listing* und *waiuku*). Für die Untersuchung des EmoRec-Korpus wurden zwei Validierungsansätze verfolgt: die interindividuelle und die intraindividuelle Validierung. Bei der interindividuellen Validierung wird das Datenmaterial aller Sprecher, mit Ausnahme eines Sprechers, für das Training verwendet. Dessen Daten werden ausschließlich für den Test benutzt. Im Gegensatz dazu wird bei der intraindividuellen Validierung nur das Datenmaterial jeweils eines Sprecher verwendet, welches in ein Trainings- und ein Testset unterteilt wird. Neben der Unterscheidung der Validierungsansätze kann auch der EmoRec-Datensatz selbst in zwei Sub-Korpora unterteilt werden, die jeweils einen Durchlauf des Szenarios enthalten. Für EmoRec I wurden gewichtete mittlere Genauigkeiten von 55.1% beziehungsweise 70.0% für inter- und intraindividuelle Validierung erzielt. Hierbei wurde nach *positiver* und *negativer Disposition* unterschieden. Im Hinblick auf EmoRec II sind aktuell nur acht Teilnehmer für automatische Analysen verfügbar, die mittels interindividueller Validierung untersucht wurden. Hierbei wurde eine gewichtete mittlere Genauigkeit von 52.9% erreicht.
Auf der Grundlage der Arbeiten am EmoRec I, wird im Rahmen meiner Forschung eine Arbeitsumgebung für die semi-automatische Annotation vorgestellt. Eine semi-automatische Annotation eines Datensatzes ist sinnvoll, da die Annotation im Normalfall ein zeit- und kostenintensiver, manueller Prozess ist, der sich so effizienter gestalten lässt. Das Vorgehen wird exemplarisch anhand des Datensatzes EmoRec veranschaulicht und es werden die erzielten Ergebnisse diskutiert. In dieser Arbeitsumgebung werden relevante, affektbehaftete Videosequenzen auf Grundlage von Audioanalysen detektiert und einem menschlichen Annotator vorgeschlagen. Dieser annotiert die entsprechenden Sequenzen nach Vorgaben des *Facial Action Coding Systems*. Durch eine semi-automatische Annotation ist es möglich, die zu betrachtenden Sequenzen zahlenmäßig zu reduzieren, was den zeitlichen Aufwand erheblich senkt. Mit der aktuellen Konfiguration der integrierten Klassifikatoren wird eine falsch positive Rate von 18.8% und eine falsch negative Rate von 38.1% erreicht. Diese Zahlen bieten Ansatzpunkte für weitere Optimierung, die in darauf aufbauenden Arbeiten durchgeführt werden können. Die Arbeitsumgebung an sich ist flexibel gestaltet, so dass eine Übertragung auf andere Modalitäten möglich ist.

Ausgehend von den bisherigen Überlegungen zur Disposition in der Mensch-Maschine-Interaktion wird das Konzept des *Involvements* eingeführt. Nach meinem Verständnis kann das *Involvement* komplementär zur Disposition verstanden werden, da dies einen bestimmten Zustand eines Nutzers und dessen Situiertheit widerspiegelt, insbesondere, ob der Nutzer an einer Interaktion partizipiert oder nicht. Dabei ist die Sichtweise eher durch technische Einflüsse getrieben und kann daher von den Betrachtungen, wie sie zum Beispiel Psychologen treffen würden, abweichen. Die Untersuchung von *Involvement* aus gesprochener Sprache ist ein relativ neues Feld, besonders unter dem Blickwinkel der Dispositionserkennung. Für die Analysen in dieser Arbeit wurde der *TableTalk*-Datensatz herangezogen. Da die audiobasierten Analysen noch am Anfang stehen, liegt in dieser Arbeit das Augenmerk auf der Annotation des Datensatzes nach Aspekten des *Involvements*. Hierbei wird auch die Reliabilität der Annotation betrachtet. Krippendorff's α (ordinal) liegt für diese Annotation bei $\alpha_o = 0.1562$. Dieser Wert wird mit Ergebnissen auf Datensätzen, die ähnliche Grundmerkmale aufweisen, verglichen. Basierend auf den Grundlagenbetrachtungen dieser Arbeit wird ein Arbeitsprogramm für weiterführende Untersuchungen auf dem Gebiet des *Involvements* entwickelt und in den Kontext der Kontrolle von Interaktionen eingebettet.

Besonders die systemische Kontrolle von Interaktionen wird in den Fokus der Wissenschaft rücken. Aus meiner Sicht wird es nur so langfristig möglich sein, geeignete und gezielte Aktionen und Reaktionen des Systems in einer Mensch-Maschine-Interaktion zu realisieren. Dazu ist es notwendig, dass ein technisches System in geeigneter Form in die Interaktion eingreift, sich assistiv auf den Nutzer einstellt und ihn damit zielführender unterstützen kann. Somit kann es gelingen, aus omnipräsenten technischen Systemen eine Technologie im Sinne eines *Companion* – eines Begleiters – zu generieren.

Abstract

THE recognition of dispositions from speech is an important issue in the analyses of Human-Machine Interactions. This thesis examines the classification of disposition under various aspects.

At first, the term *disposition* is defined and thus, differentiated from the meaning in the sense of psychology. A disposition in a technologically inspired sense is a universal description of the user's situatedness and his characteristics. Further, it includes information on the particular interaction. From this definition, several user information like his emotion and mood, his intention, etc. are subsumed by disposition. For a general overview, the terms *emotion*, *mood*, and *situatedness* are defined as well since they are highly connected to dispositions. Furthermore, a brief introduction in labelling methods is given which are used as concepts for classification in the experiments of this thesis. From these concepts, high-level dispositions like positive or negative interactions are derived.

Based on the mentioned concepts I designed, arranged, and conducted several disposition classification experiments. Considering data sets, I distinguish i) acted and ii) non-acted material. Utilising the first kind of materials, namely the EmoDB and eNTERFACE corpora, I investigated parameter sets and derived a system setup for further analyses. Transferring the conclusions to acted material, a switch towards naturalistic, that means, non-acted data sets was possible which is shown by exploratory investigations. In particular, these data sets are: LAST MINUTE and EmoRec, whereas I mainly used the latter, recorded in the context of the SFB/TRR 62.

Especially, Gaussian Mixture Models showed a remarkable good performance in classification of dispositions from speech. They are embedded into the framework of Hidden Markov Models and trained as well as tested applying the Hidden Markov Toolkit by the Cambridge University.
Both naturalistic corpora are investigated in terms of dispositions by audio analyses where I achieved recognition results of 32.0% Weighted Average accuracy on LAST MINUTE (four classes) and more than 55.1% on EmoRec (two classes, interindividual validation). In particular, EmoRec results are distinguished according to interindividual and intraindividual validation. Interindividual validation considers the speaker independently from each other and thus, trains the classifier on all samples of all users, excluding the material of one speaker

which is used for testing. In contrast, intraindividual validation utilises only samples of a certain speaker, splitting the material in training and test sets. Thus, intraindividual validation reflects the characteristics of a certain speaker in different dispositions. Applying Gaussian Mixture Models, I gain Weighted Average accuracies on EmoRec, to be specific on EmoRec I, of 55.1% for interindividual validation and 70.0% for intraindividual validation, respectively. For EmoRec II, so far, just a subset of eight participants could be analysed whereas 52.9% Weighted Average accuracy for interindividual validation was achieved.
Highly related to the audio analyses of EmoRec is the semi-automatic annotation because it was tested on this material. In the annotation framework, relevant video sequences are preclassified based on audio analyses. Those sequences are afterwards manually annotated by trained coders according to the Facial Action Coding System. As this is quite time consuming the semi-automatic annotation helps since the total amount of sequences to be annotated is reduced drastically. In the current setting of the framework, false acceptance and false rejection rates of 18.8% and 38.1%, respectively, for the identification of relevant sequences, were achieved.

Based on the considerations on basic dispositions I further, introduced the concept of involvement in conversation. Moreover, the involvement is complementary to disposition, reflecting whether the participant is in an interaction or not. Such type of analysis is a quite novel approach in the automatic analyses of speech, especially, in the context of disposition recognition. Therefore, the way of annotating a corpus, in particular, the TableTalk data set, according to involvement is discussed regarding also the reliability of the labelling. As this is an upcoming field of research, open issues are identified and also discussed in relation to the controlling of an interaction, from a system's point of view. For this, the focus is on proper reactions and on how to influence the interaction to support the system's user in a companion-like manner.

Contents

List of Figures **xiii**

List of Tables **xv**

Notations and Conventions **xvii**

1 Introduction **1**

1.1 Motivation 2

1.2 Disposition Recognition from Speech 4

1.2.1 Emotions and Moods 4

1.2.2 Disposition Recognition 6

1.2.3 Categorical and Continuous Labelling Systems 8

1.3 Inter- vs. Intraindividual Validation 12

1.3.1 Measures 13

1.3.2 Interindividual Validation 15

1.3.3 Intraindividual Validation 16

1.3.4 Connection of Both Validation Systems 17

1.4 Involvement in Conversation 17

1.5 Control of Human-Machine Interaction 19

2 State of the Art **21**

2.1 Human-Machine Interaction 22

2.2 Reviewing Disposition Recognition from Speech 25

2.3 Data Sets 28

2.3.1 Dispositional Data Sets 28

2.3.2 Data Sets of Conversations 30

2.4 Methods of Disposition Recognition from Speech 36

2.4.1 Feature Sets for Disposition Recognition 36

2.4.2 Classification Methods for Disposition Recognition 38

2.5 Automatic Analysis of Involvement 42

2.6 Summary 45

3 Data Sets **47**

3.1 Acted Data Sets 47

3.1.1 EmoDB 48

3.1.2 eNTERFACE 49
3.2 Non-Acted Data Sets 50
3.2.1 Last Minute 51
3.2.2 EmoRec I+II 53
3.3 TableTalk 56
3.4 Summary 58

4 Methods for Inter- vs. Intraindividual Disposition Recognition 59
4.1 From the Acoustic to Training Material 60
4.1.1 The Annotation Process 60
4.1.2 ikannotate 61
4.1.3 Semi-automatic Annotation 64
4.2 Features 67
4.2.1 Emotion and Disposition Recognition from Speech 68
4.2.2 Semi-automatic Annotation 73
4.3 Classifiers 80
4.3.1 Hidden Markov Models 81
4.3.2 Evaluating Parameter Sets 84
4.3.3 Recurrent Neural Networks 91
4.3.4 Comparing Hidden Markov Models and Neural Networks . 92
4.4 Fusion Aspects 97
4.5 Summary 99

5 Non-Acted, Multimodal Experiments 101
5.1 Audio Setup Only 102
5.1.1 Results on Last Minute 102
5.1.2 Results on EmoRec I 110
5.1.3 Discussion 116
5.2 Bimodal Setups 117
5.2.1 Comparison of Audio & Facial Expressions 118
5.2.2 Results of Semi-automatic Annotation 122
5.2.3 Comparison of Audio & Biophysiological Features 124
5.2.4 Discussion 125
5.3 Multimodal Setup 127
5.3.1 Experimental Setup 127
5.3.2 Audio Classification Results 129
5.3.3 Discussion in the Context of Multimodality 130
5.4 Summary 131

6 Group Involvement 133
6.1 Why Consider Involvement in Group Conversations? 134
6.2 Annotating Changes in Group Involvement 135
6.2.1 Annotation of Day1 - Group 136
6.2.2 Annotation of Day1 - Each Participant 143
6.3 Detecting Changes in Group Involvement 147
6.3.1 Automatic Detection of Changes 147
6.3.2 Systems Reaction towards a User 151
6.4 Summary 152

7 Conclusion and Future Work 155
7.1 Conclusion 155
7.2 Open Issues for Future Work 163
7.2.1 Advanced Semi-automatic Annotation 163
7.2.2 Naïve, Multiparty Disposition Recognition 165
7.2.3 Fusion Methods in Naïve Disposition Recognition 167
7.2.4 Involvement in Conversations 168
7.2.5 Disposition Recognition for Interaction Control 170

Glossary 171

Abbreviations 173

Symbols 175

References 177

List of Figures

1.1 Emotion cone according to Plutchik. 10
1.2 GEW as introduced by Scherer. 11

3.1 Workflows of the EmoRec. 54

4.1 Framework for a semi-automatic annotation. 65
4.2 Feature flow for a semi-automatic annotation. 65
4.3 Example and range in semi-automatic annotation. 67
4.4 Linear model of speech production. 69
4.5 Vowel triangle. 74
4.6 Graphical representation of the formant's bandwidth. 76
4.7 Workflow of a left-to-right Hidden Markov Model. 81
4.8 Mean accuracy dependent on the number of hidden states. 85
4.9 Mean accuracy dependent on the number of iterations. 87
4.10 Mean accuracy dependent on the feature set. 89
4.11 Structure of an SMRNN. 91
4.12 Overview of a decision level fusion architecture. 98
4.13 Overview of a mid level fusion architecture. 99

5.1 Mean accuracy values with corresponding standard deviations. . . 109
5.2 Experimental Sequences in PAD space. 111
5.3 Results of formants and bandwidths. 119
5.4 Results of intensity, pitch, and jitter. 121

6.1 Distribution of labels (group in total). 142
6.2 Combinations of numbers of assignments for each class. 143
6.3 Distribution of labels (male participants). 145
6.4 Distribution of labels (female participants). 146

List of Tables

2.1 Overview of dispositional corpora. 29
2.2 Performance of selected classifiers achieved on EmoDB. 40

3.1 Emotion's distribution in the EmoDB. 49
3.2 Sample's distribution per emotion in eNTERFACE. 50
3.3 Sample's distribution per scenario part in Last Minute. 53
3.4 Self-ratings of the participants. 56

4.1 Mean frequency values of formants. 75
4.2 Comparing of feature sets. 88
4.3 SMRNN configuration grouped by emotional class in EmoDB. . . 94
4.4 Recognition rates of HMM and SMRNN. 95
4.5 Correctness comparing SMRNNs and HMM$\Delta\Delta$. 96

5.1 Comparison of HMMs with different numbers of Gaussian mixtures.105
5.2 Recognition results of GMMs comparing MFCC and PLP features. 107
5.3 Interindividual recognition results using GMMs. 108
5.4 Interindividual and intraindividual recognition results on EmoRec I.114
5.5 Classification results according to the mixtures' number in GMMs. 122
5.6 Acceptance rate for semi-automatic annotation. 123
5.7 Classification results applying biophysiological features. 125
5.8 Sample's distribution in EmoRec II. 128
5.9 Interindividual classification results on EmoRec II. 129

6.1 Reliability values on the TableTalk corpus. 139
6.2 Distribution of samples in TableTalk for the group. 141
6.3 Reliability values for each speaker in the TableTalk corpus. 144
6.4 Classification results for detection of changes. 150
6.5 Confusion matrix for detection of changes. 150

Notation and Conventions

Personal pronoun: I want to point out that the work presented in this thesis evolved with the help and collaboration of my colleagues in Magdeburg, Ulm, and Dublin. As I am the author, the term "I" is used throughout my thesis. Nevertheless, I do not claim that this work could be done all by myself or that I am the first author of all publications reflected in the thesis. The list of publications at the end of the thesis names all the authors and co-authors that contributed to my research and I am contributing the other way around. Using the term "I" reflects further my part of the work when I am not first authored. Referring to colleagues' contributions I will use "they" (or similar terms), even if I am co-author.

Gender: Throughout this thesis I will use the male gender to specify persons, for instance actor, user, etc. Further, male personal pronouns like he and his will appear. I point out that is no discrimination but it is for reason of a better readability.

Chapter 1

Introduction

Contents

1.1 Motivation **2**

1.2 Disposition Recognition from Speech **4**

1.2.1 Emotions and Moods 4

1.2.2 Disposition Recognition 6

1.2.3 Categorical and Continuous Labelling Systems 8

1.3 Inter- vs. Intraindividual Validation **12**

1.3.1 Measures 13

1.3.2 Interindividual Validation 15

1.3.3 Intraindividual Validation 16

1.3.4 Connection of Both Validation Systems 17

1.4 Involvement in Conversation **17**

1.5 Control of Human-Machine Interaction **19**

THE way of human's interactions with systems has changed in the past years and will be changed in the future. What does it mean, the characteristic of interaction has changed? It was quite usual that an artificial system received inputs in an artificial way in the sense of human interaction. The inputs were either sensor data or commands given by keyboard and mouse. In the last years, the human way of interacting and communicating was introduced in Human-Machine Interaction (HMI). Thus, a more natural way of communication found its way into HMI, especially in the design of interfaces (cf. [Müller 2011; Carroll 2013]).

This chapter introduces the way of communication briefly and provides the reader with the basic ideas of advanced communication paradigms; in particular, emotion and disposition recognition from speech. Further, definitions which are used in this thesis are specified, especially to clarify those terms in the way they are utilised in this thesis.

1.1 Motivation

Communication and interaction are basic needs of humans. They are a general characteristics of our civilisation and do not depend on age, gender, or cultural imprint, whereas the characteristic is indeed preassigned by the cultural background. Hence, it is obvious that communication is always present in our life. Here, I do not narrow the communication towards an interaction of (multiple) humans, but include also technical devices like smartphones, computers, or even navigation devices, for instance.

The way of communication changed over the past years that means even in Human-Human Interaction (HHI) the style of interaction evolved. In the past, usually letters were written or in more advanced land-line phone calls were done. With the introduction of the first kinds of computers and networks this changed. Nowadays, it is common to use E-Mail, smartphones with different kinds of services, telephones, and video conferences. The world gets smaller in the sense that almost everybody can be reached by communication almost everywhere.

On the other hand, the technologies applied in communication remained the same over a long period, especially in the interaction with technical systems. It is quite common to use mouse and keyboard to control technical devices (cf. Section 1.5). However, in the last years, the interfaces of technical systems also have switched to a more natural handling (cf. e.g. [Karray et al. 2008; Elsholz et al. 2009; Kameas et al. 2009; Carroll 2013]). This means, touch screens and speech control have been introduced in the controlling and therefore, from my point of view, the human is more involved in the interactions. From this, it can be assumed that it is for a human much more easy to communicate or interact with a system. In general, the aspect of involvement is a quite important issue, in particular in a multi-party interaction, where the parties are either humans or technical systems. This issue is introduced in Section 1.4 and discussed in detail in Chapter 6.

Especially, speech controlling is a quite natural way of interacting with a technical system as humans are used to communicate via voice. So far, it is usual that the pure content of the message, which is further usually related to a specific domain, is analysed and the corresponding system reacts on specific commands or phrases. On the other hand, speech contains much more information as only the pure message. As Schulz von Thun already discussed in [Schulz von Thun 1981] communication has four aspects, summarised in the four-sides model: appeal, factual information, relationship, and self-revelation. In the following, I focus on the latter since this is heading to the issue of disposition recognition. Everybody experiences that, while speaking, further information, which are not directly related to the content of the utterance, are transmitted. Especially, the meta-information of self-revelation, or to be more precise the emotional and dispositional parts of this information are essential in the HHI because this provides the possibility to appraise the content. Further, the whole situation can be assessed to classify what is said; that means, the content is assessed by the additional information.

The aspects of Schulz von Thun are not only valid for HHI; this can be seen in HMI as well. Of course, the aforementioned parts are distinctly different, for instance, the factual information is in the focus. Nevertheless, the user creates a relationship with the system (cf. [Dennett 1987; Lange & Frommer 2011; Rösner et al. 2011]) and hence, it can be assumed that the four-sides model can be applied, too. From this, it is obvious that besides speech recognition the automatic interpretation of the way how something is said has to be considered. Thus, this field of research started with the analyses of emotions and corresponding states as those can be distinguished in the voice more easily. Moreover, I argue for a more general view of self-revelation (cf. [Batliner et al. 2001] where a roadmap is presented). This means that several information like the user's intention, the environment and situation, the user's emotion, etc. have to be captured by a system and hence, it has to be equipped with the possibility to react in a proper way.

In Section 1.5 I will describe the controlling of a system and in addition, a sketch of a proper system's reaction in more details. However, the question arises: what is the purpose of recognising dispositions? As mentioned before, dispositions are an essential part of the communication. Further, assessing the user's reactions in a precise way, improves overall recognition results and this leads to more robust communication at large. This includes also avoiding abortion of the dialogue. In general, the main goal is to establish a natural communication in the HMI. This thesis can only present some aspects of the ongoing research in the automatic disposition recognition from speech due to the complexity of the field. It is focused on the analyses of near real-life scenarios comparing inter- and intraindividual recognition experiments.

1.2 From Emotion to Disposition Recognition from Speech

As I already motivated (cf. Section 1.1) it is necessary to incorporate feelings, moods, and, from my point of view very important, dispositions of a user in an HMI. So far, those terms are quite vague and therefore, I introduce and arrange them from an engineer's point of view. I am aware that the usage of these term are quite different in psychology and computer sciences. However, in this thesis they will be used as specified in the following.

Further, I will discuss why emotion recognition evolved towards a disposition recognition from speech. This leads also towards different ways of categorising human behaviour, especially in an emotional or dispositional way (cf. Section 1.3).

1.2.1 Emotions and Moods

First of all, to get the meaning of the term *emotion* I concentrate on the common definition given by the Shorter Oxford English Dictionary [*Shorter Oxford English Dictionary* 2002] (cf. Definition 1.1 a)) which I will prefer here, and an extended version by Merriam-Webster [Merriam-Webster 1998] (cf. Definition 1.1 b)). Both definitions represent the common meaning of the term and are influenced by interpretations of psychologists. Therefore, they differ from a definition which is technically inspired.

Definition 1.1 *a) An emotion is a strong feeling deriving from one's circumstances, moods, or relationships with others.*
b) An emotion is a psychic and physical reaction (as anger or fear) subjectively experienced as strong feeling and physiologically involving changes that prepare the body for immediate vigorous action.

As it can be seen, an emotion is a strong reaction of a human related to several circumstances. Especially, the adjective "strong" indicates that emotions are quite expressive. Therefore, they can be recognised and distinguished relatively easy, in particular, by humans. According to the discussion in Chapter 2 expressive emotions can be classified by several types of classifiers (cf. e.g. [Schuller et al. 2009a]).

From this, I define the term emotion in a rather classifier oriented and thus, technically oriented way indicating the shortness of such an event. It means that an emotional expression is usually confined to single utterances or a few succeeding sentences. Furthermore, by this shortness I can apply methods (cf. Section 4.3) as they were utilised in speech recognition which assume stability in short periods of an audio signal. Moreover, my definition extends the term emotion towards a description that covers also reactions which occur in HMI.

Definition 1.2 *An emotion is a short occurrence of a strong feeling wheras a mood represents the background of the emotion. Emotions are direct reactions on recent actions and events. Their verbal expression is connected to single or short sequences of utterances and it can be expressed either in HHI or HMI.*

As Batliner et al. [Batliner et al. 2001] discussed (cf. also Section 3.2) expressive reactions are quite rare in an HMI which resembles to a natural way of interacting. Therefore, the evolvement of emotions has to be considered in each interaction. This is what humans are doing in each HHI. Otherwise, conversations have the danger of breakdown and the dialogue situation ends, for instance in call centres (cf. [Vlasenko & Wendemuth 2009]).

This leads to a more general term which describes the human's behaviour on longer time periods. Furthermore, *moods* are the basis of emotions (cf. Definition 1.2). As it is given in Bertelsmann Universal Lexicon [Bertelsmann 1993] mood is defined as follows.

Definition 1.3 *Mood is a longer lasting feeling connected to all experiences. Further, it is a state depending on the human's body and therefore, can be influenced either by humans or artificial systems.*

In contrast to my definition of emotions (cf. Definition 1.2) the mood is longer lasting, that means it stays stable for at least a longer sequence of utterances or, which holds in general, for the conversation in total. It is obvious that this is related to the duration of a conversation. Without restrictions, I assume that a conversation lasts from a few minutes up to some hours. With today's way of communication shorter conversations can be assumed. Therefore, a stability of a mood can be assumed. Due to that, concepts of speech recognition such as stability assumption can be transferred to a novel mood recognition in a broader sense. Hence, the term mood and the corresponding concept extends the pure classification of single emotional events towards more general aspects. Furthermore, it incorporates the personality of a user (cf. Section 1.3.3) in the conversation and by classification of such events in the HMI.

In addition to the given definition (cf. Definition 1.3) I state that each mood is also part of an HMI and hence, can influence it. Therefore, I argue for an analysis and classification of moods in advance in order to modify in the classification of emotions. As said in Definition 1.2 on the preceding page moods are the basis of every emotional expression. A better understanding will improve the classification of emotions because their occurrence is dependent on moods. Thus, even equal emotions differ in their characteristic by underlying different moods. This aspect will be discussed in Section 1.2.2.

1.2.2 Disposition Recognition

As I already introduced in Section 1.2.1 the handling of user characteristics can be ordered hierarchically. This means that one can find an order in the generalisation of user's states related to his mental situation. The hierarchy is as follows, ordered upwards: emotion, mood, and disposition.

Before I will discuss dispositions and the classification in a technical sense I consider the term *disposition* from psychological point of view, that means, give a corresponding definition which is according to [Bertelsmann 1993].

Definition 1.4 *Disposition is a natural arrangement or an acquired characteristic to practise certain characteristical occurrences or to carry out certain experiences. Further, it is the receptiveness for certain influences.*

From my point of view, this definition is too narrow. I will explain this point in the following and thus, derive a definition of disposition in a more technical sense which also gives reference to HMI.

As it will be shown in Chapter 3 in the past years one can see an evolution in the field of emotion recognition from speech. It started with the speech recognition in the 1960s where the basics towards such a technology were laid. Further based on psychologists' knowledge the methods were improved and utilised for classification of basic emotions (cf. Ekman's basic emotions [Ekman 1992]). As mentioned in Section 1.2.1, those are derived and influenced by the speaker's moods which leads to the necessity of mood recognition. However, each interaction - it does not matter if either HHI or HMI - is influenced by the situation and from this, each speaker and his corresponding behaviour is also affected by the environment and his mental state. The latter is reflected by the psychologists' definition of disposition (cf. Definition 1.4). Nevertheless, the situation of the environment and the speaker itself is quite important to understand his reactions, especially in the context of HMI (cf. Section 1.5).

This leads us to the concept of *situatedness*. In particular, it reflects the context of learning or in other words, every knowledge and feature depends on a certain context (definition and interpretation are inspired by [Simpson 2002]). Consequently, I derive the definition in a more technical way.

Definition 1.5 *Situatedness describes the entire context of an interaction and conversation including information of how many persons are in the surrounding, what is the interpersonal relationship of the interactors, the given abilities and limitations of the interactors or a technical system, the general conditions of an interaction, etc.*

In the classical emotion and mood recognition and classification the situatedness is not reflected in a proper way. In fact, it is neglected and hence, a source of information to interpret a reaction in a broader sense is not considered. As I will discuss in detail in Section 1.5 disposition recognition in general has to be considered in a natural HMI, and in particular, disposition recognition from speech, as speech is on the one hand, a natural way of interaction and on the other hand, related to the interactor's mental state, that is emotions, moods, and behaviour (cf. e.g. [Wendt & Scheich 2002; Gnjatovic & Rösner 2008]).

With these preliminary considerations and based on Definition 1.4 on the preceding page and Definition 1.5, I define the term disposition as follows.

Definition 1.6 *Disposition is the universal description of the user's situatedness and further, of his certain characteristics, including emotions and moods, that are given in a particular interaction. Furthermore, it is influenced by the knowledge and contextual information of the certain interaction that is the background of the interactor.*

From this definition, it is obvious that the disposition is influenced by the user's mental state as well as his personality and intentions. Furthermore, it shows the complexity of the task which indicates the necessity for multimodal analyses. Just by the consequent use of several modalities a handling of this issue is possible. In this thesis, I concentrate on a part of the investigation, namely the disposition from speech. Combing different modalities a universal view on the disposition can be achieved. This provides a technical system, especially as it is intended by companion systems (cf. [Wendemuth & Biundo 2012]), with a detailed view and idea of its counterpart, namely the system's user.

1.2.3 Categorical and Continuous Labelling Systems

As the terms of emotion, mood, and disposition are defined now, I introduce the systems which are used to indicate corresponding classes or values to evaluate and measure characteristics of those terms. This is usually done by applying so called labelling systems. Such a paradigm defines categories or values which are related to a state of the user. The process of annotation, which is also called labelling, will be described in Section 4.1.

It is common to distinguish three kinds of labelling systems: i) categorical, ii) quasi-continuous, and iii) continuous labelling paradigms. The order represents also the historical evolution of these systems and moreover, the complexity for the annotators in the assessment process.

Categorical Labels

Categorical labels which represent emotional states of humans have been observed for a long time. Even Greek philosophers (e.g. Aristotle and Galen of Pergamon) ordered emotional characteristics according to the behaviour of humans. They combined emotions and moods and established terms like melancholic, phlegmatic, etc. This indicates that such combination includes the origin of emotions in moods (cf. Definition 1.1 on page 4). This classification paradigm was preserved till the 19th century. At this time, a more biologically inspired analysis of emotions was introduced. For instance, Darwin started systematic observations and evaluations of emotional expressions [Darwin 1872]. For this, he applied prototypical photos to classify emotions in humans. Further, in [Darwin 1872] he also investigated emotional behaviour of animals and compared this to human's reactions. To get clear decisions he concentrated on facial expressions which were mimed by actors. By the structure of the book, Darwin distinguished two main categories, namely *low* and *high spirits* [Darwin 1872]. Further, a more fine granulated classification has been discussed, for instance, anxiety, grief, joy, love, shame, shyness, modesty, guilt, surprise, fear, etc.

In the context of such observations psychologists analysed also the behaviour and reactions of apes and humans and found so called *Basic Emotions* [Ekman 1992; Plutchik 2001]. The term defines sets of emotions which are generally shown. They are derived from numerous observations by psychologists. Unfortunately, the composition of the sets differs according to the authors of the corresponding studies. Two sets are widely used in psychology as well as in the automatic classification of emotions (cf. Section 3.1) which were defined by Ekman and Plutchik.

In the following, I concentrate only on findings that are related to humans, even if they are valid for apes, too. For instance in [Ekman 1992], the most famous and largest set of Basic Emotions is mentioned. For his observations, Ekman asked actors to play the corresponding emotions. Afterwards, these facial expressions were presented to test persons who are no actors. So, he derived the following emotions: anger, boredom, disgust, fear, joy, neutral, and sadness. Based on observations of Ekman on facial expressions Ekman & Friesen established a coding system called Facial Action Coding System (FACS) [Ekman & Friesen 1978] which describes emotional reactions in the face in a standardised way referring to Action Units (AUs). Those AUs can be used to classify emotions as it is quite hard to mentally influence the muscles' activities generating the indicating AUs (cf. Section 3.2.2). The introduction of Basic Emotions according to Ekman is presented because I mainly used these categories in my experiments (cf. e.g. Section 4.3.2).

In contrast to Ekman, Plutchik defined the following set: acceptance, anger, anticipation, disgust, fear, joy, sadness, and surprise [Plutchik 2001; Hiroshige et al. 2009]. An advanced set is given in Figure 1.1 on the next page. Nevertheless, he also relies on material generated with the aid of actors, especially, facial expressions.

As I pointed out, Basic Emotions are derived from facial expressions. Are they valid to be used in emotion recognition from speech? As several studies, for instance, [Burkhardt et al. 2005; Martin et al. 2006; Schuller et al. 2009a], show, it is possible to transfer these concepts of Basic Emotions to speech recognition. The studies applied listening tests where the participants as well as speakers were either actors or no actors. However, in both conditions Basic Emotions could be distinguished by the listeners. Inspired by these results, I concentrated in my experiments at first also on Basic Emotions and for this, I could get parameters for classifiers (cf. Section 4.3.2). These can be used in a much broader sense and on different materials, respectively, as it is shown in Section 5.1, for instance.

Quasi-Continuous Labels

Based on categorical labels a natural improvement of an annotation is the labelling with a bunch of classes. That means that still predefined classes are given, but a higher number of those. Especially, a weighting or grading of emotional classes can be established which introduces a granularity to the so-called Basic Emotions.

In [Plutchik 2001] the Basic Emotions are extended in such a way that an intensity value is assigned to each emotion. So far, such intensity holds for each emotion in total. However, it defines an order because Plutchik symbolised the intensity value in a cone (cf. [Plutchik 2001; Hiroshige et al. 2009] and Figure 1.1).

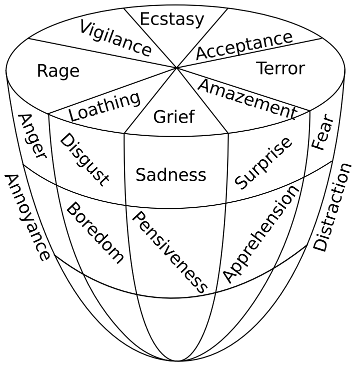


Figure 1.1: Plutchik's cone of emotions where single emotions are ordered by their quality (as the segments of a circle) and intensity (depth of the cone). The figure is take from [Siegert et al. 2012b]. Credits to Kim Hartmann and Ingo Siegert for permission to use it.

A more continuous way of interpreting emotions was introduced by Scherer as he ordered the emotions according to several dimensions, for instance in [Scherer 2005]. Furthermore, he extended the set of emotions which are either not as expressive or are more related to a natural behaviour of interactors. That means, they are more difficult to observe and to recognise. This paradigm is reflected in the Geneva Emotion Wheel (GEW) as introduced in [Scherer 2005]. A realisation of this model is given in Figure 1.2 on the facing page and was also used in the annotation tool called interdisciplinary knowledge-based annotation tool for aided transcription of emotions (ikannotate) (cf. [Böck et al. 2011b]). According to [Scherer 2005] the dimensions are *Pleasure* and *Control*. However, other experimental measures can be used to quantise emotions. Further well known sets are combinations of *Pleasure*, *Arousal*, and *Dominance* (cf. [Bradley & Lang 1994; Mehrabian 1996; Grimm et al. 2007]). Usually, two of these labels are posed in a two dimensional map like in the GEW.

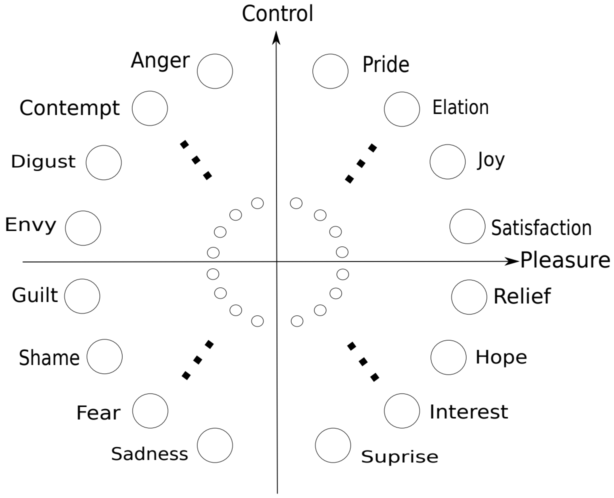


Figure 1.2: Geneva Emotion Wheel related to this introduced in [Scherer 2005] and used in ikannotate (cf. Section 4.1) [Böck et al. 2011b]. The granularity of the intermediate levels for each emotional expression is defined by the designer of the labelling process. In [Böck et al. 2011b] two additional labels are introduced, namely neutral and other (cf. Section 4.1).

Whether Plutchik's or Scherer's way of labelling, both can be applied by using, for instance, ikannotate to do the annotation on utterance level (cf. Section 4.1).

As I explained, such labelling paradigms are in between a full categorical and a continuous annotation, I call them quasi continuous labels because for each emotional expression like 'in high spirits' (cf. [Scherer 2005]) a mapping to a numerical value can be found. For instance, Hartmann et al. are working on a mathematical transformation and mapping in this paradigm [Hartmann et al. 2012]. In contrast, an empirical approach towards a suitable mapping is given in [Hoffmann et al. 2012].

Continuous Labels

The most abstract form of annotation is the continuous labelling. In this, no expressions like *anger*, *shame*, etc. or predefined classes are given. Rather full continuous numerical values in all dimensions will be assigned to any kind of human behaviour reflecting the emotional state. Therefore, not only a *time continuous* annotation, as it is used in categorical and quasi-continuous labelling, but also a *continuous labelling in human's reactions* is conducted. Nevertheless, usually a predefined range of the values is given which reflect the maxima. Such labelling can be done using, for instance, the tool FEELTRACE [Cowie et al. 2000].

A schema for continuous labelling is Self-Assessment Manikin (SAM) [Bradley & Lang 1994] which is based on the Pleasure-Arousal-Dominance (PAD) space introduced in [Mehrabian 1996]. According to this, reactions are arranged within a three-dimensional space. Along each dimension the influence of a situation towards the user is measured. Therefore, eight interesting clusters can be generated representing the different kinds of situations, for instance, low dominance, low arousal, and high pleasure. I mention that sometime the *Pleasure* dimension is substituted by *Valence* which is more or less due to the customs of some authors. In this thesis, I use the PAD space as introduced by Mehrabian.

Of course, the PAD is not the only way of measuring emotional states in a continuous manner – it depends more on the matter of observation–, however it is the most popular one. In my experiments I also applied the PAD space, to be more specific a subset (cf. Chapter 5), as they are related to the EmoRec I+II (EmoRec) corpus [Walter et al. 2011] (cf. Section 3.2.2). Further, this space is utilised in ikannotate applying the SAM rating given by [Bradley & Lang 1994; Grimm et al. 2007].

In general, continuous labels have the capability to reflect dispositions (cf. Definition 1.6 on page 7) as, on the one hand, they have no fixed classification. On the other hand, which is more important, they directly incorporate the situation and environment of an interactor by dimensions which reflect characteristics influencing the user. Especially, in the dimension of *Dominance* the surrounding of the person is part of the measure as the user is either controlling, or controlled by the situation. As Batliner et al. argue for more natural corpora, I argue heavily for continuous labels to handle disposition in a proper way. For the purpose of usage in HMI control I refer to Section 1.5 and also for further development to Section 7.2.5).

1.3 Inter- vs. Intraindividual Validation

From the general reflections of emotions and disposition as well as the different labelling paradigms, I will consider the types of recognition validations which are of interest in HHI, but rather, in HMI. Namely, I distinguish inter- and intraindividual validation approaches which are connected with the well-known concepts of cross-validation and Leave-One-Speaker-Group-Out (LOSGO) or Leave-One-Speaker-Out (LOSO). I will use the former names of the methods, because they represent a more general intention of the validation concept. This means, the more technical term cross-validation and LOSGO/LOSO are, from my point of view, too focused on the engineer's view. But, disposition as such (cf. Definition 1.6 on page 7) which comes from situatedness (cf. Definition 1.5 on page 7) is broader, especially, considering the surrounding of an interaction. Therefore, I introduce the corresponding terms (cf. Section 1.3.2 and 1.3.3) interpreted in a technical sense. Further, to evaluate the findings, suitable measures of accuracy have to be defined.

1.3.1 Measures

At first, I declare the measures which will be used throughout this thesis to evaluate the performance of the experiments. This is related to the validation method as well as the structure of the material itself. I concentrate on common performance measures as known from speech recognition, hence, the focus is on the material. As discussed in more detail in Chapter 3, data sets are either balanced or unbalanced in the number of samples per class[1]. This is reflected by the following two measures.

The *Weighted Average accuracy (WA)* is the measure which reflects the class-wise accuracy. Therefore, the accuracy is calculated for each class separately and afterwards averaged over all classes. Moreover, this is the reason why WA is more accurate and furthermore, applied in the context of unbalanced data sets.

Let x_i be the number of correctly classified samples, y_i be the number of all samples in class i, and c be the number of all classes, then WA is calculated by

$$\mathrm{WA} = \frac{\sum_{i=1}^{c} \frac{x_i}{y_i}}{c} \tag{1.1}$$

In contrast, *Unweighted Average accuracy (UA)* is based only on the samples correctly classified without regarding the classes. This means, it is the ratio of correctly classified samples and the number of all available samples. UA is thus computed by

$$\mathrm{UA} = \frac{\sum_{i=1}^{c} x_i}{\sum_{j=1}^{c} y_j} \tag{1.2}$$

[1] The term class has to be considered more generally, not only in the sense of Section 1.2.3.

where x_i is the number of correctly classified samples per class i and y_j is the number of all samples in the class j. Therefore, $\sum_j y_j$ is the total number of samples in the material summed over all classes and $\sum_i x_i$ is the total amount of correctly classified samples over all classes. It is usually applied if the data set is balanced in the number of samples per class which also means that WA and UA are more or less equal.
A roughly similar definition of the measures is, for instance, given in [Olson & Delen 2008].

In the following I will discuss characteristics of these two accuracy measures with respect to their relationship. In general, both measures are linked with each other: will be one increased (e.g. UA), the other one is decreased (e.g. WA) and the other way around. In fact, if a given data set is balanced in the sense of sample distribution; that means, the number of samples per class c_i is almost equally distributed over all classes, UA and WA are roughly equal in the absolute value. Usually, corpora are unbalanced with respect to the samples' distribution (cf. e.g. description of data sets in Chapter 3) and thus, both measures are different in their values. In the case of unbalanced data sets a high UA value can be achieved if the class with the most samples is classified in a good way. On the other hand, classifying the class with the fewest samples results in a high WA. This effect can also be seen in the experimental results presented in Section 5.1.2. Both described characteristics can be derived directly from the corresponding equations (cf. Equation 1.1 and 1.2). Given these considerations, especially two case in the measures' relation are important and hence, will be further discussed.

- UA > WA which is the preferable situation. In this case, WA ranks the performance of the classifier better because the effect of a well recognised class with a large amount of samples is decreased. Thus, the influence of all classes is reflected properly.
- UA < WA and thus, the class with the fewest number of samples dominates the classification performance. In general, this should be avoided for a proper ranking of the corresponding classification results.

To overcome this problem in statistics and other research communities like information retrieval and artificial learning several other measures were introduced (for instance in [Dietterich 1998; Makhoul et al. 1999]). Especially in speech recognition, accuracy measures like *F-Measures*, *Precision*, *Recall*, *Likelihood Ratio* etc. are known for the evaluation of results (cf. [Jiang 2005]). These were also transferred to the evaluation of classifiers dealing with emotion recognition from speech (cf. [Grimm et al. 2007; Grimm et al. 2008; Vlasenko & Wendemuth 2009; Schuller et al. 2010a]) providing also their own characteristics. As I do not use such additional confidence measures in the my experiments, I do not introduce these.

1.3.2 Interindividual Validation

Every person generates a model of his counterpart to react on the other's behaviour, hopefully proper. Unfortunately, in the beginning of each interaction, usually no information of the counterpart is given. Therefore, at first, each participant has to rely on a model and hypothesis derived from it which reflects a general behaviour in a particular situation. That means, from several previous interactions a prototypical model is established to indicate a possible course of interaction. This holds for HHI as well as for HMI and further, it should be true for technical systems like described in [Wendemuth & Biundo 2012] that react like companions with a user.

Based on these considerations, a validation method has to be defined which deals with the requirements of unseen users. It means that any kind of system in an HMI can only rely on material which was previously 'processed' dependent on previous interactions. This is the technical complement to an universal model or strategy used by a human.

In speech recognition such validation method is call LOSO which means that material of one user is not used in the training process and only presented to the system in the test. An extension of LOSO is LOSGO applying the procedure to a group of speakers.

I redefine this validation strategy to an interindividual validation method. As in the standard approach the material of one user is left out for the training and is used in the test, only. But, while usually the average of the speakers' characteristics is learned, in the advanced HMI (cf. e.g. [Wendemuth & Biundo 2012; Böck et al. 2012b]) personality is an additional and important feature. Therefore, I will use this term which is inspired and based on psychology. The difference in the LOSO and interindividual validation will be more clear in the context of intraindividual validation (cf. Section 1.3.3).

1.3.3 Intraindividual Validation

In HHI it is common that after a small period of introduction the interactors adapted to each other. This is related to a switch from an universal model (cf. Section 1.3.2) of the counterpart towards an individual model reflecting the significant characteristics of the partner. Hence, while interacting the validation of the partner's behaviour is adapted, thus narrowed, and consequently improved as it is based on the observations of a single person.

This can be transferred to a technical system as well by using only material which was collected from one user. Therefore, the training as well as the testing is based on this specific person and reflects his characteristics, especially, the personality. Thus, I will use the term intraindividual validation to indicate the particular validation method. Especially, in long term interactions as intended in [Wendemuth & Biundo 2012], an intraindividual validation is to favour.

Intraindividual validation is carried out as follows: the collected material of a speaker is arbitrarily split into two sets. The training set, mainly 90% of the material, is used to train or adapt a classifier. The remaining data is afterwards presented in the test of the system. With Artificial Neural Networks (ANNs), for instance, the method slightly changes as an additional validation set is needed, used for parameter tuning of the ANNs. Hence, the splitting is normally 80% training, 10% validation, and 10% test. To get statistical significant results the splitting is repeated ten times and afterwards the results of each run are averaged. An extension of this procedure is the stratified intraindividual validation transferred from the principles of stratified cross-validation. For this the splitting is done in such a way that in the test set the overall distribution of samples to the classes given in the data set are represented (cf. e.g. [Diamantidis et al. 2000]). This means that an equal portion of samples from each class are added to the test set. Especially in unbalanced corpora in terms of samples' number per class, this approach provides a better ratio between training and test sets. Usually, in my experiments I applied the stratified intraindividual validation; referring to it only as intraindividual validation.

Another approach which is related to intraindividual validation is the so-called cross-validation usually applied in speech recognition. In contrast to intraindividual validation the material of all speakers is used to generate training and test sets. The procedure to get these sets is the same as for intraindividual validation and further, the stratified cross-validation (cf. e.g. [Diamantidis et al. 2000]) can be applied as introduced before. The advantage of this method is that certain side-effects like age and gender can be reduced or even eliminated and a result which is based on a general observation is achieved. Especially in explorative experiments this can help to get indications for suitable feature sets and classifier parameters. Indeed, in some cases where only a small amount of material per speaker is available, this method might be better to assess a technical system as the result is therefore based on more material.

1.3.4 Connection of Both Validation Systems

Both methods, as introduced in Section 1.3.2 and 1.3.3, are based on the validation approach commonly used in speech recognition. The main difference is that I concentrate on the personality or the individual behaviour as a criterion in the validation of classification results. Moreover, intraindividual validation narrows the assessment to a single speaker reflecting the idea of a technical system which is adapted to one user and for this, can be seen as a step towards companion systems (cf. [Wendemuth & Biundo 2012]).

Especially systems having a long term interaction with a specific user should switch the validation methods. As discussed in [Böck et al. 2012b], in the beginning of an interaction the system should rely on a general model of any kind of user. To assess its performance the interindividual validation is applied while interacting data material is recorded and can be used to adapt the system (a framework to do so is proposed in [Böck et al. 2012a]). Therefore, a switch in the validation approach towards intraindividual validation has to be done. Hence, the methods are on the one hand complementary but in case of technical companion systems they are successors in the process of evaluation of classification results. Nevertheless, in this thesis I will usually compare the classification results according to both validation methods.

1.4 Involvement in Conversation

So far, I introduced emotions and more importantly, disposition in Section 1.2 and further discussed the differences in the validation methods (cf. Section 1.3) which are influenced by the personality of a user as an additional perspective. In this Section I give a description of an aspect which directly evolves from the disposition; namely involvement. As Oertel et al. state important characteristics that make “a conversation a naturally interactive dialogue are the dynamic changes involved in spoken interaction." [Oertel et al. 2011b]. In contrast, in [Antil 1984] the involvement is discussed in a much broader sense which influences the behaviour of human begins. Therefore, I consider the involvement under the focus of being a disposition and derive a definition (cf. Definition 1.7) that is more general as, for instance, in [Oertel et al. 2011a] which is more fixed on task-oriented conversations.

Definition 1.7 *Involvement reflects that somebody is generally participating in something. In particular, in the context of speech, this means, a user is participating in a conversation or interaction.*

As in Definition 1.7, in particular, I understand by *involvement* that a speaker is participating in a conversation or interaction whereupon it does not matter whether this is an HHI or an HMI. In the definition of disposition (cf. Definition 1.6 on page 7) the situatedness and further information like the background are mentioned. Being involved in an interaction is a crucial point in the analysis of the user's behaviour. Otherwise, a misinterpretation of actions is most likely.

Before discussing the influence of the involvement as a characteristic of communication, I will consider the issue: what are the characteristics of involvement? Being involved or participating in something represents a large variation of behaviour. Usually this is investigated by psychologists who can interpret small differences in human behaviour to derive conclusions for the course of an interaction. In general, it reflects the behaviour of a person interacting with a communication partner. It does not matter whether this is another person or a technical system. Participating is manifested in gestures, facial expressions, and speech. For example, a person gets more involved in a conversation if he leans forward and starts talking. As from this example, it is obvious that the question of involvement is an issue which can be faced only with multiple modalities, that is, using not only speech but also video features, etc. to detect and classify participation. The methods for this and how to perform involvement recognition will be discussed in Section 4.4 and Chapter 6.

As I said in Section 1.1 it is important to detect involvement, especially in multi-party conversations. Multi-party means that at least three subjects are participating in an interaction where each subject is either a human or a technical system. In such scenarios (cf. Section 3.3), the detection of involvement indicates which participants have to be considered as being part of the interaction. From the point of view of a technical system, that means to distinguish which subjects have to be considered. As in the HHI the attention is focused on subjects who are of interest and then their reactions are analysed. Further, information on how to react on disturbances, for example, are retrieved from it [Novick et al. 2012]. In this way, a technical system can derive information to react to specific situations properly, for instance, who is talking to whom. To be specific, even in a dyadic (only two subjects interact) conversation one party can be out of the interaction and therefore is no longer involved. However, in this thesis I just consider multi-party interactions and hence, assume that in two-subjects conversation both partners are participating.

Multi-party or group interactions are usually analysed by psychologists (cf. e.g. [Wageman 1995; Keyton & Beck 2009]) observing the dynamics, which is the behaviour, of the group and the conversation itself. From this, several sets of conversations, like meetings or teaching scenarios, can be derived. A brief overview is given in [Böck et al. 2013b]. From the technical system's perspective it is more important to distinguish different constellations of an interaction and involvement. As discussed in [Kühn & Koschel 2010] these are: i) involvement of each participant, ii) dyadic or sub-group interactions, and iii) involvement of the group in total. Again, I will concentrate in this thesis on the latter case, only. The way such observations can influence the HMI is discussed in Section 1.5 and Chapter 6.

1.5 Control of Human-Machine Interaction

In Section 1.1 I already mentioned the overall goal of automatic disposition classification, namely improving the system's capabilities to recognise the user's disposition to react in a proper way. This is essential to avoid any kind of breakdown in the course of the interaction or, in particular, the dialogue. As I am not interested in the global design of interfaces or systems, I will discuss the way of controlling a system not in detail.

In the beginning of HMI, it was common to submit commands by mouse and keyboard to a system and the interfaces were designed in a minimalistic manner (cf. [Carroll 2013]). As it is known, especially the design of interfaces has changed dramatically towards more fancy styles and appearances (cf. [Rogers et al. 2011; Carroll 2013]). Nevertheless, the way of how to give instructions did not change as much. I do not believe that this style of interaction will be detached totally by new technologies, but nowadays the way of interaction is much more manifold. New technologies were developed, for instance, touch screens, and methods were implemented like speech and gesture recognition, so that the controlling of a technical system is not only manifold but also multimodal (for the purpose of controlling cf. [Panning et al. 2012; Krell et al. 2013]; for the purpose of gestures cf. [Appenrodt et al. 2010]). This is also due to the development of the technical capabilities like small cameras, high quality but small microphones, improvement of the computational power in small devices, etc.

On the other hand, controlling a system means that the user needs a concept how to control it. Even with speech, gesture, or facial expression recognition, control is mostly not as intuitive and natural as it should be. This is due to the predefined signs and gestures to get a reaction of the system or the vocabulary of the speech recognition which is reduced to a specific domain. All in all, the user is asked to adapt to the system. The goal is to change this which means the system is adapting to the user (cf. [Wendemuth & Biundo 2012]) and not the other way around. As in Chapter 2, especially in speech recognition universal models and several adaptation methods are applied to obtain systems which are more general in their reactions. This is also reflected in the output of technical devices which becomes multimodal, too, (cf. [Jaimes & Sebe 2007]). An outlook concerning the power of multimodal recognition which also leads to an appropriate system's reaction that further could influences the user's way of interaction is presented in Section 7.2.5.

So far, the process of controlling is a task which assumes several conditions of the user. First of all, willingness and good will to interact. Further, an idea how the system works in general, that is, what is the purpose and what can the user get from it. And finally, the user adapts to the capabilities of the technical system. To overcome the latter case, I will present results and approaches in this thesis to enable systems which interact more natural and human-like as it is discussed in [Wendemuth & Biundo 2012] and which finally are integrated into human life as companions.

Chapter 2

State of the Art

Contents

2.1 Human-Machine Interaction 22

2.2 Reviewing Disposition Recognition from Speech 25

2.3 Data Sets 28

2.3.1 Dispositional Data Sets 28

2.3.2 Data Sets of Conversations 30

2.4 Methods of Disposition Recognition from Speech 36

2.4.1 Feature Sets for Disposition Recognition 36

2.4.2 Classification Methods for Disposition Recognition 38

2.5 Automatic Analysis of Involvement 42

2.6 Summary 45

SO far, I introduced the general idea of this thesis, namely the automatic detection of dispositions from speech. For this, the main concepts and definitions were given and discussed, especially, in a technical sense. I stated that the term *disposition* subsumes various aspects of a user as well as his behaviour in an interaction. In particular, it reflects the emotions, moods, and situations of a certain user. Furthermore, from these considerations I discussed the detection of involvement and its changes in a conversation. Using these information a user-centred HMI and further, a system influenced HMI could be established.

In the following, the state of the art in disposition recognition from speech will be presented which is used to position my research in the community. It is obvious that this can only be a spotlight of the research activities over the past due to the enormous amount of publications. Further, a few communities like speech recognition or classifier development are already established for a long time; that means, for 40 years or even longer. Thus, I mainly concentrate on issues that are in the context of my research and present only particular highlights, being aware that covering the overall aspects of the community is not possible in this thesis due to the multitude of related topics. Furthermore, a brief introduction on HMI will be given as my research is integrated in the context of it.

2.1 Human-Machine Interaction

In Section 1.5, I discussed and motivated the importance of HMI and further, described roughly how this can influence a conversation. In Section 2.1 and Section 6.3.2, the influencing as well as the controlling of conversation from a system's point of view will be considered. Therefore, a brief overview of HMI and its state of the art is necessary to position the discussed aspects properly. The automatic emotion recognition and further, the automatic disposition recognition can be seen as an important part of a proper HMI. Therefore, I review the developments in HMI focusing on the main aspects which are related to automatic disposition recognition.

Carroll presents in [Carroll 2013] a historical overview on HMI. The community gained importance mainly in the late 1970s when the Personal Computer (PC) became popular. Before, the usage of computers was focused to researchers and enthusiastic hobbyists who were interested in the functionality and the calculation power of these machines. But with the PC the focus shifted to productivity applications and interactive games (cf. [Carroll 2013]). The research community on HMI developed in this context. In the following decades, the style of software and thus, interfaces of PCs, which are an integral part of HMI, changed from an "Unix-like" appearance – which is command based – to an "Apple Macintosh style" – being mainly graphical based – [Carroll 2013]. In the mid 1980s E-Mail and network services were introduced which started novel communication possibilities; "people were not interacting *with* computers, they were interacting with other people *through* computers" [Carroll 2013]. Hence, new communication paradigms and mechanisms arose. This was and is linked with a continuous development of interfaces and devices (cf. [Shneiderman & Plaisant 2010; Rogers et al. 2011; Carroll 2013]). Moreover, this is also set in relation to affective computing (cf. [Picard 2000; Höök 2013]). As it is stated in [Carroll 2013] the HMI community is indeed a community which is connected to various other research disciplines (cf. [Carroll 2013]). Mainly, it is linked with computer sciences, but also with information technology, psychology, design, cognitive sciences, etc. From these, studies on HMI are greatly influenced (cf. [Shneiderman & Plaisant

2010; Carroll 2013]). Nevertheless, most of these studies deal with the aspects of the functionality of a system and its usability (for detailed information of these terms cf. [Shneiderman & Plaisant 2010]). These issues were widely investigated over the past decades.

A more technically inspired review on the state of the art in HMI is given in [Karray et al. 2008]. The authors distinguish the approaches according to the characteristics of the interfaces used in the controlling of systems. They mention i) switch based devices using buttons and keyboard, ii) pointing devices applying touch screens, trackballs, graphic tablets, etc., iii) audio based systems utilising speech recognition, and iv) haptic devices used for motion detection in virtual reality, disability assistive tools, etc. Moreover, the HMI systems can be distinguished according to the modalities applied in the approach, namely visual, audio, sensor, and multimodal HMI (cf. [Karray et al. 2008]). In [Karray et al. 2008] furthermore, methods to equip devices with several kinds of functionality are discussed.

Especially, the multimodal based HMI shows promising approaches as stated in [Jaimes & Sebe 2007].

> Multimodal interfaces have been shown to have many advantages [...]: they prevent errors, bring robustness to the interface, help the user to correct errors or recover from them more easily, bring more bandwidth to the communication, and add alternative communication methods to different situations and environments. [Jaimes & Sebe 2007]

This quotation reflects the potency of multimodal oriented interfaces and further, of multimodal based HMI. By doing so, both the user and the system react on multiple modalities like vision, audio, etc. Moreover, for the observation of the user also biophysiological measurements as heart rate, skin conductance, etc. are introduced as an additional modality. For this, the authors state that the HMI will become more human-centred since the way of interaction is getting more natural. Such multimodal observations provide the opportunity to investigate also affective states of a user (cf. [Jaimes & Sebe 2007; Höök 2013]). From my point of view, it is thus just a small step towards a dispositional observation of the user in the HMI.

In general, the community of HMI was and is an emerging field of research. This is also reflected by the numerous survey and review articles cited in [Jaimes & Sebe 2007; Lew et al. 2007; Karray et al. 2008; Ratzka 2010; Shneiderman & Plaisant 2010; Rogers et al. 2011]. Furthermore, this can be derived from the historical evolution as well, which was already discussed and is described in more detail in [Carroll 2013]. Moreover, the various kinds of applications which are highly linked to HMI indicate the importance of this research. In the following a brief collection of topics and applications is given, referring to [Jaimes & Sebe 2007; Karray et al. 2008]:

- portable and wearable devices like smartphones, navigation devices, bracelets recording biophysiological features (cf. [Höök 2013]), or glasses providing information (cf. [Carroll 2013])
- virtual environments
- virtual agents (e.g. cf. [Traum et al. 2012]) and robots (e.g. cf. [Merten et al. 2012])
- ubiquitous computing and ambient spaces including ambient information (additionally cf. [Elsholz et al. 2009]) and intelligent homes and offices (additionally cf. [Kameas et al. 2009])
- support of elderly people (e.g. cf. [Merten et al. 2012]) or disabled users
- medical support
- education, games, entertainment, and arts.

Especially in the context of virtual agents and ubiquitous computing, the goal is an intelligent HMI. In [Lew et al. 2007] a research perspective is drawn, stating that the community should be "interested in intelligent interaction where the computer understands the meaning of the message of the user and also the context of the message" (cf. [Lew et al. 2007]). For this, the HMI have to evolve from a "good old HMI" [Müller 2011] towards a novel approach integrating cognitive aspects. Therefore, in [Müller 2011] three statements towards such intelligent HMI are discussed, namely

1. "[HMI] research is a special case of cognitive systems research."
2. "Successful interaction requires resistance to the other agent."
3. "Resistance is essential even for non-cognitive [HMI] systems."

This can be seen as a kind of roadmap to investigate HMI systems under novel aspects. Such analyses are also part of the research community dealing with companions (cf. [Wilks 2005; Wilks 2006]) and companion technologies (cf. [Wendemuth & Biundo 2012]). The aim is to consider an HMI incorporating aspects of affective computing (cf. [Picard 2000]) and also emotion as well as disposition recognition to enable systems with companion-like characteristics. This is amongst others the goal of the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (SFB/TRR 62). The ideas are described in [Wendemuth & Biundo 2012]. In this thesis, I deal with one specific aspect, namely the disposition recognition from speech, that can be integrated in the context of HMI to enable systems to assess their user more properly. As this is the main idea of the thesis, I will consider the current approaches in the disposition recognition from speech in the following.

2.2 Reviewing Disposition Recognition from Speech

The disposition recognition from speech is based on the ideas and methods from speech recognition and emotion recognition from speech. In this Section, I briefly sketch the evolution of the research communities culminating in disposition recognition. The developments in the community are also described in several survey papers which will be referenced in the following.

In [Anusuya & Katti 2009] a detailed review of the developments in speech recognition is given. The authors state that the starting point of automatic analyses of speech is in the early 1920s with a toy called 'Radio Rex'. This commercial system reacted on a spring released by a frequency of 500Hz which is roughly related to the first formant of the vowel [e] in 'Rex' (cf. Table 4.1 on page 75). During the following years and decades, automatic speech recognition systems were improved and thus, quite significant results could be achieved (cf. Section 6 in [Anusuya & Katti 2009]).

Inspired by speech recognition and its methods, in the 1990s emotion recognition from speech was developed (cf. [Dellaert et al. 1996; Schuller et al. 2011c]). This is also related to the foundation of the *affective computing* (cf. [Picard 2000]) and its research community (an overview is given in [Höök 2013]). The very first beginning of the computational emotional speech analysis is marked by the work discussed in [Bezooijen 1984; Tolkmitt & Scherer 1986] where certain acoustic features were investigated using statistical methods. The authors of [Ververidis & Kotropoulos 2006] and [Iliou & Anagnostopoulos 2010] give an overview of the developments in the field of emotion recognition from speech. The psychological tradition of technical emotion recognition is reviewed in [Fragopanagos & Taylor 2005] relating the research in automatic emotion recognition also to theories of emotions as established in psychology. Moreover, in [Schuller et al. 2011c] basic ideas of the community which were pursued over the years in emotion recognition from speech are presented. Additionally, [Schuller et al. 2011c] reflects the work related to investigations of more naturalistic HMI in the sense of speech recognition, for example, analysing realistic emotions, utilising naturalistic data sets, etc. These ideas are based on position papers of, for instance, Batliner (cf. [Batliner et al. 2000; Batliner et al. 2001]). And indeed, while reviewing the literature, a shift in the analyses from acted emotions towards naturalistic emotions and dispositions can be seen (cf. [He et al. 2009; Zeng et al. 2009; Schuller et al. 2010a; Vlasenko et al. 2011a; Schuller et al. 2011c; Planet & Iriondo 2012]).

The issue of investigating more naturalistic emotions is also reflected by the emotion recognition challenges arranged since 2009. Moreover, these challenges show the development which was conducted in the community to improve the recognition of emotions and dispositions from speech. In 2009, the first emotion challenge was organised in conjunction with the Interspeech-2009 (cf. [Schuller et al. 2009c]). As a start, categorical emotion recognition was implemented as the challenge's task using the AIBO corpus (cf. [Batliner et al. 2008]) as data set. General issues that could be learnt from this challenge are presented in [Schuller et al. 2011c]. Three main aspects were seen as a kind of conclusion (cf. Section 3.6 in [Schuller et al. 2011c]): i) most of the participants used Mel-Frequency Cepstral Coefficients (MFCC) as the main features for their recognisers, ii) the investigation of other features might be worthwhile, and iii) the recognition of emotions from speech is a complex task and thus, the community might benefit from cooperations with other disciplines like video processing, psychology, etc. From these considerations two approaches were pursued, namely conducting emotion recognition challenges with broader senses of investigations and generating the Audio/Visual Emotion Challenge (AVEC). In terms of emotion recognition challenges these are: the "Interspeech paralinguistic challenge" analysing age, gender, and affect at Interspeech-2010 (cf. [Schuller et al. 2010b]), the "Interspeech speaker state challenge" at Interspeech-2011 where the speaker states under different situations like fatigue, alcoholic influence, etc. are considered (cf. [Schuller et al. 2011d]), the "Interspeech speaker trait challenge" at Interspeech-2012 focusing on the personality of a speaker as well as likeability (cf. [Schuller et al. 2012]), and finally, the "Interspeech computational paralinguistic challenge" at Interspeech-2013 with its subtitle 'Social Signals, Conflict, Emotion, Autism' (cf. [Schuller et al. 2013]). Regarding the topics of these challenges, it is to notice that there is a continuous evolution in emotion recognition from speech towards research which analyses issues related to more naturalistic situations, culminating in aspects of social signals and diseases like autism. Therefore, the topics of the challenges reflect also the cutting edges of the research community.

The other important lesson learnt from the first emotion recognition challenge is that the recognition of emotions is a complex task, in general. Thus, the emotion of a human or, more specifically, a user should be observed with more modalities as this provides the possibility of a comprehensive user observation. Since this thesis deals mainly with emotion and disposition from speech, I describe the state of the art in multimodal analyses just briefly. A survey of recognition methods and findings according to audio-visual emotion analyses is given in [Zeng et al. 2009]. Additionally, [Bousmalis et al. 2013] reflects the importance of gestures and head movements in nonverbal behaviour. To establish a kind of benchmarking of audio-visual feature based classifiers the AVEC corpus was founded in 2011 (cf. [Schuller et al. 2011a]). This challenge provides sub-challenges for audio, video, and audio-visual analyses of naturalistic emotional material since the SEMAINE corpus (cf. [McKeown et al. 2012]) was used. In 2012, the challenge's task was amongst others extended to a continuous emotion recognition and in 2013 additionally, the detection of depression levels is a sub-challenge.
On the other hand, biophysiological features, and classifiers based on those, support a comprehensive observation of a user (cf. [Nasoz et al. 2003; Kim & André 2008; Walter et al. 2013]). Especially, in the context of establishing a kind of ground truth in the identification of emotional behaviour, biophysiological features supply evidences since those signals are hard to control consciously. Therefore, they are called 'honest signals' in biophysiology. This issue will be also discussed in this thesis in Section 5.2.3.

In general, it can be seen that emotion recognition is an emerging domain of research. It is the same with disposition recognition as it is based on emotion recognition. Since disposition recognition and, in particular, the disposition recognition from speech is a quite novel approach, to the best of my knowledge, the literature is not as detailed as for emotion recognition. From the definition of disposition (cf. Definition 1.6 on page 7), the work done in emotion recognition can be included in the disposition recognition. On the other hand, the PhD thesis of Stefan Scherer laid foundations towards a conversational disposition recognition (cf. [Scherer 2011]). In this context, my thesis is aiming to incorporate the disposition recognition from speech in a naturalistic HMI. Currently, various approaches are pursued in this direction; for instance, the modelling of user's moods (cf. [Siegert et al. 2012a]) or describing human's emotions mathematically (cf. [Hartmann et al. 2012]). In general, the issue of disposition recognition is a comprehensive topic of future research that will influence the HMI strongly in next years, especially, in the sense of generating automatic systems that can be seen as companions (cf. [Wendemuth & Biundo 2012]).

2.3 Data Sets

Usually, systems for speech recognition as well as disposition recognition from speech are based on classifiers. These have to be trained to recognise or classify given utterances in a proper way. The training is based on data sets which are suitable to cover the characteristics that should be recognised. In the following, I briefly review the corpora that can be used for the generation of disposition recognition systems. Furthermore, in Section 2.3.2 data sets in the context of conversation analyses are considered.

2.3.1 Dispositional Data Sets

In general, I distinguish two types of data sets: i) corpora which are only suitable to train speech recognition systems and ii) data sets that also supply dispositional information. The second category's corpora could be used for training of more naturalistic speech recognition systems as well. They provide, for instance in Last Minute or EmoRec, speech, recorded under naturalistic conditions; that means, afflicted with accents, slurring of words, etc.

Speech Recognition

Corpora that are only used in speech recognition were mainly recorded or generated in the 1990s, for instance, Resource Management and TIMIT in 1993, Polycost 250 in 1996, etc. Such data sets were usually recorded under controlled conditions; that means, in most cases laboratory conditions with almost no noise or even anechoic chambers were used to obtain clear speech samples. In [Anusuya & Katti 2009] the authors present a list of data sets which are commonly used in the speech recognition community. Related to speech recognition is the research field of speaker identification and verification from speech. For this, speech samples are applied to extract information that are connected to characteristics of a certain user. For classification purposes speech recognition data sets can be utilised as well. Moreover, it is also possible to apply corpora generated in the context of speaker recognition in speech recognition. For this purpose, in [Melin 1999] an extensive list of corpora is given.

In this thesis, the recognition and classification of dispositions from speech is in the focus. Therefore, the pure speech recognition data sets were considered here briefly and emotional or dispositional corpora are investigated more deeply.

Disposition Recognition

Table 2.1: Overview of data sets which provide dispositional characteristics. Besides the name of the corpus and the reference where to find a description, the type of collection is given. For this, I distinguish acted (act) and non-acted (n-act) corpora (cf. Chapter 3). Furthermore, for each corpus it is marked if it can be used for disposition recognition as well (Disp. rec.). This listing is adopted from [Vlasenko 2011] and actualised.

Corpus	**Reference**	**Type**	**Disp. rec.**
DES	[Engbert & Hansen 1996]	act	–
SUSAS	[Hansen & Bou-Ghazale 1997]	act/n-act	x
EmoDB	[Burkhardt et al. 2005]	act	–
eNTERFACE	[Martin et al. 2006]	act/n-act	–
SmartKom	[Wahlster 2006]	n-act	x
ABC	[Schuller et al. 2007a]	n-act	x
AVIC	[Schuller et al. 2007b]	n-act	x
AIBO	[Batliner et al. 2008]	n-act	x
VAM	[Grimm et al. 2008]	n-act	x
EmoRec	[Walter et al. 2011]	n-act	x
SAL	[McKeown et al. 2012]	n-act	x
Last Minute	[Frommer et al. 2012b]	n-act	x

Based on the methods and data sets of speech recognition, corpora which contain emotional and dispositional characteristics were generated. In the late 1990s this started with, for example, the Danish Emotional Speech database (DES) (cf. [Engbert & Hansen 1996]). In [Ververidis & Kotropoulos 2006], in particular in Table 1, an exhaustive listing of available data sets related to emotion recognition is given. In addition, [Schuller et al. 2009b; Schuller et al. 2010a; Vlasenko 2011] review the existing emotional speech data sets. Anagnostopoulos et al. present also an estimation of emotion's distribution in various data sets, stating that the most appearing emotion in corpora is *anger*, within approximately 85% of the data sets (cf. [Anagnostopoulos et al. 2012]). In Table 2.1 on the previous page, I present a collection of the mainly used corpora in the community. This selection contains also data sets which are up-to-date in the sense of disposition recognition, namely EmoRec (cf. [Walter et al. 2011]), Last Minute (cf. [Frommer et al. 2012b]), and Sensitive Artificial Listener (SAL) (cf. [McKeown et al. 2012]). The mentioned corpora also reflect the aspect that currently most data sets were generated using multiple modalities to cover the speakers' overall reactions. This aspect is also introduced in [Zeng et al. 2009] where the authors furthermore present data sets which can be analysed in audio-visual investigations. Moreover, they describe why a multimodal user observation is reasonable. Mainly, it is due to the different occurrences of dispositions in various modalities. From the selection of data sets (cf. Table 2.1 on the preceding page) I used four corpora in my experiments, namely the Berlin Emotional Speech Database (EmoDB), eNTERFACE'05 (eNTERFACE), Last Minute, and EmoRec. Those will be presented in Chapter 3 in more details.

Highly related to the issue of data sets is the aspect of naturalistic dispositions included in these. As discussed in [Batliner et al. 2000; Batliner et al. 2001] and Section 3.2, naturalistic corpora should be in the focus of analyses and thus, such kind of material should be generated with priority. Naturalistic means that the dispositions are not played by an actor but are obtained in a kind of a non-acted scenario. The community considered the remarks of Batliner et al. in [Batliner et al. 2001] and reacted properly by recording naturalistic corpora. As it can be seen from Table 2.1 on the previous page, especially, after 2006 naturalistic data sets have been generated. Usually, the corpora applied in disposition recognition reflect HMIs where just one user is participating in this interaction. In contrast, for investigations of involvement (cf. Section 1.4) multiparty corpora are necessary. An overview of those will be given in the following Section.

2.3.2 Data Sets of Conversations

In the previous Section, I considered the data sets which are commonly used in disposition recognition from speech. Therefore, I gave only a brief overview. In contrast, corpora which are suitable to investigate aspects of conversations, especially, group conversations are not as widespread, yet. Hence, such data sets are described in more detail. In this context, two kinds of corpora can be distinguished: i) data sets recorded for a certain purpose and ii) data sets supplied for the whole community. As this thesis is focused on technical aspects of conversations, in particular, of involvement, I neglect corpora which are analysed in psychological studies and concentrate on those that are already applied for technically inspired issues.

Individually Generated Data Sets

In several cases, the generation of a data set is linked to a certain task or a current issue of investigation. Therefore, corpora are recorded that are not widespread in the sense of being given to a distributor like ELRA or LDC, but available from the collector on request. I will call these data sets individually generated corpora because they were collected for a certain, individual purpose. Besides task depended reasons, such data sets are not commonly distributed since they reflect only a limited aspect of research or highly depend on a language, for instance, and thus, are not in the scope for many researchers.

Dillon recorded three participants who listened to music of Johann Sebastian Bach (cf. [Dillon 2005]). The issue was to evoke an emotional engagement while listening to the music. For this, the participant's reaction is related to the characteristics of the music pieces like its tempo, articulation, sound level, etc. The whole setup is highly related to musical analyses.

In relation to the engagement, in [Gustafson & Neiberg 2010] a data set is presented that covers dyadic conversations between an entertainer of a radio show and his listeners. Both partners discuss via telephone about various topics linked to the show. During the interaction, the entertainer attempts to keep the caller engaged in the conversation. As this is a telephone-based setting only audio recordings are available. The corpus contains 73 calls of Swedish conversation. The main issue which was analysed on the data set is, how humans prosodically align each other during a telephone call tracking also the engagement of the partners.

Analysing engagement and involvement – for this, the community is working on distributed data sets (cf. next Section) – culminates in providing socio-feedback (cf. [Rasheed et al. 2013]). For this purpose, in [Rasheed et al. 2013] a real-time system is developed that investigates acoustic cues in dyadic HHIs. To train the classifiers, a corpus of 42 face-to-face conversations is recorded which focuses on issues of team building. The data is collected with a lapel microphone for each participant and the recording is afterwards automatically classified according to levels of involvement, dominance, and impedance.

The aspects of engagement and socio-feedback are indeed related to involvement but are not in the focus of this thesis. Therefore, I will guide the attention towards data sets which reflect involvement. These are introduced in the following Section.

Distributed Data Sets

So far, I considered data sets which were recorded by the authors of the corresponding papers reflecting their research needs and interests. Furthermore, the corpora are mainly focused on engagement that is not in the focus of this thesis. Hence, I will introduce in the following, corpora which can additionally be utilised to study aspects of involvement. Furthermore, they are also distributed by the collectors either via organisations like ELRA and LDC or available as free download via internet. To the best of my knowledge, the following list of data sets reflects all corpora which are available for the community, yet. As I am dealing mainly with speech processing, I grouped the data sets according to the modalities, namely audio-only and multimodal recordings; that means, besides audio samples, material of further modalities is provided. For each data set its main characteristics are briefly introduced; for detailed descriptions I refer to the corresponding papers.

Audio-only Data Sets In the collection of conversational data sets, to the best of my knowledge, only one distributed corpus can be found that relies on an unimodal setting, namely audio recordings.

ISCI Meeting Corpus In [Janin et al. 2003] the *ISCI Meeting Corpus* is introduced, providing English spoken material of speakers from all over the world gathered in a meeting room. Most of the participants are native English speaker (28 participants) whereas 25 participants are from non-English speaking countries. All of them have an academic background and are in the age of 20-62 years. In each session ten speakers interact with each other talking about one of four topics. The four categories are: i) understanding of academic problems, ii) data set recorder meetings, iii) issues of robustness in signal handling, and iv) network services. In total, 75 meetings are included in the corpus.

The equipment of the recording is as follows: Each participant wore a headset microphone which recorded the signal close to the speaker. Further, four omnidirectional microphones are posed on a table recording the whole situation. In addition to the audio samples, for each session a transcript is provided by the distributor.

Multimodal Data Sets In contrast to the ISCI Meeting Corpus, the following data sets supply at least two modalities. In all of the corpora participants used English to communicate with each other. Providing an overview, I concentrate on the main characteristics of the data sets and refer to the corresponding papers for further details.

ISL Meeting Corpus In [Burger et al. 2002] the *ISL Meeting Corpus* is introduced that contains 104 meetings with eight participants each. The participants – 18-70 years old – are mainly native American English speakers, but also a few of them are from different non-English speaking countries. However, the recordings are in English. The data sets consists of 103 hours of discussions on different topics related to various scenarios. The topics are prearranged to keep the recordings under controlled conditions, dealing with i) planning issues, ii) military topics, iii) games, iv) topic discussion in general, and v) chatting. There are no information about the distributions of the topics available.
The audio-visual recordings were done in a laboratory environment using eight table-mounted microphones and in addition, one lavaliere microphone for each participant. Furthermore, three standard cameras were used to observe the group in total. Additionally to the audio-visual material, meta-information like scenario type, participants' information, etc. are collected which allows a grouping of the session. Besides the transcription of the material, the data set was processed according to the formalities of Verbmobil.

M4 Corpus Another corpus that deals with prearranged or scripted scenarios is the *M4 Corpus* presented in [McCowan et al. 2003]. The recorded group meetings have a talk-like style; that means, one participant gives a monologue and thus, is mainly in action. In contrast, the group is reacting on the talk by taking notes, showing agreement or disagreement, etc. In total, eight participants are involved in each meeting where the constellations vary for each session. There are no detailed information about the participants themselves or the topics of the

meetings available.
The M4 Corpus provides also audio-visual recordings which are fully synchronised. For the audio data collection a table-mounted microphone array as well as 24 lapel microphones were used. The video streams were recorded with three TV cameras. In total, 60 meetings with five hours of data were collected. Besides the audio-visual information, the content of a whiteboard and a projection is in the view of one camera. Though, these are not provided as additional modalities.

AMI Meeting Corpus Linked to the M4 Corpus is the *AMI Meeting Corpus* (cf. [Carletta et al. 2005]). Both data sets provide audio-visual recordings of meetings that incorporate additional modalities. In contrast to the M4 Corpus, in the current corpus the whiteboard and the projector contents are separately recorded and thus, can be seen as further modalities. The audio recordings' setting is as follows: Besides headset microphones for each participant, omnidirectional table-mounted microphones were used. Furthermore, microphone arrays with four and eight omni-directional units and a binaural manikin collect audio samples. The whole group is monitored by two to three cameras that are either placed in the corner of the meeting room or in an overhead position. Additionally, for each participant close-up cameras are utilised. The material in total is transcribed and aligned with the audio streams. Moreover, annotation and labelling according to aspects like what is named, topics, activity of the group – what is done –, gestures, emotions using FEELTRACE (cf. [Cowie et al. 2000]), etc. is available. So far, no labelling regarding involvement is provided.
The scenario of the data set is related to investigations of group behaviour. Therefore, the four participants play a role over several sessions, namely i) project manager, ii) marketing expert, iii) interface designer, and iv) industrial designer. The goal is: designing a TV remote control whereas none of the participants is a professional designer. As this is a development task several meetings in different phases of the project take place. Hence, this data set provides possibilities for comparative studies.

VACE Multimodal Meeting Corpus Military meetings and thus, military topics are in the focus of the *VACE Multimedia Meeting Corpus* introduced in [Chen et al. 2005]. The scenario mainly concentrates on wargames and military activity planning like planning of rocket launches. The total number of meeting participants is not specified. The only information is that at maximum eight participants can sit around the provided table.

The fully time synchronous recordings were collected with two to six table-mounted microphones, one earset microphone for each participant, and ten camcorders mounted on an overhead rail system. Additionally, nine infrared cameras are utilised to collect the motion of the participants. Due to the enormous amount of cameras, which are configured in a way that each participant is in the focus of at least two cameras, and the time synchrony in the recordings a 3D tracking of the motions and gestures is possible. This is a distinguishing characteristic of this corpus.

Multimedia Database The *Multimedia Database* by [Campbell et al. 2006] is a collection of meetings recorded over 12 months. Finally, 12 sessions were selected to be included in the current collection. Each session presents scripted business-like scenarios which deal mainly with planning and progress of business activities. The meetings were recorded in various locations and thus, different recording environments are provided. Besides the varying number of participants, which is in the range of 4 to 12, also the length of each session is different.
The technical setup of the recording was kept fixed over all sessions. For visual data collection a table-mounted camcorder with a 360-degree lens was used. Further, for audio recordings four microphones in a windmill configuration were mounted on a table centred between the participants. Additionally, in some sessions a stereo microphone was used.

TableTalk The *TableTalk* corpus is introduced in [Campbell 2009] and further information will be given in Section 3.3.

D64 The latest corpus, the *D64* (cf. [Oertel et al. 2010]), is, besides TableTalk, the corpus which is most related to the investigations of this thesis regarding the involvement in a group conversation. Further, it is linked to the TableTalk corpus (cf. [Campbell 2009] and Section 3.3) as both have a similar setting. Both data sets provide material which contains non-scripted interactions in a colloquial style. In particular, in D64 the participants chat about topics of culture and politics. The data set is split in two sessions recorded on two different days where each session lasts 4 hours. Four and five participants (mainly the authors of the paper [Oertel et al. 2010]), respectively, sit together in an apartment, that means, in a naturalistic environment.
The apartment was just slightly modified to fit the recording equipment, but

keeping the naturalistic conditions of an informal environment. Two 360-degree cameras mounted on a table in the centre of the room as well as seven additional video cameras recorded the group and the participants' behaviour. For the purpose of a better movement detection reflection markers were adjusted on the participant which do not restrict the mobility. Further, headset and lapel microphones for each participant were applied to collect audio samples. In general, the data set was processed according to [Campbell 2009] in terms of video feature extraction. Additionally, the material is labelled with levels of involvement as described in [Oertel et al. 2011b].

2.4 Methods of Disposition Recognition from Speech

Up to now, the general idea of emotion recognition from speech and its evolution (cf. Section 2.2) was briefly discussed. Furthermore, data sets which are suitable to train classifiers for various kinds of dispositions were introduced. To establish proper recognition systems for dispositions, feature sets which can be automatically extracted from speech and which contain meaningful information are necessary. These will be discussed in the following Section. Moreover, the classifier itself that represents the core of each recogniser will be considered. A detailed description of the methods applied in my experiments is given in Chapter 4 and thus, the following collection is intended to be an overview, only.

2.4.1 Feature Sets for Disposition Recognition

It is common in the community to distinguish prosodic and spectral features. Prosodic features are mainly reflecting the characteristics of a voice and speaking style. On the other hand, spectral features represent the characteristics of the speech signal itself.

In the beginning of speech recognition the focus with respect to features was mainly on the spectral ones. In Table 4 in [Anusuya & Katti 2009] an overview of widely used features is given including wavelets, Linear Predictive Coding (LPC), and MFCC. LPC and MFCC are also commonly used in emotion recognition from speech. This was shown by various researchers, for instance, [Vogt & André 2005; Ververidis & Kotropoulos 2006; Vlasenko et al. 2008; Schuller et al. 2010a;

Schuller et al. 2011c; Sapra et al. 2013]. In [Ververidis & Kotropoulos 2006] as well as in [Schuller et al. 2011c] spectral feature sets are introduced and related to the usability in emotion recognition from speech. Especially as an result of the first INTERSPEECH emotion recognition challenge, it is shown that most of the systems are relying on MFCC based feature sets (cf. [Schuller et al. 2011c]). In my studies, I usually rely also on MFCCs (cf. [Rabiner & Juang 1993]) and Perceptual Linear Predictive Coefficients (PLP) (cf. [Hermansky et al. 1991]) features representing the spectral characteristics of the signal (cf. Section 4.2.1).

On the other hand, prosodic features are also used to evaluate the voice's characteristics of a speaker. These are mainly duration, pitch, intensity, speaking rate, and voice quality features. The term voice quality features includes features like noise-to-harmonics ratio, jitter, shimmer, etc. These methods are reviewed in [Ververidis & Kotropoulos 2006] and [Schuller et al. 2011c]. A quite exhaustive overview of applied features and the corresponding feature selection process is given in [Anagnostopoulos et al. 2012] where also the extraction methods for features are briefly described. In [Ververidis & Kotropoulos 2006] additionally vocal tract features like formants and bandwidth, speech energy, and the Teager Energy Operator are introduced. In my experiments on semi-automatic annotation (cf. Section 5.2.2) formants, their bandwidths, intensity, pitch, and jitter are used as additional prosodic features as introduced in Section 4.2.2. In [Fragopanagos & Taylor 2005] the authors report about empirical studies on prosodic features considering several emotions. These studies are limited to anger, boredom, happiness, and sadness, whereas a more extensive examination of prosodic features and emotions is given in [Cowie et al. 2001]. It was found that, especially, pitch is a discriminative feature for these emotions. Nevertheless, Fragopanagos & Taylor concluded:

> In fact due to their empirical nature, these observations tend to have a certain degree of variance or even disparity depending on the researchers and the material used for the studies. [Fragopanagos & Taylor 2005]

In addition to the aforementioned features, functionals applied to these are widely used (cf. [Vogt & André 2005; Schuller et al. 2010a; Schuller et al. 2011c]). Usually, mean, minimum and maximum of a feature, and various quantiles are applied as functionals to the values of the considered features. In particular, an overview is shown in [Schuller et al. 2010a; Schuller et al. 2011c].

Considering the presented features, [Zeng et al. 2009] discussed these in the context of audio-visual analyses and further, in a sense that detaches from emotions towards broader investigations. For this, features are more related to dispositions. Such issue is also reflected in publications like [Bencherif et al. 2012] regarding gender effects in the recognition of traits and [Scherer et al. 2012] where multimodal recognitions of user states are investigated. These aspects are further regarded in [Scherer 2011] leading towards disposition recognition in HMI. The presented works support the assumption that the established features from speech recognition and emotion recognition from speech can be transferred to the disposition recognition from speech.

Finally, as a kind of conclusion about the features which are used in emotion and disposition recognition from speech, it can be stated that:

> [The features'] relative contribution can also vary greatly, depending on the database being analysed: For instance, for data based on scripted speech, [prosodic] features are normally of no value, apart from some specific applications such as data mining in movie archives. On the other hand, as the register comes closer to spontaneous/real-life speech, these features can gain considerably in importance. [Schuller et al. 2011c]

2.4.2 Classification Methods for Disposition Recognition

Based on the corpora and extracted features which were introduced in the previous Sections, classifiers can be trained and afterwards evaluated. Usually, the training process and thus, the performance of a classifier was linked to the utilised data set. In [Schuller et al. 2009a; Schuller et al. 2010a] the authors showed that this limitation could be overcome by cross-corpus training and evaluation. For this, a data set with a certain characteristic but less training material is just used to adapt a classifier which was already pretrained on another corpus providing a suitable amount of samples. Investigations by Schuller et al. support the assumption that such a cross-corpus handling seems to be independent of several kinds of classification methods. Of course, this is related to the individual classifier and thus, the currently applied methods in emotion and disposition recognition from speech will be reviewed.

In the beginning of speech processing and recognition simple comparison operations were applied (cf. [Anusuya & Katti 2009]), using analogue filter banks and logic circuits. Further developments in the community led to more complex systems which are able to recognise words. An advanced kind of classifier, which fits the conditions for speech recognition quite well, was introduced by the Hidden Markov Model (HMM) (cf. [Cave & Neuwirth 1980; Rabiner & Juang 1993]). Nowadays, HMMs are a quite popular classifier method.

As it is stated in [Schuller et al. 2011c] various other methods, for instance Support Vector Machines (SVMs), ANNs, etc., were transferred from speech recognition to emotion recognition from speech. The most commonly used classification approaches are collected in [Schuller et al. 2011c] and also reflected in [Vlasenko 2011]. Additionally, in [Zeng et al. 2009] and [Bousmalis et al. 2013] they are related to multimodal investigations of emotions and further, to first ideas of disposition recognition.

In the following, the methods are sketched briefly, providing also a rough idea of the classifier's performance.
In [Anagnostopoulos et al. 2012] a marvellous overview of commonly used classification methods is given and additionally, a collection of emotion recognition results is presented. This paper also provides a selection of references for each approach that is quite suitable for further reading. The following considerations are based on [Schuller et al. 2011c] and [Anagnostopoulos et al. 2012] to which I refer for corresponding references. Additionally presented methods are referenced at the specific point.
Common classification approaches can be divided into two classes; namely single classifiers and hybrid classifiers including ensemble as well as voting methods. So far, it is quite common in the community to apply single classifier architectures to recognise emotions from speech. The most popular classifier is the HMM using HMMs as internal models. A Gaussian Mixture Model (GMM) is a probabilistic model that applies convex combinations of multi-variate normal distributions to model the characteristics of a given probabilistic density function and thus, they are a special type of a production model. In contrast, HMMs are characterised by a twofold production process. HMMs model the temporal evolution and finally, produce a corresponding output. Another common statistical method is the SVM which aims to find a separation in a transformed feature space by applying a kernel function. Special cases of SVMs are Twins SVMs and Fuzzy-Input Fuzzy-Output SVMs (cf. [Borasca et al. 2006; Thiel et al. 2007]). Inspired by the computational intelligence's community several classifiers are used in emotion recognition from speech, learning relations between presented inputs and corresponding emotional classes. Most of them are based on ANNs (as additional

Table 2.2: Performance of selected classifiers achieved on EmoDB. The classifiers are as follows: Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN), C4.5 algorithm (C4.5), and Random Forest (RF). The list is adapted from [Anagnostopoulos et al. 2012] representing only values obtained by speaker-independent experiments. It is to notice that the values itself are hard to compare amongst the different classifiers since the evaluation lack uniformity in the validation measures (cf. [Anagnostopoulos et al. 2012]).

Classifier	Classification performance
GMM	81.0%
HMM	78.4%
SVM	88.6%
ANN	55.0%
C4.5	61.5%
RF	48.0%

reference cf. [Fragopanagos & Taylor 2005]) and Fuzzy Sets. Furthermore, Evolutionary and Genetic Algorithms are applied to tune parameter sets of the classification approaches. In the ANN approach various architectures are considered, mainly, Multi-Layer Perceptrons (MLPs), Probabilistic Neural Network, Deep Neural Networks (DNNs), and Simple Recurrent Networks (SRNs), more specifically Segmented-Memory Recurrent Neural Networks (SMRNNs) (cf. [Chen & Chaudhari 2009]). Furthermore, Echo State Networks (ESNs) (cf. [Jaeger 2001]) provide a powerful approximation of dynamic processes like emotion recognition. Based on ideas of artificial intelligence, decision tree approaches classify given samples according to predefined criteria. A well-known algorithm in this context is the C4.5 algorithm which is used within the decision trees' method. Linking several decision trees results in a Random Forest which is an Ensemble Classifier. Further modelling techniques – inspired by linguistic observations – are Bag-of-Words (BoW) and n-gram models. BoW is a numerical representation of a text given in a vector space modelling. It is widely used in the automatic categorisation of documents and natural language processing. Therefore, it is also suitable for emotion classification from text and speech. The n-gram models are probabilistically based predictors which estimate the next item in a given sequence. Again, this can be applied to text and speech. To get an impression of the classifiers' performance, Table 2.2 presents classification results by different architectures. The Table is adapted from [Anagnostopoulos et al. 2012] –

presenting also the references to the corresponding recognition results – and concentrates on results achieved on EmoDB. It is to notice that the values are just an overview and no direct comparison is possible since the way the classifiers were evaluated lack uniformity. Nevertheless, the classification performance indicates the potency of each method.

The second category of classification methods are hybrid classifiers that also incorporate ensemble and voting approaches. Therefore, these are also related to aspects of fusion. The way of combining different kind of classifiers is manifold. In [Zeng et al. 2009; Schuller et al. 2011c; Anagnostopoulos et al. 2012; Schels et al. 2012] various approaches for fusion architectures and combinations of different classifiers are presented. The proper approach is highly depending on the classification task and the material which is to be handled. In [Kuncheva 2004] the levels where fusion could be applied are discussed, namely on i) feature level – combining different kinds of features –, ii) mid level where the final decision is learned, and iii) decision level applying a predefined combination rule to the output of several classifiers. On the other hand, ensemble techniques like Boosting, Bagging, and Stacking help to improve the performance of classifiers, especially, in combination with voting methods like maximum accuracy methods. In [Anagnostopoulos et al. 2012] a collection of results is presented, achieved with hybrid and ensemble methods. From this, it can be concluded that fused architectures usually perform better than single classifiers as the former combine the advantages of each method.

Finally, as the different classification methods are on hand, it is necessary to train such recognisers. For this, several frameworks and toolkits are available (cf. [Nguyen 2009]). The most common toolkits for HMMs are the Hidden Markov Toolkit (HTK) (cf. [Young et al. 2009]) developed at the Cambridge University and openEAR (cf. [Eyben et al. 2009]) designed at the Technical University Munich. In my experiments, I usually applied the HTK (cf. [Young et al. 2009]). For other classifiers, a powerful tool is WEKA (cf. [Witten & Frank 2005]) providing statistic and artificial intelligence methods.

The presented approaches and architectures are widely used in the community of emotion recognition from speech. As it is denoted in [Scherer 2011] most of the methods can be also applied in the context of disposition recognition and thus, be transferred to the disposition recognition from speech. As already stated in Section 1.4 the disposition of a user can also be a relevant information for the assessment of a conversation and the other way around. Therefore, it is interesting to investigate this aspect. In Chapter 6 this is done on a specific topic – namely a group conversation – and in general, in the following Section, focusing on the involvement of a user participating in a conversation.

2.5 Automatic Analysis of Involvement

As already discussed in Section 1.4, involvement in a conversation is an important source of information and thus, worthwhile to be analysed. Moreover, such analysis is an issue related to dispositions since aspects of situatedness (cf. Definition 1.5 on page 7 and [Simpson 2002]) and personal behaviour are reflected by the involvement. In the sense of psychological investigations, the involvement has already been regarded for a longer period (cf. [Krugman 1965]). In this context usually the group's reaction as well as the interactions of the group's participants are under investigation by psychologists which is subsumed by the term *group dynamic* (cf. [Wageman 1995]). The research on group discussions (cf. e.g. [Lamnek 2005; Keyton & Beck 2009; Kühn & Koschel 2010]) fits in the group dynamic analyses. By the community various sets of conversations are defined that include also different kinds of constellations between the communication partners. For instance, the most obvious grouping is i) a monologue by one participant, ii) a dyadic interaction, iii) a sub-group interaction, or iv) a conversation of the group in total. This assignment is directly reflecting the number of participants who are interacting in such a conversation that is significant for an evaluation of the group discussion. Furthermore, different hierarchical constellations influence the conversation and interaction (cf. [Goodwin & Goodwin 2000; Keyton & Beck 2009]). On the other hand, several setups are under investigation representing different types of situations in daily life. These are, for example, evaluations in office environments (cf. [Veitch & Kaye 1988]), medical or health care environments (cf. e.g. [Banister & Begoray 2004]), group meetings (cf. [Carletta et al. 2005]), and teaching situations (cf. [Kang 2012]). Usually, the analyses are mainly done manually and the automatic part is low. Even with the upcoming technical analyses, the psychological investigations are relevant – even if they are manually driven – to establish fundamental theories and ground the automatically derived results.

In [Krugman 1965] the analysis of involvement with dispositional characteristic was introduced, namely the influence of TV commercials on the involvement of the audience. For this, Krugman started furthermore the observation of involvement also in a technical related manner, from my point of view. In [Antil 1984] the focus is linked to commercials and certain products as well whereas the approaches are still founded in manual analyses. Besides an updated state of the art, Antil opts for an operationalisation of the involvement's analyses. Nevertheless, the considerations are still grounded in psychology.

With [Hartwig et al. 2002] the focus of analysis slightly changed from psychology towards a technically inspired research. Furthermore, the authors reviewed the observations on involvement up to that point and concurrently used deception detection as a test case for involvement. For this, manual investigations are accompanied by statistical analyses. A technical investigation of involvement was started in [Wrede & Shriberg 2003] where four classes of user behaviour in an interaction were examined, namely being amused, disagreeing, being involved, and other (cf. [Wrede & Shriberg 2003]). They discuss the involvement in the context of emotions located in "Hot Spots" [Wrede & Shriberg 2003] that mark a situation which is relevant for the interaction. This is also investigated in [Oertel 2010]. Related to the work of Wrede & Shriberg is the research presented in [Yu et al. 2004] that also utilises the recognition of emotions to establish a detection of engagement in remote conversations. In both publications feature sets are examined which are suitable to fit the particular issues whereas it is, so far, unclear if those features are universal and thus can be transferred to other data sets or situations. Especially in [Yu et al. 2004], first classification results on utterance-level – to the best of my knowledge – based on their feature set are presented using HMMs, achieving 75% and 51% performance results on acted and spontaneous speech, respectively, having five discrete classes representing the numerical level of arousal, valence, and engagement (cf. Table 1 in [Yu et al. 2004]). For detection of engagement in continuous speech an accuracy of 63% was obtained (cf. Table 2 in [Yu et al. 2004]).
In relation to remote conversations, in [Gustafson & Neiberg 2010] the engagement in a radio broadcast show is considered while the entertainer is talking to listeners linked-in via telephone. They found that in such an interaction the behaviour and speaking style is aligned to each other during the conversation. Furthermore, prosodic cues and backchanneling is investigated as well.
Considering the presented research activities, the involvement is so far not integrated in a broader framework. Moreover, it is 'assumed' to be somewhat separate. In my research, I handle the involvement more generally, being integrated in the framework of dispositions. On the other hand, I do not distinguish between en-

gagement and involvement as from my point of view, both terms reflect similar research issues and thus, can be used synonymously.

Like in [Yu et al. 2004] I argue that for a general analysis of involvement, a multimodal observation of the user and the group is necessary. As discussed in [McCowan et al. 2005], in addition to psychology, the analysis of a group conversation including the involvement is an emerging domain. It is highly related to various research communities like speech processing, image, or video processing, and information fusion (cf. [McCowan et al. 2005]). Furthermore, such kind of analyses combines several levels of perspectives, namely social, psychological, synchronous, and individual perspectives (cf. [McCowan et al. 2005; Schuller et al. 2007b; Campbell 2009]. On the other hand, the issue will be more complex if interactions with multiple users and a technical system are observed (cf. e.g. [Harris & Rudnicky 2007; Kumar et al. 2007]). Thus, an automatic analysis of group conversations is, up to now, focused on the group in an HHI setup. Nevertheless, from such information and findings advanced technical systems (cf. [Wendemuth & Biundo 2012]) might benefit.

Consequently, involvement in a multimodal setup is considered in [Oertel et al. 2011a] and [Oertel et al. 2011b] using audio and video recordings of a group interaction. The analyses are done based on the D64 corpus (cf. [Oertel et al. 2010] and Section 2.3.2). The whole scenario as well as the observation is inspired by [Campbell 2009]. In [Oertel et al. 2011a] results on D64 are presented according to the modalities and further, by combination of these. The authors achieved 74.4% mean accuracy at its best at which the involvement is grouped in ten grades. This is comparable to the findings on TableTalk (cf. Section 6.3). So far, the most capable system is shown in [Rasheed et al. 2013] which allows a real-time analysis in terms of behaviour detection "with respect to involvement, dominance, and impedance" by classifying levels of involvement.
My work is related to the context of involvement detection; to be specific, it is the detection and classification of changes in the involvement. For this, the TableTalk corpus (cf. [Campbell 2009]) is used where results are given in Section 6.3.

To conduct reasonable classification and detection experiments, a proper preprocessing in the sense of labelling has to be applied to the corpora which is also the case with involvement. Reviewing the literature regarding the labelling of involvement, so far, no common classes or categories are established. For instance, in [Krugman 1965; Antil 1984] the observed categories are mainly related to the involvement towards a product. The engagement in a conversation in a technical sense is labelled in [Yu et al. 2004; Gustafson & Neiberg 2010]. Moreover, [Wrede & Shriberg 2003; McCowan et al. 2005; Oertel et al. 2011a] give labels for involvement in various granularities. The most detailed labelling approach is presented in [Oertel et al. 2011a] classifying levels of involvement. They discuss the processes and describe a kind of user's manual which is given to the labellers. Further, in [Oertel et al. 2011b] the reliability of the labelling, based on the aforementioned ten grades, is shown with $\kappa = 0.56$ for 30 labellers. The labelling for TableTalk which is presented in Section 6.2 is connected to the procedure in [Oertel et al. 2011a]; though, in my work changes on involvement are regarded. The most abstract categories, namely interest, dominance, and impedance, are given in [Rasheed et al. 2013] which aim on socio-feedback of a system. This kind of research is linked to the idea of companion-like systems (cf. [Wendemuth & Biundo 2012]) for which this thesis provides parts of analyses that will be discussed in the following Chapters.

2.6 Summary

In the current Chapter the developments in the HMI as well as in fundamental evolutions in disposition recognition from speech have been reviewed. For these, also remarkable achievements have been introduced and discussed. Furthermore, a general overview of the automatic analysis of involvement has been given. This user characteristic shown in conversations and interactions was integrated in the user's disposition recognition and further, linked to the context of HMI. In addition to the methods used in feature extraction and classification, data sets for both disposition recognition from speech and involvement analyses have been introduced. As stated in Section 2.4.2, the performance of classifiers is highly depending on the applied corpora. Thus, in the following Chapter the data sets used in my investigations will be introduced in greater detail.

Chapter 3

Data Sets

Contents

3.1 Acted Data Sets 47

3.1.1 EmoDB 48

3.1.2 eNTERFACE 49

3.2 Non-Acted Data Sets 50

3.2.1 Last Minute 51

3.2.2 EmoRec I+II 53

3.3 TableTalk 56

3.4 Summary 58

IN this chapter I introduce the data sets which were used in the experiments of this thesis. As it is common in the research community on emotion recognition from speech, I distinguish two types of data sets: i) acted and ii) non-acted corpora. It is obvious that this is just a small selection of data sets which are used for emotion and disposition recognition from speech. A more advanced collection of material can be found, for instance, in [Campbell & Reynolds 1999; Anusuya & Katti 2009; Zeng et al. 2009; Schuller et al. 2010a; Schuller et al. 2011c; Anagnostopoulos et al. 2012]. Further details are also given in Section 2.3.

3.1 Acted Data Sets

In the beginning of speech recognition and, especially, emotion or disposition recognition from speech data sets with acted material were used to train and test any kind of recognition and classification systems since such corpora supply controlled conditions in the recording and further, provide a kind of ground truth in the emotions label that is necessary for the validation of recognition systems.

Acted material means that the participants of the recording were either actors or naïve speakers who were asked to react in a specific manner, this is acting the favoured emotion. Therefore, a controlled situation was created where the output, that is the intended classification result, was predefined and hence, the evaluation could be done in an easy way. Further, it can be assumed that in acted data sets the expressiveness in the emotions is high (cf. [Batliner et al. 2000]). In particular, as it was yet unclear which method and features are useful in the recognition process, such acted material was a suitable starting point to evaluate those research issues. In this thesis, I also started with observations which were based on data sets recorded under predefined situations. Nevertheless, according to [Batliner et al. 2001] who heavily argued for non-acted data sets, I switched towards corpora that are more natural (cf. Section 3.2) applied to the methods presented in Chapter 4.

3.1.1 EmoDB

The EmoDB [Burkhardt et al. 2005] is widely used in emotion recognition from speech (cf. [Schuller et al. 2010a]). However, it was intended to get suitable material to investigate the speech synthesis process afflicted with emotional characteristics (i.e. the set of Ekman's Basic Emotions [Ekman 1992] is realised, namely, anger, boredom, disgust, fear, joy, neutral, and sadness). Due to the high quality of the recordings it became a kind of benchmark data set to test methods and classifiers, for instance [Schuller et al. 2009a; Xiao et al. 2009; Albornoz et al. 2010; Iliou & Anagnostopoulos 2010; Schuller et al. 2010a; Böck et al. 2010].

According to [Burkhardt et al. 2005] the recoding conditions are as follows: the recordings were done in an anechoic cabin using a Sennheiser MKH 40 P 48 microphone and a Tascam DA-P1 portable DAT recorder. The sampling rate was 48kHz which was afterwards down-sampled to 16kHz. A special characteristic of the data set is that the content of the utterances is not related to the emotional expression and, further, due to the acted situation, the data does not represent the speaker's disposition. The ten sentences which are spoken in German can be found in [Burkhardt et al. 2005].

Further, the authors applied listening evaluations to the material to assure high quality. They selected only samples with at least 80% recognition rate and 60% naturalness. For this a subset of approximately 493 samples from totally

800 was collected. The total number of utilised samples varies from publication to publication as some authors reselected the material for their studies.

For my studies I relied on the subset given by Burkhardt et al. and therefore, used 493 samples. Due to the selection process the distribution of the samples according to emotions is unbalanced (cf. Table 3.1). For this both measurements, WA (Equation 1.1) and UA (Equation 1.2) (cf. Section 1.3.1), have to be applied to assess the results of any classification done on EmoDB.

Table 3.1: Emotion's distribution in the EmoDB.

Emotion	Number of samples	Overall time
anger	127	05:34.77
boredom	79	03:39.97
disgust	38	02:08.73
fear	55	02:03.70
joy	64	02:43.89
neutral	78	03:04.32
sadness	52	03:27.43
total	493	22:42.81

3.1.2 eNTERFACE

The data set eNTERFACE was recorded during a summer school in 2005, which had the same name, held at the University Louvain, Belgium. The participants of this school served as the actors in the recordings. Besides audio, video recodings were done to generate a data set of facial expressions simultaneously. For this thesis, I concentrate on the audio part and hence, present only information which are related to this material. Video settings and analyses can be found in [Martin et al. 2006]. In eNTERFACE, 42 participants (34 male and 8 female) from all over the world (cf. [Martin et al. 2006]) were recorded, thus, most of the speakers were non-native, nevertheless, the language is English. As the authors were interested in Basic Emotions they applied the ideas of Ekman & Friesen [Ekman & Friesen 1978] that led to the following emotional categories used in eNTERFACE derived from facial expressions: anger, disgust, fear, happiness, sadness, and surprise. Similar to EmoDB participants were asked to act the emotions. In contrast, by the setup of the recordings, the speakers were introduced to a specific emotion by listening to a short story which describes an emotional situation, and afterwards they were asked to react on it with any kind of utterance. This means, the emotion was induced but with the participant's knowledge. The reaction was assessed by two experts according to the expressiveness, especially, whether this is an unambiguous emotion regarding the predefined sets.

The data set was recorded with a standard mini-DV digital video camera and a "high-quality microphone" [Martin et al. 2006] which is not further specified. In total, each participant provided several recordings for each emotion and therefore, 1170 samples are available. Only the audio material is used throughout the experiments which leads to a slightly different number of available samples (cf. Table 3.2) in comparison to [Martin et al. 2006] as their counting is based on the video samples.

For the experiments in this thesis the full set of samples was used. As the distribution of the material (cf. Table 3.2) is much more balanced as in EmoDB it is expected that WA and UA are not significantly different.

Table 3.2: Sample's distribution per emotion in eNTERFACE.

Emotion	Number of samples	Overall time
anger	200	10:51.56
disgust	189	08:44.16
fear	189	08:47.28
happiness	205	08:36.32
sadness	195	09:56.20
surprise	192	08:22.00
total	1170	46:30.24

3.2 Non-Acted Data Sets

As introduced in Section 3.1 I applied the methods which are used in this thesis to non-acted material as well. Why should such naïve data sets be considered? Batliner et al. [Batliner et al. 2000; Batliner et al. 2001] discussed the importance of real-life material in detail. Especially, the fact that dispositions in general, but emotions in particular, are expressed in a subtle manner in real-life interactions pushes the research community towards non-acted data sets. Especially, in [Batliner et al. 2000] a comparison of three sets, namely acted, read, and real-life emotions, is given. Further, some emotions identified by Ekman [Ekman 1992] are not represented in naïve HMIs, for instance disgust (cf. e.g. [Siegert et al. 2011]).

When switching from acted to non-acted corpora several aspects have to be considered. Of course, the methods and the parameters of the classifiers have to be adapted to the data sets and thus, for the usage in real-life applications. This needs some effort and research. But, an important aspect is to get a feeling for the recognition results. That means, in acted material depending on the classification methods accuracy values of more than 75% are reached (cf. e.g. [Schuller et al. 2009a; Böck et al. 2010; Glüge et al. 2011]) with tuned classifiers quite easily. In contrast, as, for instance, discussed in [Batliner et al. 2000; Schuller et al. 2009a; Schuller et al. 2010a] such high accuracy values cannot be achieved using material which is not as expressive. Especially, Batliner et al. compares the different conditions and found a decreasing in the recognition results [Batliner et al. 2000]. From this, lower accuracy values are expected for real-life, non-acted data sets. On the other hand, the results have to be set in correlation to either the chance level, which is a pure probabilistic decision, or the 'ultimate system': human assessment. Considering the performance of humans in the classification of dispositions, even those cannot get an accuracy of 100%; that means, all dispositions given in a corpus can be recognised correctly. I encourage the reader to perform an experiment: listen to near real-life data as presented in Section 3.2.1 and 3.2.2 and do the classification job by yourself. Then the complexity of such task will be obvious and the achieved classification results can be ranked accordingly. Therefore, I argue that any classification result have to be related to the performance of the human and in fact, applying this point of view, the classification methods achieve significant results.

3.2.1 Last Minute

A scenario which represents naïve HMI is the Last Minute corpus [Rösner et al. 2012]. Whereas in [Rösner et al. 2012] information of the data set itself is given, [Frommer et al. 2012b] describes the technical equipment and realisation of the recordings and further, first analyses done on this material. This data set was recorded in the context of the SFB/TRR 62 at the Otto von Guericke University Magdeburg in a Wizard-of-Oz (WoZ) setting. This means, the system's actions are simulated by an operator which is usually called wizard. The data was recorded with six cameras (different types and manufacturers) and two Sennheiser ME 66 shot-gun microphones as well as a Sennheiser headset. For a subset of users also biophysiological signals like heart rate, skin conductance level, etc. were taped using the NEXUS-32 amplifier and the Biobserve software.

The story of the Last Minute scenario is a follows: the participants were asked to evaluate a novel HMI system which is highly adaptable to its user and hence, a kind of a companion (for the purpose of the systems setup cf. [Rösner et al. 2012] and for the ideas related to companion technologies cf. [Wendemuth & Biundo 2012]). To perform a system's adaptation towards a user the interactions reflecting the personality and the personal reaction of a user in different situations are necessary. Therefore, in the beginning several personal information like name, age, gender, etc. are enquired. Afterwards, the participant has to be engaged in a computer-based virtual task, preparing his luggage for a journey which he has won. To get different emotions and dispositions the user is faced with several barriers (cf. [Rösner et al. 2012]) that are implemented in the scenario called *baseline*, *challenge*, *listing*, and *waiuku*. The barriers are followed by a specific situation that I also refer to by the barrier's name. Further, the material of these situations are utilised in my experiments. A detailed description of the situations which are also representatives of class labels can be found, for instance, in [Siegert et al. 2012c] and thus, are introduced briefly at this point.
Baseline is the first part of the scenario and follows the collection of user data. It is assumed that the participant is at this moment relaxed and the first excitement has gone. As the name suggests, it is the baseline to calibrate the methods and classifiers to the specific user.
Challenge is the label for the first task occurring in the scenario. The predefined limit of the luggage is reached and hence, the user has to rearrange the baggage. This includes unpacking and selecting of other items.
Listing describes the itemising part of the scenario. Here, the participant is mainly engaged listening to the description of items which are already packed into the luggage. Hence, this situation is mainly driven by the system.
Waiuku reflects the last section of the scenario. The participant comes to know the destination of the journey. So far, he believes that this will be a summer trip, as the recordings were made in summer, but the final destination is in New Zealand which means it is winter there. The participant has to re-pack his luggage accordingly. This puts the participant under pressure as also the time limit expires which was set since the trip should start almost immediately after the experiment.

As mentioned before, the Last Minute is a kind of naïve HMI where in total data of 130 participants was recorded and accordingly, 56 hours of material are available. To keep this near real-life characteristic, the pushing of the wizard is kept minimal. Nevertheless, its influence can be recognised according to the user's reactions. For the sake of a proper experimental design, psychologists were involved in the planning of the scenarios. Moreover, the single recordings were done in a predefined structure to highly ensure the same conditions for each participant (for specifications cf. [Frommer et al. 2012a]). Finally, questionnaires filled by each participant reflect the user's personality traits and their ideas about this particular HMI.

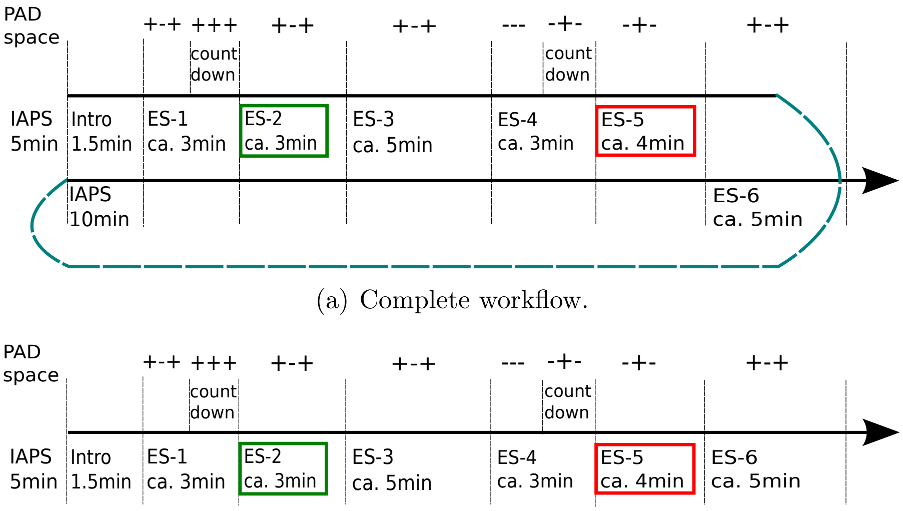
For the experiments in this thesis I used a subset of twelve participants for which at that moment full transcriptions and annotations were available. Further, for those users the full set of modalities are on hand (namely audio, video, and biophysiological signals) which is important for multimodal aspects (cf. Section 5.2 and 5.3). Altogether, I ended up with 219 audio samples given in Table 3.3 where the distribution according to the single participants is approximately equal. In total they indeed differ for the single scenarios.

Table 3.3: Sample's distribution per scenario part in Last Minute.

Scenario	Number of samples	Overall time
baseline	69	01:39.14
challenge	47	01:30.49
listing	39	00:59.49
waiuku	64	02:15.93
total	219	06:25.05

3.2.2 EmoRec I+II

In [Walter et al. 2011] another data set with near real-life HMI is introduced containing induced dispositions (cf. also Section 3.2.1). For the experiments in inter- vs. intraindividual disposition classification I mainly used this material, called EmoRec. As in Last Minute (cf. Section 3.2.1) this data set is recoded in a WoZ manner and was collected in the SFB/TRR 62 by the Medical Psychology Group at the Ulm University. The focus of this material is on multimodal data collected with enriched biophysiological signals, namely skin conductance level, respiration, blood volume pulse, heart rate, and electromyogram (cf. [Walter et al. 2011; Böck et al. 2012b]). The equipment was a Canon MD215 camcorder for the video recordings and a Sennheiser ME66 shotgun microphone with a sampling frequency of 44.1kHz collecting audio samples. The audio material is divided into two parts according to the recording conditions as in the first one the internal microphone of the camcorder was used only. Nevertheless, the sampling frequency is equal in both cases. The collection of the biophysiological signals was done with the Biobserve software applying the Nexus-32 amplifier. The signals of all modalities were postprocessed to ensure synchronicity on the material.



(a) Complete workflow.

(b) Workflow of the EmoRec I subpart.

Figure 3.1: Workflows of the EmoRec data set. EmoRec I is the subpart of the full set (cf. Figure 3.1(a)) which is just one pass of the scenario. The green and red frame indicate the Experimental Sequences with positive and negative dispositions, respectively.

According to [Walter et al. 2011] the scenario of the data set is as follows which is also visualised in Figure 3.1. The full scenario consists of two runs repeating the same steps in the experimental setup where the first run represents a stand-alone data set, called EmoRec I (cf. Figure 3.1(b)). Further, the both runs as well as the subsequences were recorded in a predefined structure to highly ensure the same conditions for each participant (for specifications cf. Table 1 in [Walter et al. 2011]). The whole story is to interact with a mental trainer simulated by the popular game ‘Concentration’. The idea of the game is to find corresponding pairs of cards given on a deck. In correlation to the disposition which should be induced the level of difficulty of the deck varies. Furthermore, other types of reactive elements in an HMI are varied to push the participant towards dispositions, namely misunderstandings, time delays, etc. which are controlled by a wizard.

As introduced in [Walter et al. 2011] the participant is routed through several parts of the PAD space [Mehrabian 1996] (cf. Section 1.2.3). This is indicated by the + and − symbols in Figure 3.1 on the preceding page where a + means high and − low values in the corresponding dimension. Each change of the position in the PAD space is linked to a new deck and therefore, a new Experimental Sequence (ES). In Figure 3.1(a) on the facing page as well as in Figure 3.1(b) on the preceding page two ESs are marked that represent prototypical characteristics of positive (ES-2) and negative (ES-5) dispositions. This is the reason why most of the results presented in this thesis are based in these two ESs.

In total, data of 125 participants were recorded and postprocessed. Unfortunately, due to the high effort of manual annotation just a subset of the data is so far coded with FACS. This is done by a trained coder according to guidelines given by [Ekman & Friesen 1978]. As such manual annotation is a limitation for data sets I proposed in [Böck et al. 2013a] a framework to reduce such effort and hence, get the annotation of facial expressions faster. However, to be consistent in the analyses, especially, for fusion aspects, so far, I have to consider the particular subset which is already annotated. I ended up with 20 participants (ten women and ten men) who are all native Germans and to the best of my knowledge no actors. The recorded material has a total length of $20 \cdot 7 = 140$ minutes (cf. times of ES-2 and ES-5 in Figure 3.1(b) on the facing page). Eliminating all non speaking parts, this results in a speech lenght of round 24:03 minutes in total. As psychologists accompanied the data recordings and set up the scenario it can be assumed that the participants react in an emotional way and thus, show the intended dispositions. Nevertheless, besides the recordings each participant has to indicate his personal disposition in the PAD space utilising the SAM rating [Bradley & Lang 1994]. The results are quite significant, this means, show that the induction of the disposition worked. They are presented in Table 3.4 on the next page according to [Böck et al. 2012b], whereas analysis of SAM ratings are provided by psychologists who co-authored the publication.

Since EmoRec as well as Last Minute (cf. Section 3.2.1) are generated in the same project (SFB/TRR 62) the set of participants was not disjunct. A subset of 8 participants was recorded in both scenarios. This was intended to collect material as a kind of test samples to validate findings from one scenario in the other. Therefore, both data sets are linked and can be used to examine naïve HMI with different settings but similar participants. The samples of theses 8 participants is included in the material used for the experiments done on the two data sets.

Table 3.4: Self-ratings of the participants during the ES' in PAD space according to SAM (cf. [Böck et al. 2012b]).

	ES-2			**ES-5**		
Participant	**P**	**A**	**D**	**P**	**A**	**D**
112	7	3	7	4	7	4
114	8	3	8	3	7	3
118	7	5	7	4	7	6
125	7	7	8	1	1	9
127	8	3	8	2	7	3
129	8	4	7	4	7	4
208	7	3	7	1	9	1
212	9	9	4	3	6	4
215	9	2	9	7	3	9
219	9	1	9	2	7	2
225	8	5	9	8	5	9
226	7	6	7	2	7	4
308	7	4	7	3	6	4
423	7	3	7	7	1	5
427	8	5	2	3	3	2
506	7	7	6	3	7	7
510	7	4	8	5	6	5
511	7	3	7	3	6	2
518	9	5	7	1	7	2
602	7	3	7	2	8	2
mean	7.2	4.3	7.1	3.4	5.9	4.4

3.3 TableTalk

[Campbell 2009] introduces a corpus called TableTalk which supplies a naïve group conversation. In contrast to the other data sets (cf. Section 3.1 and 3.2), the corpus is an HHI providing near real-life conversations with a non-fixed domain. This multimodal, multi-party corpus contains recordings with at least four participants on three different days. On one day, an additional participant joined the constellation. She is an native English speaking Australian woman. Since she is not occurring in the two other days I suggest to skip this recording for analyses as thus, comparisons between recordings of the same persons on two different days can be done. One advantage of TableTalk is that any discussion is not prescripted and further, on a non-fixed topic. Most of the time it is a colloquial chat including slurring of words, backchannels, and filler words.

Independent of the day, the conversations are recorded with one shotgun microphone and one 360-degree camera which were both positioned centred on a table. It is important that due to the single microphone, only one channel for all participant's voice activities is available. Therefore, it is not directly possible to investigate a single speaker in the audio recording as the voice activities are several times mixed and thus, highly overlapping. This is the main disadvantage of the TableTalk data set. On the other hand, the video signal provides good material for analyses. Moreover, the data set is already postprocessed according to participant's movements, especially, in terms of face and body detection (cf. [Campbell & Douxchamps 2007; Campbell 2009]).

For experimental results presented in this thesis, I concentrated on the first part of TableTalk which is called *day1* lasting 34:34 minutes in total. In this HHI four participants (two female, two male) are chatting about youth culture in Japan, movies in general, and plans for evening events. The conversation is in English, however, only one participant is a native English speaker; the participants are citizens of the following countries: Belgium, Finland, Great Britain, and Japan.

Due to the postprocessing according to [Campbell 2009] a general framework for analyses was given. To utilise the advantage of given video features I decided to apply the given frame rate of 100ms and adapt this also for audio investigations. Doing so resulted in a set of 20541 samples which are to be processed. Each sample is extracted with a sample rate of 100ms and neither overlap between samples nor a windowing is given. Given the 20541 samples after smoothing this results in a total time of 34:24 minutes.
In contrast to video material, the audio was not preprocessed to be applied in a classifier. This was part of my experiments. Further, the data set in total had no annotation with respect to involvement which I understand as a kind of disposition. In Section 6.2 the process of preparing the data set is presented as well as reliability analyses of the annotation are mentioned. These are also compared to data sets with similar conditions in the sense of the annotation process and labels.

3.4 Summary

The data sets which were presented in this Chapter are prototypical representatives of acted and non-acted corpora.
Especially, the acted material, namely EmoDB, supplies expressive emotions which can be clearly identified by human beings and also by automatic systems. With eNTERFACE a switch towards realistic corpora is started as no actors are playing emotions like they might occur in interactions. On the other hand, these data sets are still with the HMI context. Considering Last Minute and EmoRec (cf. Section 3.2), these corpora supply naturalistic participant's behaviour that is related to his disposition. Furthermore, they are recorded directly in an HMI context. Therefore, they are suitable for my further investigations. Moreover, with TableTalk a HHI is on hand that provides dispositions and involvement cues which influence the conversation.

As the data sets are presented and their characteristics are introduced they can be applied to develop automatic systems, which are part of technical devices, to recognise emotions and dispositions from speech in HMI. To achieve this aim suitable classification methods and feature sets have to be investigated, adapted, and improved. These aspects are discussed in Chapter 4.

Chapter 4

Methods for Inter- vs. Intraindividual Disposition Recognition

Contents

4.1 From the Acoustic to Training Material 60

4.1.1 The Annotation Process 60

4.1.2 ikannotate 61

4.1.3 Semi-automatic Annotation 64

4.2 Features 67

4.2.1 Emotion and Disposition Recognition from Speech 68

4.2.2 Semi-automatic Annotation 73

4.3 Classifiers 80

4.3.1 Hidden Markov Models 81

4.3.2 Evaluating Parameter Sets 84

4.3.3 Recurrent Neural Networks 91

4.3.4 Comparing Hidden Markov Models and Neural Networks . 92

4.4 Fusion Aspects 97

4.5 Summary 99

THE process of disposition recognition is based on classifiers and on validation methods (cf. Section 1.3) applied to it. Validation is an important issue to assess the quality of recogniser systems that are founded on classifiers. As motivated in Chapter 1 and discussed in Chapter 3, the system's performance is highly related to the material which was used for training. In particular, the training material's quality further depends on the preparation process. Hence, I will present methods for the annotation of data sets and especially a tool called ikannotate [Böck et al. 2011b] developed by Ingo Siegert and me.

As mentioned, based on annotated data sets suitable classifiers can be trained. In particular, I will concentrate on HMMs and a certain kind of ANNs. Especially, for the HMMs I investigated feature and parameter sets. The development was done on acted material, but as discussed in Section 3.2 the focus of research is on non-acted data material. However, I found that both parameter sets can be transferred from acted to non-acted corpora. Moreover, as a special type of production models μ an additional kind of classifier, namely GMMs, are investigated.

Finally, the fusion of different methods will be discussed since the process of disposition recognition is in fact multimodal. It is assumed that with multimodal approaches the variety of dispositional behaviour could be handled better. Nevertheless, even in a single modality fusion of different kinds of classifiers, for instance, might improve recognition results. As this is not the main focus of this thesis, fusion is introduced only briefly. Corresponding results will be discussed in Section 5.3.

4.1 From the Acoustic to Training Material

4.1.1 The Annotation Process

The annotation of a data set is both a quite challenging issue and quite time consuming. So far, it is done most of the time fully manually by well-trained annotators or experienced non-professionals, this means persons who are familiar with assessing human behaviour like psychologist but are not labellers by profession. Therefore, in the latter case, a larger number of annotators are asked to judge the material and afterwards, the reliability - also called interrater reliability - is computed to evaluate the labelling quality.

To preprocess a given data set for emotion recognition from speech, usually three steps are necessary; namely *transcription*, *annotation*, and *labelling.*
In *transcription* the spoken utterances are transferred to a textual notation. This is purely done by writing down what is said with mispronunciations, elliptic utterances, etc. The second step is the annotation. In general, the term *annotation* is used to define the process of adding prosodic and paralinguistic signs to a text

like in the Gesprächsanalytisches Transkriptionssystem (dialogue analytic transcription system) (GAT) [Selting et al. 2011]. Sometimes timing information are added as well. Finally, further information like emotions and dispositions are appended during *labelling*.
In the speech community usually annotation and labelling are used synonymously.

No matter whether the corpus is labelled by a professional or non-professional annotator, both need a tool for doing this. So far, usually text editors are used or systems which are adapted to professionals, for instance, the tool Folker (cf. [Schmidt & Schütte 2010]). Nevertheless, an emotional, multimodal labelling cannot be done with it; just a transcription and annotation. The formerly can be done with, for instance, Atlas [Meudt et al. 2012], but it does not provide support for neither transcription nor annotation. Hence, we developed a tool which allows non-professionals to transcribe, annotate, and label all in one, given data sets using audio recordings (cf. Section 4.1.2).

Furthermore, labelling is still a time consuming issue. One solution, which is pursued especially in psychology, is to establish a data set where the class labels, for instance, are fixed by experimental design (cf. Last Minute in Section 3.2.1 and EmoRec in Section 3.2.2) or posed by an actor (cf. Section 3.1). Nevertheless, even with given classes the necessity to mark interesting parts of the material is still on hand. This means that, for example, the part where a dispositional behaviour occurs has to be assigned manually. Further, annotating features, in particular, facial ones like AUs in FACS, is so far mainly hand-crafted. For this case, I proposed a framework which is related to a semi-automatic annotation of features to reduce the manual effort (cf. [Böck et al. 2012a; Böck et al. 2013a]). This is discussed in Section 4.1.3.

4.1.2 ikannotate

As I introduced, several tools like Folker [Schmidt & Schütte 2010] and Atlas [Meudt et al. 2012] provide different capabilities in the preprocessing of data material. Unfortunately, to handle material in a continuous way, that is, having the same output formats etc., none of those systems is appropriate. Based on the idea of such an continuous handling Ingo Siegert and I developed a tool called ikannotate. It is released, published at the ACII 2011 in [Böck et al. 2011b], and also demonstrated at the ICME 2011 [Böck et al. 2011a]. As this thesis not mainly deals with annotation as such, I will introduce the tool briefly, only. Nevertheless, it was used to prepare several experiments which are referred to.

The main advantage of ikannotate is that for each processing step the same data structure, namely XML, is used to store the relevant information. This is done on utterance level which provides the opportunity to extract any information according to a certain utterance by parsing only one document and getting those in a compact form as they are handled en bloc. Therefore, sorting functions and statistical analyses can be performed easily. Technical details can be found in [Böck et al. 2011b].

Transcription

The transcription is done on utterance level based on the audio material of the data set. This is reasonable because dispositions usually change slowly (cf. Definition 1.5 on page 7 and Definition 1.6 on page 7), that means, one utterance is long enough to be covered by one disposition. Each sentence is therefore the base unit for the next steps in the process of data preparation, namely annotation.

To do the transcription the operator enters the spoken utterance which will be automatically stored. Further, the sentence is assigned to a speaker which provides the possibility to extract separate information for each participant of the interaction. Moreover, this assignment is utilised in the annotation process to cover and visualise the course of the interaction, or more specifically, of the dialogue.

Annotation

The more important aspect in the preprocessing of a corpus, especially, for the purpose of speech recognition and disposition recognition from speech, is the annotation of the material with prosodic features. Furthermore, paralinguistic characteristics are also added to the transcribed utterances. At the end, an enriched document is generated by the annotator.

To mark paralinguistic and prosodic characteristics various systems are developed by linguists like GAT [Selting et al. 2011], codes for the human analysis of transcripts (CHAT) [MacWhinney 2000], or semi-interpretative working transcript (HIAT) [Ehlich & Rehbein 1979]. In ikannotate the first system is realised which provides a comprehensive covering of the features which occur in HHI, but more importantly, in HMI. Furthermore, the system can be divided by design in three subsets covering different degrees of granularity in the number of features. Those are i) minimal (smallest information), ii) medium (enhanced information), and iii) full (containing all information) (cf. details in [Selting et al. 2011]). Due to this concept the annotation process can either be focused or, after full annotation, the content be reduced towards specialised analyses which need only a few entities.

The main advantage of ikannotate is that the annotator has not to care about the specialised signs which are used and defined to mark the corresponding prosodic and paralinguistic features. These are inserted by selecting the characteristics by clear words. For this, even untrained annotators, or those experienced in other annotation systems, can do the marking.

Labelling

Similar to the annotation process the user is supported while labelling; that means he can select the (sub)classes by clicking at the corresponding item.

According to the labelling systems which are discussed in Section 1.2.3, ikannotate comes with an implementation for each of these systems. For the categorical approach the Basic Emotions according to [Ekman 1992] are realised. As mentioned in Section 1.2.3 the GEW proposed by Scherer (cf. [Scherer 2005]) as a good example for quasi-continuous labelling. Thus, in ikannotate it is implemented, as also displayed in Figure 1.2 on page 11. Finally, SAM is used for continuous labelling. For this, the method according to [Bradley & Lang 1994] gives the possibility to label in PAD space. This is also discussed in [Grimm et al. 2007]. Of course, the self-rating is just a part of the system's power. Moreover, it can be used to assign external ratings as well. Details of the implementation are given in [Böck et al. 2011b].

Additional Features

So far, I introduced the main components of ikannotate. Further, the tool provides additional features which are helpful to get a meta-analysis of the corpora.

As discussed in [Stefan et al. 2010] the distribution of an emotional state or a disposition over the utterance is an additional source of information. To reflect this fact, the labeller can mark the maximum of the (emotional) reaction. This is done on word level. Nevertheless, the range of the maximum can be extended to several words. According to [Stefan et al. 2010] only one maximum per utterance is reasonable.

The second add-on is the self-rating of the labeller according to his certainty about the labelling. In every labelling step the user is asked to assign a degree of uncertainty to the label given in this step. This has two advantages: First of all, the assessments can be evaluated and finally, ranked taking into account the uncertainty. Further, especially, in the fusion of those decisions (cf. Section 4.4) the degree of uncertainty is an important element the process. It influences the way a result, for instance, in the combination with Dempster-Schafer Theory (DST), is achieved. To the best of my knowledge, ikannotate is the only tool which provides this feature, yet.

Additional sub-tools are integrate in ikannotate; but due to the focus of this thesis I will skip the explanation here and refer to [Böck et al. 2011b] for further information.

4.1.3 Semi-automatic Annotation

As I introduced, the process of annotation as well as labelling is a quite time consuming task that is therefore also quite expensive. To get a feeling for it I will give a brief example: the TableTalk corpus (cf. Section 3.3) was already preprocessed with a Face Tracker but neither with prosodic and paralinguistic nor dispositional markings. As I was interested in the involvement this has to be labelled to explore the data set accordingly. Several labellers – in general, more than two – are necessary for valid markings. They have to be recouped for their effort. Moreover, the process as such took roughly 80 man-hours for 34:34 minutes of recordings. This examples shows the importance of a support in the procedure of data preparation to reduce the manual effort.

The ultimate goal would be to get a system which provides a fully automatic preprocessing as discussed in Section 4.1.2 of a corpus, for instance, the EmoRec data set. So far, such a system has not been implemented. In [Böck et al. 2012a] and [Böck et al. 2013a] I introduced a framework for a semi-automatic annotation of AUs in FACS. For this, relevant video sequences are automatically marked based on acoustic classification. The framework will be introduced in this Section and results are presented in Section 5.2.2.

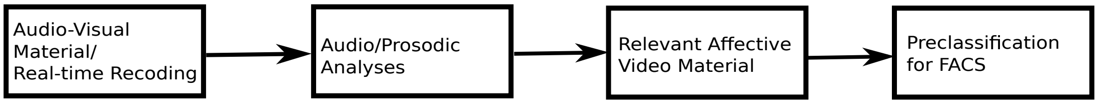

Audio-Visual Material/ Real-time Recoding
Audio/Prosodic Analyses
Relevant Affective Video Material
Preclassification for FACS

Figure 4.1: Workflow of the framework for a semi-automatic annotation (cf. [Böck et al. 2013a]). It is based on audio, especially, prosodic features to identify relevant sequences in a video stream.

The idea was inspired by forced alignment which is a common technology in speech recognition and the work by Looze et al. who used prosodic features for dynamic analyses of mimicry [Looze et al. 2011]. The main workflow is visualised in Figure 4.1 where each block is a successive step towards AU preselection or if possible a preclassification of facial expressions. The preselection is the goal I aimed on, however in the future a preclassification might be possible, though it strongly depends on the processed data set. A more pessimistic estimation is that the preselection is more likely.

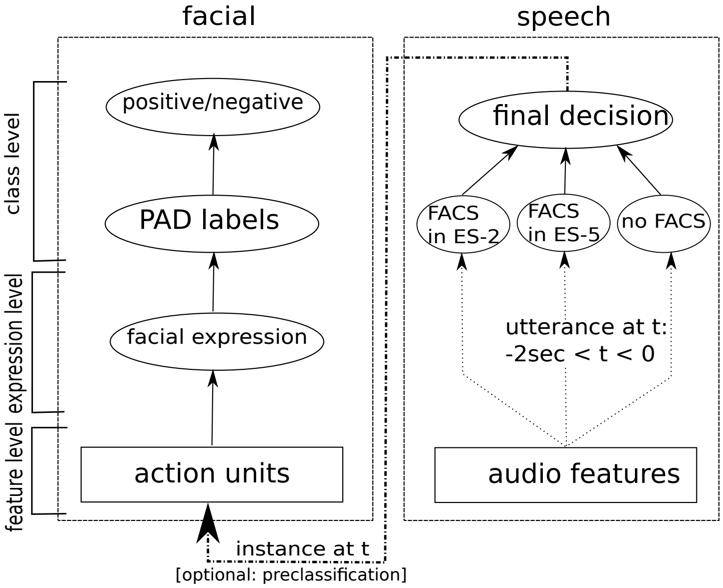


Figure 4.2: Flow of features in the framework for a semi-automatic annotation of EmoRec (cf. [Böck et al. 2013a]). Based on audio features a categorisation from speech's point of view can be given. At time t this decision is given to the facial part. Using this information a hierarchical structure to get classes can be established. In the speech analysis a preclassification of the utterances is possible, marking these if they are related to facial expressions (FACS) in an experimental sequence (ES).

Based on the features (cf. Section 4.2.2) which were determined on non-acted material (cf. Section 3.2), in particular, on the data set EmoRec (cf. Section 3.2.2), the audio material of a corpus is evaluated. If the speech sample is identified as related to a dispositional event the audio time is logged. At this point, I do not distinguish between emotional and real dispositional events. Since I assume synchrony in time throughout the different recorded modalities, in particular, audio and video, the time stamp of an audio event is equal to this in the video channel. Especially, in Last Minute (cf. Section 3.2.1) and EmoRec (cf. Section 3.2.2) the assumption is true (cf. [Frommer et al. 2012b; Walter et al. 2011]). The so extracted time information can be used to indicate ranges in the video sequence where emotionally/dispositionally relevant AUs might occur. The annotator has just to assess a short sequence instead of the whole video which reduces the time effort. In Figure 4.2 on the previous page details of the framework in the sense of features can be found. Based on audio features a classification and further, a final decision is generated. This information is used to indicate sequences which are relevant for analyses of facial expressions. From AUs facial expressions can be constructed and finally, classes can be derived. Optionally, the annotator can be informed with a preclassification based on audio analysis.

In the setup GMM (cf. Section 4.3.1) are used as classifiers. For their training only those speech samples are used which are less than two seconds before an AU (cf. Figure 4.3 on the facing page). The audio material is grouped whether it belongs to an ES or not (cf. Figure 4.2) and from this corresponding GMMs are trained. As discussed in Section 1.2.1 and from Definition 1.2 on page 5 emotions are short events and therefore, I selected this threshold (cf. [Koelstra et al. 2009] as there is also a correlation between emotional reactions and the measurement of those in electroencephalogram (EEG)). Indeed, doing such semi-automatic annotation for dispositions an appropriate threshold has to be determined.

Common systems which provide support for FACS coding are based on the analysis of faces in video. So, why do we have to switch to another source of information? In several cases these systems fail because of non-perfect recording conditions like glasses, fringes, or sensors. Further, they are also influenced by the lighting conditions. All of these disadvantages can either be considered while collecting the data, or they will never change. As discussed in [Böck et al. 2013a] a feature is selected which is independent from such factors; namely speech.

Pure video-based FACS coding yields a hit rate, which is a measure for correctly annotated AUs, of roughly 76% by manual marking [Limbrecht-Ecklundt et al.

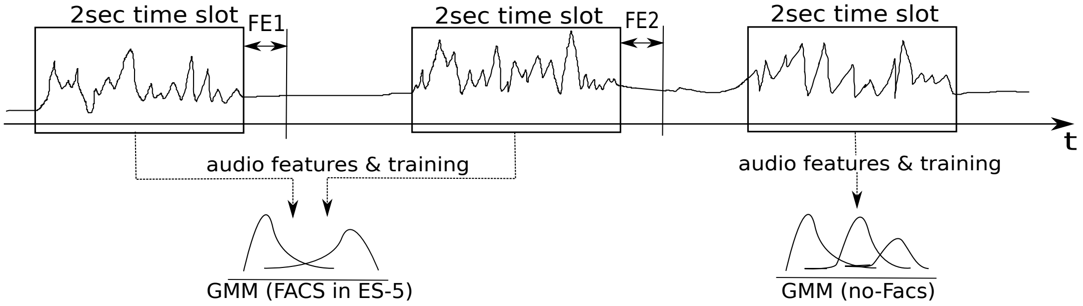


Figure 4.3: Course of a speech sample marking the occurrence of two facial expressions (FE). Only those samples are used in training that are less than two seconds before the event; in this figure exemplary visualised with ES-5. In case of no facial expression the material is used to train a GMM for the class belonging neither to ES-2 nor ES-5 (cf. [Böck et al. 2013a]).

2013]. Additional indicators help to guide the annotator through the process as they focus his attention to relevant sequences. In manual annotation, a lot of concentration is spent on material and sequences not containing any significant information. For this, reducing the manual, boring effort might improve the hit rate. So far, this is a matter of research and has to be proven in larger contexts with several independent annotators (cf. Section 7.2.1).

4.2 Features

In this Section only features applied for recognition tasks and semi-automatic annotation are introduced. An overview of extracted features is given, for instance, in [Vlasenko et al. 2008; Schuller et al. 2009a] and I am aware of the multitude of features which can be used in the processing of speech. Nevertheless, it was one of my research interests to show the potency of small feature sets to obtain recognition systems which are based on low level computational complexity. With several features like global means (cf. e.g. [Schuller et al. 2009a]), movement in contours (cf. [Vlasenko et al. 2012]), or more general, statistical features, this is not easily feasible as statistical information has to be gathered over time. Of course, if small periods are used those characteristics can be applied as well. As discussed in Section 4.3.2 and shown in [Böck et al. 2010] even small subsets can lead to respectable results.

I will explain and characterise the extracted features as it is necessary for my work. Cited references will give information in more details. Further, in Section 4.2.1 mainly the acoustic features are introduced whereas in Section 4.2.2 additionally prosodic methods are discussed which were only used in semi-automatic framework.

4.2.1 Emotion and Disposition Recognition from Speech

Relying on measurements based on the speech signal, acoustic features are those which mostly describe the characteristics of a voice. In contrast, prosodic (cf. Section 4.2.2) and statistical features are high-level features (cf. [Schuller et al. 2010a]) derived from expressions and functionals. Acoustic features are also called Low-Level Descriptors (LLD) (cf. [Schuller et al. 2010a]) evolving and changing over time. To incorporate the features' development in time usually derivatives, especially the first and second derivatives, are added. These are known by Delta (D) – first derivative – and Acceleration (A) – second derivative – representing the differences between two successive speech samples' characteristics. This trick has to be used for classification methods like HMMs (cf. Section 4.3.1) that work on a short-time context. In contrast, advanced ANNs, namely SRNs, can do this by design (cf. Section 4.3.3).

In the context of acoustic feature extraction, usually, windowing techniques are applied that allow to handle the speech signal in a short-time context. In speech recognition typically a Hamming window (cf. Equation 4.1) is used. Let n be the input's index and N is the window size, a Hamming window is calculated as follows

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{4.1}$$

Furthermore, an adaptation towards an equal loudness is done for the speech samples to avoid classification side-effects based on different levels of loudness. For this usually a log function is applied to the signal.

In the following, I introduce in detail three acoustic feature sets related to my experiments. The LPC and PLP were investigated amongst others in [Böck et al. 2010] and Section 4.3.2. There I found that those are not as predictive for general aspects of emotion and further, disposition recognition, as MFCC. Therefore, I concentrated on the latter in further experiments. This is also motivated by the general use of MFCC in speech and emotion recognition, for instance, shown by

[Vlasenko & Wendemuth 2009; Hübner et al. 2010; Muda et al. 2010; Schuller et al. 2011c].

For the purpose of introduction, I refer to the following books [Rabiner & Juang 1993; Gold & Morgan 2000; Wendemuth 2004] that are mainly used as foundation. If I differ from these, I will give the appropriate reference at the corresponding position.

Linear Predictive Coding

Well-known in communication technology, the LPC is an approach reducing the amount of data which has to be transmitted via any kind of channel. Especially, in mobile telephony only the necessary characteristics of a voice are sent to the receiver where the corresponding acoustic is finally generated again. LPC is a simple modelling technique relying on the source-filter model by Fant [Fant 1960]. The important feature is the vocal tract's characteristic which varies for each human and each spoken item. This is predicted and afterwards, only the so estimated coefficients are transmitted.

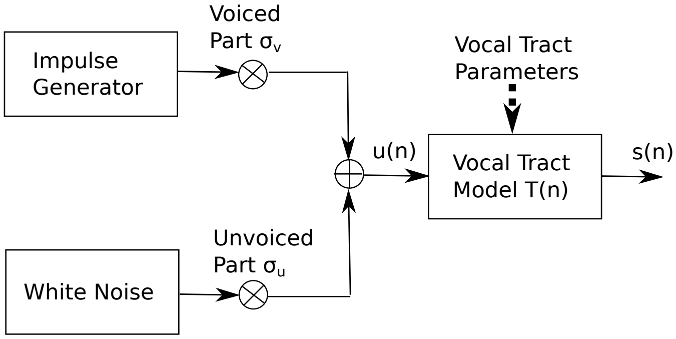


Figure 4.4: Standard linear model of speech production (cf. [Wendemuth 2004]) which is divided into a voiced and an unvoiced part of speech. The vocal tract parameters are estimated by LPC. In addition, the model holds in parts for speech recognition, too; namely the modelling of the vocal tract.

The functionality of the source-filter model as well as its components is given in Figure 4.4 according to [Wendemuth 2004]. The system in Figure 4.4 is used to model the speech production process. The important part is the vocal tract model representing the characteristics of human's speech apparatus neglecting glottis, nasal tract, and lip's radiation. Vocal tract characteristics can be used to describe the voice and further, the content of the utterances themselves.

The speech signal $s(n)$ is computed by an excitement $u(n)$ and the transfer function $T(n)$. Furthermore, $u(n)$ is amplified with σ according to the voiced part (cf. Figure 4.4 on the preceding page). For this, the speech signal can be written as

$$s(n) = \sigma u(n) + \sum_{i=1}^{p} a_i s(n-1) \tag{4.2}$$

In Equation 4.2, the coefficients a_i are included which are the vocal tract parameters; in the LPC they are called predictor coefficients. After Z-Transformation the transfer function can be written as

$$T(n) = \frac{S(z)}{\sigma U(z)} = \frac{1}{A(z)} \tag{4.3}$$

with $A(z) = 1 - \sum_{i=1}^{p} a_i z^{-i}$. As in both Equations a_i's are utilised, they are estimated by minimising the error function solving $\frac{\partial E}{\partial a_i} = 0$ with

$$E = \sum_{n=0}^{N-1} \left(s(n) + \sum_{i=1}^{p} a_i s(n-i) \right)^2 \tag{4.4}$$

where N is the length of the sample in time. This results in a covariance matrix based system of equations. For this, the coefficients a_i can be estimated by either a covariance or an autocorrelation approach.

To obtain the LPC spectrum, the coefficients a_i are concatenated in a vector and a Discrete Fourier Transform (DFT) is applied with a length of N. If the number of coefficients is smaller than N, zero padding is necessary. After this procedure the formants of the certain voice can be observed. Formants are an important and characterising feature of the speech signal.

Perceptual Linear Predictive Coefficients

The main idea of this feature extraction approach is to model the human auditive perception. Therefore, a mel scaled spectrum is applied. From this, the coefficients can be calculated from the error which occurs between both series.

The main steps in calculation of the PLP features are as follows:

1. compute the Fast Fourier Transform (FFT) of the speech signal
2. do a critical-band integration and resampling

3. preemphasise the spectrum according to an equal-loudness curve
4. apply the power law of hearing
5. compute the Inverse Discrete Fourier Transform (IDFT)
6. apply Levinson-Durbin Recursion to solve the linear equations' system

To calculate the power spectrum usually the FFT is computed on the speech signal. Before doing so, the signal is preprocessed with a window, for instance, a Hamming window (cf. Equation 4.1). Such spectrum is integrated in band filters based on a frequency scale that is called the mel scale which has a twofold characteristic: below 1kHz it is roughly linear and otherwise, logarithmic. In contrast to LPC, this analysis is oriented on the human's way of perceiving sounds and speech. The integration is done with triangular or, in case of PLP, trapezoidal windows applying the warping function of Schroeder (cf. Equation 4.5) where the frequency ω is given in radians per second.

$$\Omega(\omega) = 6ln\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{\frac{1}{2}}\right\} \tag{4.5}$$

After adapting the signal to an equal loudness the spectral amplitudes are compressed. In PLP the cube root is used to handle this aspect. This is also known as the power law of hearing. By doing so, on the one hand, the human's sense of hearing is reproduced, and on the other hand, amplitude variations are reduced. For PLP, the IDFT leads to coefficients which are similar to autocorrelation ones as an autoregressive modelling is applied to the signal. It should be noticed that the values of the power spectrum are all real and even. Therefore, it is only necessary to consider the cosine components. To smoothen the resulting signal, in PLP an autoregressive model is deployed to the compressed spectrum whereas no additional filtering is applied. In this context, this model is derived from linear equations which are constructed on the basis of autocorrelations of the previous time step. This leads to a better noise robustness as it can be achieved by the method of cepstral transformation.

In addition to this standard PLP approach, Hermansky et al. invented RASTA-PLP [Hermansky et al. 1991] which has a more robust characteristic against convolutive disturbance and further, can be applied online. As I did not used RASTA-PLP in any experiments I do not introduce this method, and just refer to, for instance, [Hermansky et al. 1991].

Mel-Frequency Cepstral Coefficients

The main aspect in the computation of MFCC is the separation of the excitation and the resonance frequencies.

Like in the description of PLPs I sketch the method to compute MFCC first.

1. window the input signal
2. apply any kind of Fourier Transformation
3. compute the absolute spectrum and do a mel frequency warping
4. apply a logarithmic function to this spectrum
5. reduce the wave band by mel scale and utilise rectangular filter banks
6. compute the Cosine Transformation to obtain the cepstrum

As in PLP the speech signal is first windowed by, usually, a Hamming window (cf. Equation 4.1). To compute the spectrum a Fourier Transformation is applied. It is to be noticed that for further computation only the absolute amplitude spectrum is used. Before the logarithmic function is applied a mel frequency warping is done to adapt the signal to the auditory perception of humans. For this, the signal is warped by Equation 4.6, where f is the current frequency.

$$\mathrm{mel}(f) = 2595 \lg \left(1 + \frac{f}{700}\right) \tag{4.6}$$

With the mel scale the abscissa is, on the one hand, transformed from a frequency in Hz to the mel scale in mel and, on the other hand, compressed in its range (cf. [Wendemuth 2004]). Afterwards, an additional filtering with triangular filters is carried out; each one according to a fixed band structure. Finally, as proposed in [Davis & Mermelstein 1980] the cepstrum of a speech signal – that is $\log(\mathrm{FFT}(f))$ which is called cepstrum – is computed by applying the Cosine Transformation on the mel scaled logarithmic spectrum. This is done according to the following equation (cf. Equation 4.7)

$$\mathrm{c}(q) = \sum_{m=1}^{M} \mathrm{mel}(k) \cos\left(\frac{(2m+1)\pi q}{2M}\right), \quad q = 1, \ldots, \frac{M}{2}, \tag{4.7}$$

where M is the sequence length and $\mathrm{mel}(k)$ is calculated as in Equation 4.6.

Usually, after computing the standard MFCC the parameters can be adapted to the characteristics of a speaker. Especially, the Vocal Tract Normalisation (VTN) is a common procedure to eliminate the differences in the vocal tracts of various speakers. This can be also done for a group of speakers which means that an adaption towards a group's vocal tract characteristic is arranged.

Skipping VTN, mel frequency warping, and filtering Molau et al. are presenting an approach to compute MFCC directly on the power spectrum [Molau et al. 2001]. In my experiments I am usually using the HTK developed at the Cambridge University (cf. [Young et al. 2009]) to extract the MFCC.

4.2.2 Semi-automatic Annotation

The framework of semi-automatic annotation is presented in Section 4.1.1 where the idea of reducing the manual effort to assign FACS units is discussed. So far, this framework is universal; that means it can be used independent of a corpus. The features I present in the current Section are suitable to achieve the aim to be applicable universally. However, the optimisation of the framework's features is still a matter of research. Nevertheless, respectable results were already achieved (cf. Section 5.2.2). Up to now, the experiments are based on the EmoRec corpus introduced in Section 3.2.2.

In the framework, I concentrated my research on the usability of prosodic features, where those used for the training of the classifiers applied in the framework will be introduced in the following, to give indicators for sequences which are worth to look at from an annotator's point of view. In Section 5.2.2 I will discuss the results of this approach, but let me mention some general findings. With these prosodic features I got quite respectable results for sequences in the EmoRec data set [Böck et al. 2013a]. Especially, in the ES-5 facial expressions could be identified. With the counterpart, namely ES-2, the recognition was more difficult due to the small amount of training and test material.

In contrast to, for instance, Björn Schuller who is heavily arguing for large feature sets (cf. [Schuller et al. 2009b; Schuller et al. 2009a; Schuller et al. 2010a; Schuller et al. 2011b]), I argue the other way around. From my point of view, it is more important to find meaningful, prosodic features that give a key of understanding how dispositions are perceived. This is more or less a bottom up process. Fortunately, the engineering community is not alone in analysing features. Psychologists like Klaus Scherer are investigating conversations and interactions under various aspects (cf. also Section 1.2.2). Influenced by [Scherer 2005; Schuller et al. 2008b] I selected a subset of prosodic features which were found to be meaningful in linguists' analyses; namely formants, bandwidth, pitch, intensity, and jitter. Usually, speaking rate and number of pauses are suitable paralinguistic features, too. But, due to the data set that has a command style those are to be neglected as the speakers are too focused on the task.

In the following, I introduce these prosodic features that are applied in the current framework and further, used in the experiments (cf. Section 5.2.2).

Formant and Bandwidth

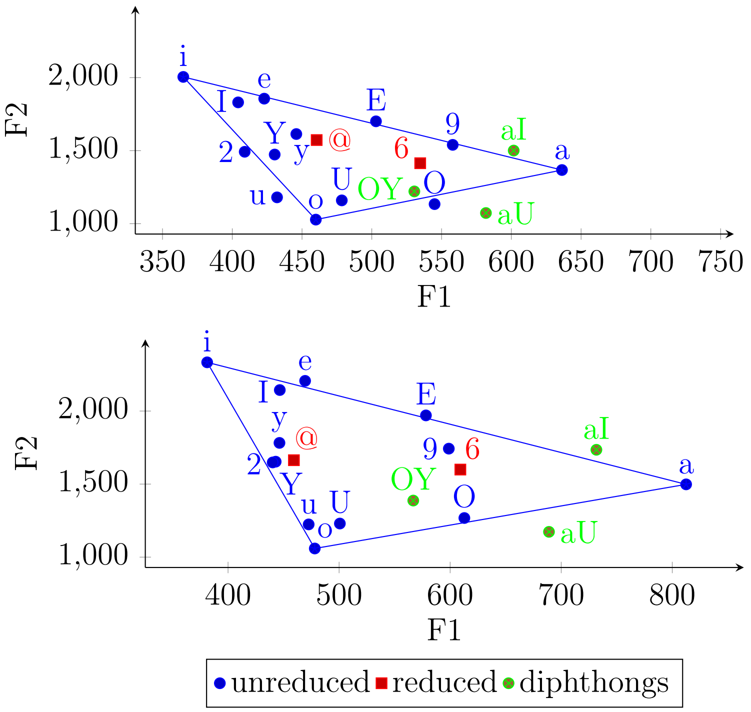


Figure 4.5: Vowel triangle in F1-F2 space (in Hz) for male (top) and female (bottom) speakers adapted from [Vlasenko et al. 2011b].

Formants are defined as “the spectral peaks of the sound spectrum” by Fant (cf. [Fant 1960]) and therefore, are computed based on LPC (cf. Section 4.2.1). With this method the characteristic features of the vocal tract can be evaluated by calculating the coefficients a_i of the $s(n)$ (cf. Equation 4.2) where the formants can be estimated by finding the peaks in the LPC filtered signal. The spectral peaks can be extracted after the DFT was applied. A common method to extract formants from a speech signal is the Burg algorithm [Burg 1975]. Burg relies heavily on the stationarity of the current input signal. And indeed, one can assume that for short time intervals the speech signal is stationary. The analysis of formants is usually applied on vowels as due to the voiced characteristic they are proper to investigate the vocal tracts characteristic. In fact, formants are heavily related to the vocal tract (cf. [Gold & Morgan 2000]) as they represent the harmonics of the vocal tract, which are the resonance frequencies given in certain frequency ranges (cf. [Gold & Morgan 2000]).

Analysing the formants whose typical values for vowels, namely for English speaker, are given in Table 4.1, F1 and F2 are related to the identification and hence, understanding of a vowel. Therefore, they are important to carry an information. In contrast, F3 and higher formants describe the sound of a voice, mainly its timbre (cf. [Bross 2010]). Besides this, formants reflect the gender and age of a speaker and further, they are also influenced by the emotional and dispositional speaker's state [Vlasenko et al. 2012]. Due to the wide range of variabilities in formants they are usually visualised in a vowel triangle in the F1-F2 space (cf. Figure 4.5 on the preceding page). Especially, when observing emotional reactions of a speaker a shift of formants can be seen (shown by [Scherer 2005; Goudbeek et al. 2009] for HHI) even in HMI (cf. e.g. [Vlasenko 2011]).

Table 4.1: Mean frequency values in Hz of the first (F1) and second (F2) formant of English vowels according to [Fry 1992].

Vowel	**F1**	**F2**
iː	300	2300
i	360	2100
e	570	1970
a	750	1750
aː	680	1100
o	600	900
oː	450	740
u	380	950
uː	300	940
ʌ	720	1240
əː	580	1380

The bandwidth of a formant is also known as frequency range. It is defined as the range which is in the frequency interval of a formant with −3dB of the formant's power. A graphical interpretation is given in Figure 4.6.

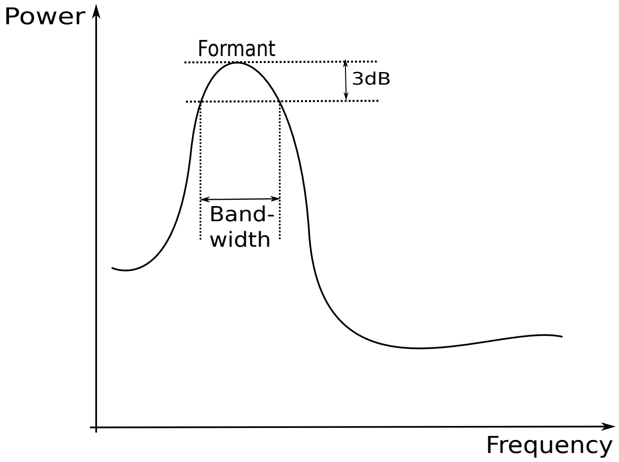


Figure 4.6: Graphical representation of the formant's bandwidth.

Intensity

The intensity is measured by using the amplitude of the sound wave. The higher the amplitude, the louder the sound; that is, the higher the intensity. In other words, the intensity value is the ratio of the signal's power, which can be seen as the energy E in a given time interval, and the area effected by the wave. Hence, the intensity I is calculated as

$$I = \frac{E}{\text{Area}}. \tag{4.8}$$

In Equation 4.8 Area is the variable referring to any kind of area and thus, the particular equation of area calculation has to be filled in. Usually, the area is defined by the system which is used to measure the intensity. In a biological sense, this is the system of the internal ear, whereas technically it is the membrane of the microphone. Further, from Equation 4.8 the unit of I can be given which is $\frac{\mathrm{W}}{\mathrm{m}^2}$.

With the standard way of extracting intensity, small values are obtained and no relation to the human's sense of hearing is given. To overcome this effect the intensity level IL is defined as follows

$$IL = 10 \log \frac{I}{I_0}. \tag{4.9}$$

With the logarithmic function the connection to the hearing behaviour of humans is introduced. Further, the intensity level is normalised to the threshold of hearing which can be realised with human ears; that is $I_0 = 10^{-12}\frac{\mathrm{W}}{\mathrm{m}^2}$. The unit of IL is defined as dB.

The main disadvantage of the intensity is that the acoustic pressure is reduced quadratically with the distance. From this, $I \sim \frac{1}{r^2}$ highly depends on the distance r between the speaker and the microphone recording the signal. Thus, the value of the intensity is not equal all the time since in naturalistic environments the distance between a moving speaker and the microphone could not be fixed. Hence, in naturalistic recordings intensity usually varies much since the distance cannot be controlled satisfactorily as it is possible in acted recordings.

Pitch

The detection of pitch is indeed the estimation of the fundamental frequency F0. It has been seen as a difficult topic in audio signal processing for several years, but can be considered as solved (cf. [Rabiner et al. 1976]). In particular, there are fortunately different methods for pitch detection in speech analysis available. As there is a large community which is dealing with F0 estimation several methods are on hand. Therefore, I introduce those briefly and refer to [Gerhard 2003] for details, which provides a basic overview.

> Since pitch is a perceptual quantity related to F0 of a periodic or pseudo-periodic waveform, it should suffice to determine the period of such oscillation, the inverse of which is the frequency of oscillation. [Gerhard 2003]

The previous statement is a description of pitch as well as a kind of definition. From it, the way of computing the pitch value can also derived. The methods of pitch detection can be divided into three main classes: i) time domain, ii) frequency domain, and iii) statistical methods.

The time domain methods are further subdivided in *event rate detection*, *correlation*, and *phase space* approaches.

> The theory behind [*event rate detection*] methods is that if a waveform is periodic, then there are extractable time-repeating events that can be counted, and the number of these events that happen in a second is inversely related to the frequency. [Gerhard 2003]

Each method is looking for specific kinds of events in the time domain, such as the following ones:

- **Zero crossing rate.** The main idea is that the information of pitch is based on the specific spectral content of a waveform. Especially, the number of zero crossing events of the waveform itself – that is, counting how often the waveform is crossing the zero per time unit – is an indicator for the characteristic of the waveform and thus, for the fundamental frequency.
- **Peak rate.** It counts the number of peaks per second given in the waveform. In fact, only the positive peaks are considered and from these, the frequency of the waveform is estimated. Otherwise, the method is similar to zero crossing.
- **Slope event rate.** Assuming that the waveform is periodic, the corresponding slope will be periodic, too. Hence, the information can be extracted analogous to zero crossing or peak rate. Using the waveform's slope instead of the original waveform can be more informative or may lead to more robust detection of the events.

The *correlation* in the time domain approach is implemented, for instance, in Praat (cf. [Boersma 2001]). There are three main methods which use the time delay of the signal given by succeeding frames of the waveform whereas the correlation is a measure of similarity between two waveforms.

- **Autocorrelation.** The autocorrelation s_{XX} uses the same waveform and computes the similarity between two frames shifted in time. In Equation 4.10 s_{XX} is computed as defined in [Wendemuth 2004]. For this, the $x[n]$ is the input sequence and ergocity is assume.

$$s_{\mathrm{XX}}(\kappa) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{+N} x[n]x[n+\kappa]. \tag{4.10}$$

- **Cross-correlation.** In contrast to the autocorrelation, cross-correlation measures the similarity between two different sequences or waveforms, respectively. In general, it is computed similar to Equation 4.10 as follows (cf. [Gerhard 2003])

$$s_{\mathrm{XY}}(\kappa) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{k=-N}^{+N} x[n]y[n+\kappa]. \tag{4.11}$$

where $x[n]$ and $y[n]$ are two correlated sequences. By this, the first peak of s_{XY} refers to the period of the waveform.

- **Yin estimator.** This approach is looking for a balance between autocorrelation and cancellations. Therefore, it introduces a difference function (cf. Equation 4.12 and [Gerhard 2003]) to minimise the differences between waveforms. The method was developed by de Cheveigné and Kawahara (cf. [Cheveigné & Kawahara 2002]).

$$d(\tau) = \sum_{i=1}^{N} (x[j] - y[j+\tau])^2 . \tag{4.12}$$

With the *phase space* a short-time history of a given waveform can be investigated. From this, cycles which are repetitive can be observed where the phase space is a plot of the waveform against its slope at time t.

Another approach to extract F0 is to use the frequency domain instead of time domain. It is divided as well in three subcategories describing how to handle the task: i) filter-based methods, ii) cepstrum analysis, and iii) multi-resolution methods.
The filter-based methods apply different filters to the signal having different centre frequencies. The result is the comparison of all filters where the one that has the best match of a spectral peak and the centre frequency gains the highest value. Two well-known realisations of the approach are the optimum comb filter, applying "many equally spaced passbands" [Gerhard 2003], and the tunable filter where the user influences the centre frequency to narrow it.
The cepstrum analysis investigates the Fourier transform of the logarithm by the magnitude spectrum of the waveform. Thus, the number of relevant peaks with a corresponding quefrency; that is the logarithmic frequency, is reduced. The highest peak is assumed to be the fundamental frequency.
Finally, the multi-resolution method combines outputs of an algorithm on higher and lower resolutions or larger and smaller time windows to improve the estimation of F0.

Statistical methods as Neural Networks and Maximum Likelihood Estimators are used but are not as widely spreaded as the others. Therefore, I just refer to the literature, for instance, [Gold & Morgan 2000; Gerhard 2003].

Jitter

Jitter and shimmer are acoustic characteristics of the voice which are given in the signal (cf. [Farrús et al. 2007]). Temporal fluctuations in a signal are called jitter, whereas shimmer is related to the variation of the amplitude; namely the vibration of the voice. Both features are connected to the dispositional state of a user since they were seen as an indicator for negative emotions, in particular, fear and sadness (cf. [Scherer 2001; Schuller et al. 2009a; Looze et al. 2011]). Especially, jitter is related to fear, mainly. Since I found only jitter relevant in my experiments (cf. Section 5.2.2 and [Böck et al. 2012a]) I just introduce this feature.

To compute the jitter value the fundamental frequency F0 has to be extracted, first. This can be done with LPC as it is described in Section 4.2.1. In general, there are several methods to calculate jitter: i) computing it in an absolute or relative manner, ii) as relative average perturbation, or iii) as a five-point period perturbation.

As I used only the absolute method for feature extraction in my experiments I just introduce this. Again, the computation is based on F0 and from that its period length is extracted. Since jitter is the cycle-wise variation of the fundamental frequency it is computed as follows (cf. [Farrús et al. 2007])

$$\text{Jitter}_{\text{absolute}} = \frac{1}{N-1} \sum_{i=1}^{N-1} | T_i - T_{i+1} |, \tag{4.13}$$

where T_i is the corresponding F0 periods' length and N is the total number of extracted F0 periods. Given Equation 4.13 jitter can be interpreted as the average of the absolute differences between two consecutive periods. In the calculation of jitter only the differences – no variances – are considered (cf. [Farrús et al. 2007]). In the same way jitter's extraction is realised in Praat [Boersma 2001].

4.3 Classifiers

In emotion recognition from speech several classifiers are established (cf. Section 2.4.2). In the community the transferring of existing methods to the recognition of disposition from speech is increasingly considered. This thesis will supply approaches which lead in this direction. For my research, I concentrated on the two methods: i) a Hidden Markov Model (HMM) (cf. Section 4.3.1) and ii) a Simple Recurrent Network (SRN) (cf. Section 4.3.3); in particular, a Segmented-Memory Recurrent Neural Network (SMRNN). Both methods are introduced in the corresponding Sections. As mentioned before (cf. Section 2.4.2) other approaches are available, for instance, types of SVMs, ESNs, or DNNs. Since those are not used in any experiments presented in this thesis, I do not introduce these methods.

In addition to the description of classifiers, I present results of experiments to evaluate parameter sets for HMMs (cf. Section 4.3.2). Further, in Section 4.3.4 Recurrent Neural Networks are compared against Hidden Markov Models in terms of performance in emotion recognition from speech which indicates also the usability in disposition recognition from speech.

4.3.1 Hidden Markov Models

An HMM is characterised by a twofold production process where a temporal evaluation takes place and finally, an output is produced. Therefore, any HMM is a finite state automata which can be described by a five tuple $H = (S,K,s_0,A,B)$ where $S = \{s_1, \ldots, s_n\}$ is the set of states, $K = \{k_1, \ldots, k_m\}$ is the output alphabet, s_0 is the initial state of the HMM, $A = \{a_{ij}\},\ i,j \in S$ is the transition probability and $B = \{b_i(o_i)(o_l)\},\ i \in S,\ l \in K$ is the production probability of the model, respectively (cf. [Manning & Schütze 1999]). If every state s_i is also generating an observation, then $i = l$ and the probability of an observation o_i can be written as $b_i(o_i)$. Further, any state may produce a collection of Gaussian mixtures. A visualisation of an HMM is given in Figure 4.7 on this page.

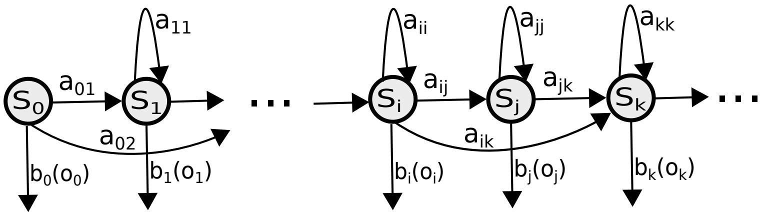


Figure 4.7: Workflow of a left-to-right Hidden Markov Model (cf. [Glüge et al. 2011]).

As HMMs are known in the community of speech recognition and are also investigated to be feasible for emotion recognition, I do not consider all details of the model but refer to the corresponding literature (cf. [Manning & Schütze

1999; Nwe et al. 2003; Wendemuth 2004; Schuller et al. 2009b; Böck et al. 2010; Schuller et al. 2010a; Vlasenko et al. 2012]). Briefly, the training of an HMM is related to two issues: i) finding the probability of a given observation and ii) finding the best state sequence matching an observation.
Assuming that an observation sequence O and a model $\mu = (A, B, \Pi)$ are given, where Π is the set of initial probabilities, the probability $P(O \mid \mu)$ can be computed. Given a state sequence S the observation probability can be written as (cf. [Manning & Schütze 1999])

$$P(O \mid S, \mu) = \prod_{i=1}^{N} P(o_i \mid s_i, s_{i+1}, \mu) \tag{4.14}$$

Using Bayes' Rule and incorporating the initial, transition, and production probabilities Equation 4.14 can reformulated to Equation 4.15 which will be used in classifiers. In fact, by utilising the forward-backward algorithm this probability can be calculated quite efficiently.

$$P(O \mid \mu) = \sum_{S} \pi_{s_i} \prod_{i=1}^{N} a_{i,i+1} b_i(o_i) \tag{4.15}$$

The second issue is related to the question, how to find a proper sequence of states that fit the presented observation. In other words, it is the probability of a state s_i given a sequence of observations O and a model μ; that is $P(s_i \mid O, \mu)$. As it is intended to estimate the best match of states Equation 4.16 has to be solved (cf. [Wendemuth 2004]) where $\mathbf{q}_{\max}$ is the most likely state sequence.

$$\mathbf{q}_{\max} = \arg\max_{\mathbf{s_i} \in S} P(q = s_j \mid o_1 \ldots o_t, \mu). \tag{4.16}$$

This estimation is done with the Viterbi algorithm (cf. [Viterbi 1967]) which evaluates Equation 4.16 efficiently.

Beside the mathematical foundations of HMMs, it is of interest how the parameters of the model can be interpreted in the context of emotion and disposition recognition. The input sequence results from speech; to be precise, the input is a sequence of features (cf. Section 4.2) extracted from speech samples. For each emotion or disposition which should be recognised a model is generated; that is, each model is representing exactly one user's state of emotion/disposition. From this, the observation sequence O is mapped to a sequence of such states. This can be realised by concatenating several observations while an utterance, usually consisting of several words, is processed; this is called word-level recognition. Of course, the recognition can also be done on phoneme-level or utterance-level. In most of my experiments, I did the recognition on utterance-level assigning the output of an HMM to the utterance in total. Hence, $P(O \mid \mu)$ (cf. Equation 4.15) can also be interpreted as $P(D \mid M)$ where D is the sequence of either dispositions or emotions – it is obvious that in the current case $D = O$ – and $M = \{\mu_i\}$ is a set of models where i is the index of models representing the dispositional states of a user. Thus, the probability to observe a disposition given a model is computed. In emotion recognition from speech usually the sequence of states is used to model the characteristics of the emotional speech and hence, each emotion is represented as a single model reflecting the certain temporal characteristics of the specific emotion (cf. e.g. [Vlasenko et al. 2007]). Finally, the model with maximum log-likelihood (related to Equation 4.16) is selected to be the result of the recognition.

Furthermore, three types of HMMs can be distinguished. Firstly, the standard model has a sequence of states and is handled as described above. The second kind of HMM is an one state model. To deal with flexibility Gaussian mixtures are used as model for the single state. This kind of model is called Gaussian Mixture Model (GMM) whereas each Gaussian mixture approximates the characteristics as a Gaussian distribution according to Equation 4.17 with K is the number of components, c is the weight, and g is the density function. An examples of Gaussian mixtures is given in Figure 4.3 on page 67.

$$b_j(\mathbf{x}) = \sum_{k=1}^{K} c_{jk} g_{jk} \mathbf{x} \tag{4.17}$$

The third kind of architecture is a combination of an HMM and a GMM where in each HMM's state Gaussian mixtures are applied as models. Comparing the first two approaches, both are suitable to handle disposition recognition from speech. On the other hand, the multiple states HMM can handle dynamics better than the single one which is focused on the characteristics given by the data and modelled by the mixtures. Details on parameter sets will be discussed in Section 4.3.2 giving remarks and recommendations for training as well as construction of HMMs/GMMs.

4.3.2 Evaluating Parameter Sets for Hidden Markov Models on Acted Material

The experiments as well as the results which are discussed in details are published in [Böck et al. 2010] and further, used through several experiments presented in Chapter 5 which are also published at particular conferences.

I conducted the experiments presented in the following on two data sets; namely EmoDB (cf. Section 3.1.1 and [Burkhardt et al. 2005]) and eNTERFACE (cf. Section 3.1.2 and [Martin et al. 2006]). In [Böck et al. 2010] an additional data set is used; namely the SmartKom Database (SmartKom) [Wahlster 2006]. The experiments on this data set were work of David Philippou-Hübner and hence, I do not in detail report on it in this thesis. The findings will only be used for comparison (cf. [Böck et al. 2010]).

As already discussed in Section 3.1, to evaluate feature sets and recognition methods the community concentrated in the beginning on acted material. For this, a ground truth in the sense of emotional expressions is given. In both corpora, namely EmoDB and eNTERFACE, the utterances are quite expressive and thus, the allocation to several classes is easy. Further, the material is well selected and preprocessed by the distributors. Hence, I could rely on the provided material and investigate the parameters of the HMMs without any influences of possible interference of side-effects caused by realistic material. Both data sets have distinct classes based on Basic Emotions (cf. Section 1.2.3).

Number of Hidden States

In the experiments, the classification labels are given for the whole utterance. Therefore, the temporal evolution of the emotion has to be modelled by the classifier. As discussed in Section 4.3.1 this can be done by using an HMM incorporating several states. They reflect the temporal characteristics of the utterance as well as the emotion. As it is common from speech recognition (cf. [Kwon et al. 2003; Schuller et al. 2008a; Wöllmer et al. 2009]) for each emotion one model μ_i is applied and the final decision is according to the maximum log-likelihood (cf. Section 4.3.1). At first, I investigated the number of hidden states for each HMM. The training and testing of all subexperiments are done utilising HTK developed at the University of Cambridge [Young et al. 2009].

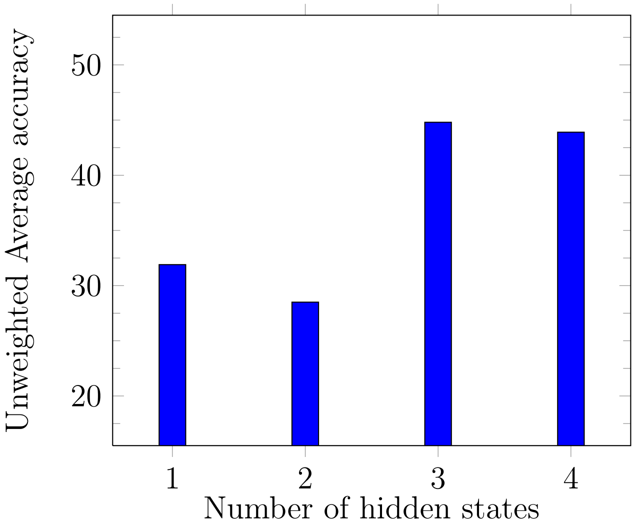


Figure 4.8: Mean Unweighted Average accuracy in percent dependent on the number of hidden states for eNTERFACE having the MFCC_0_D_A feature combination [Böck et al. 2010]. As with more than four hidden states the accuracy is still decreasing, I did not display further values.

As given in Figure 4.8 I inspected the accuracy, to be precise the Unweighted Average accuracy, depending on the number of states. From my point of view, it does not matter whether Unweighted Average accuracy or Weighted Average accuracy is used since I am not interested in an absolute accuracy in sense of comparing results but getting parameters which are suitable to solve the task, Unweighted Average accuracy is feasible for evaluation. The best performance on eNTERFACE is achieved with three internal hidden states. Applying more states, the performance decreased. Using only one or two states did not cover the complexity of the characteristics necessary to classify utterances in total. From this finding, I conclude that reflecting emotional characteristics in whole utterances of near real-life recordings is covered best by three internal states. I did the same survey with the EmoDB. Due to the expressiveness of the emotions no difference was found varying the number of hidden states. The performance kept constantly at 70.9%. For this, I recommend to use a simple classifier setup with acted material; that means, using models with one internal state.

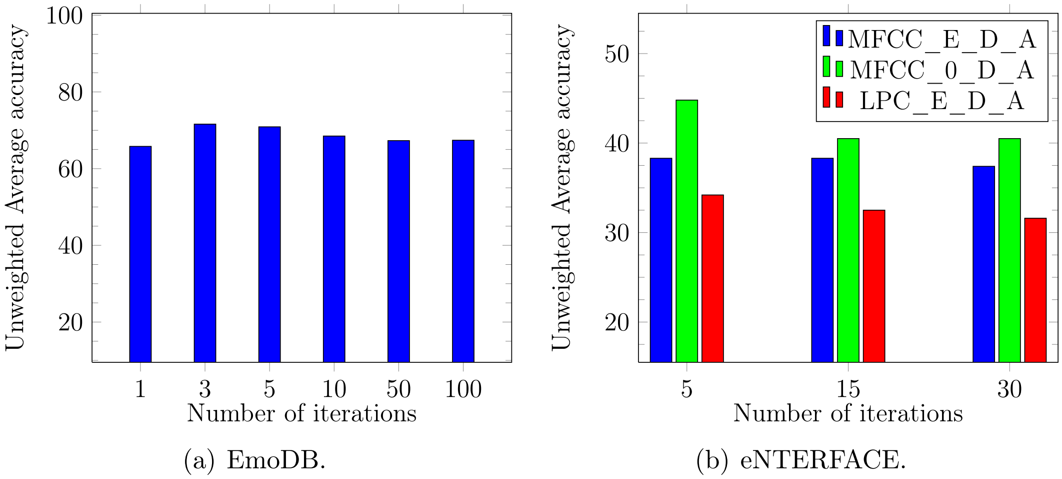
In contrast, having only short statements like ‘C 2’ (cf. EmoRec in Section 3.2.2) the influence of the temporal evolvement is marginal. Due to the shortness of the utterances effects like lexical and semantic influences are suppressed which leads to a focused emotion and disposition characteristic. Thus, HMMs have to be constructed in a different way; usually, they are ‘reduced’ to GMMs. This is investigated in Section 5.1 using non-acted material.

Number of Iterations

Another parameter which influences the performance of HMMs is the number of iterations; that is, how often the training material is presented to a model. One iteration accords one run of the training set in total. In [Böck et al. 2010] I reported that the performance increases if the number of iterations increases. For the training of HMMs a number of iterations is necessary so that the models can converge towards an optimum since HMMs are trained with the Baum-Welch algorithm – a kind of an Expectation-Maximisation (EM) algorithm. If the optimum is achieved, an increase of the iteration’s number results in a decrease of performance since the models lose the capabilities to generalise. This effect is related to the overfitting which might occur in the training of ANNs. These considerations can be seen in experiments (cf. [Böck et al. 2010] and Figure 4.9 on the facing page). As reported in [Böck et al. 2010], for instance, the best results were obtained with five iterations on SmartKom by 50% Unweighted Average accuracy.

I did the same analysis for EmoDB and eNTERFACE (cf. Figure 4.9 on the next page). In comparison to the results of SmartKom I extended the range of iterations up to 100. A subset of the results can be found in Figure 4.9(a) on the facing page for EmoDB. All experiments are done using a three state model and MFCC_0_D_A features (MFCC with derivatives and 0^{th} cepstral coefficient). The decrease in the performance is not as significant as in SmartKom but the maximum is reached with three iterations.

In contrast to this, with eNTERFACE the best performance is obtained with five iterations (cf. Figure 4.9(b) on the next page). Thus I conclude that the more realistic the material is, the more iterations are necessary. In fact, since the characteristics of the emotions are not as expressive, the model needs to process the training material more often to adapt the model’s parameters, especially the transition probabilities A and the production probabilities B. On the other hand, using too many iterations leads to a kind of overfitting. As discussed before, after several training iterations the model converges towards an optimum which results from the EM algorithm. If the HMM is further trained, the model is adapted



(a) EmoDB.

(b) eNTERFACE.

Figure 4.9: Mean Unweighted Average accuracy in percent dependent on the number of iterations of presentations of training material for EmoDB having the MFCC_0_D_A feature combination (cf. Figure 4.9(a) on the current page). In contrast, the results for eNTERFACE (cf. Figure 4.9(b) on this page) are grouped according to three different feature sets, reflecting the same characteristic as with EmoDB (cf. [Böck et al. 2010]).

too much to the training material. This results in a decreased capability in generalisation and hence, in a worse performance.

From Figure 4.9(b) it can be seen that the influence of the iteration's number is given with any kind of spectral features. Due to the characteristics of the different feature sets (cf. Section 4.2.1) the influence is not equal which can be seen in the absolute performance. Further, it is affected by the so-called additional term, which is either the Energy term E or the zeroth cepstral coefficient. Both are investigated in the following Subsection.

I summarise for the number of iterations: For realistic material five iterations should be used to train HMMs. If the material is acted, less number of iterations – in the case of EmoDB, three iterations (cf. Figure 4.9(a)) are suitable to cover the emotion's characteristics. These conclusions can be drawn since the training process and thus, the number of iterations are independent from further characteristics of the material like noise, etc.

Comparing Feature Sets

As most of my experiments are dealing with spectral features (cf. Chapter 5) I concentrate in this Subsection on those feature sets and compare their performance. As stated in Section 4.2.1, usually the standard features are enriched with the first, called Delta, and second, called Acceleration, derivatives. These cover the temporal evolvement of the speech signal. Furthermore, two kinds of additional parameters are assessory to the aforementioned sets: i) the Energy term E and ii) the zeroth cepstral coefficient 0^{th}.
Extracting the Energy term from a speech signal means computing the logarithm of the signal energy (cf. Equation 4.18 according to [Young et al. 2009]).

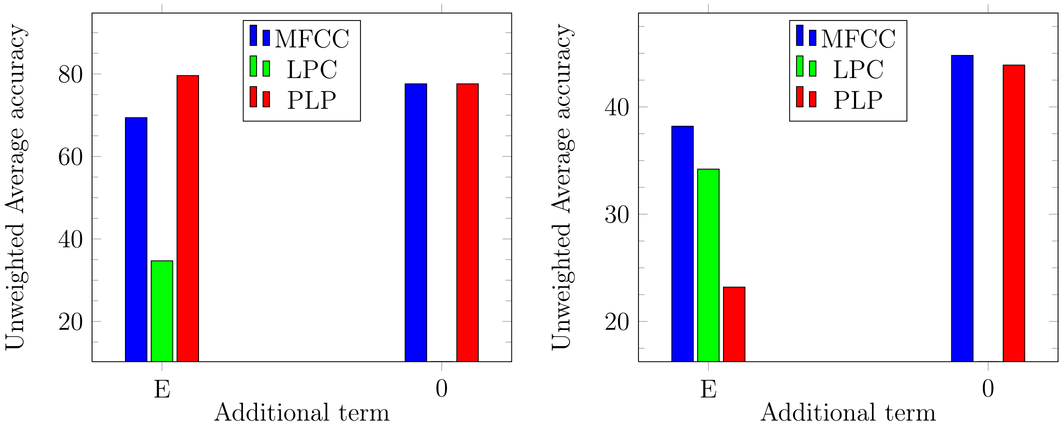
$$E = \log \sum_{n=1}^{N} s_n^2, \tag{4.18}$$

where N is the number of speech samples s_n. It indicates the energy given in the utterance. As already discussed in the introduction of the intensity feature, the energy is depending on the distance between the speaker and the recording microphone. Thus, the energy value varies with the distance and therefore, gives reliable indications for disposition only under controlled conditions.
The zeroth cepstral coefficient is the first cepstral coefficient which can be computed from an audio signal and thus, is representing the basic characteristics of a voice and the speaking style, respectively. For details of the extraction process I refer to Section 4.2.1.

Table 4.2: Feature sets which are compared to be suitable in emotion recognition. The two feature sets are grouped by the additional term which is either Energy (E) or zeroth cepstral coefficient (0). MFCC are the Mel-Frequency Cepstral Coefficients, PLP the Perceptual Linear Predictive Coefficients, and LPC are the Linear Predictive Coding coefficients. D and A are the first and second derivatives of the features, respectively (cf. [Böck et al. 2010]).

Feature sets 1	**Feature sets 2**
MFCC_E_D_A	MFCC_0_D_A
PLP_E_D_A	PLP_0_D_A
LPC_E_D_A	---

In Table 4.2 on this page all combinations of features and additional terms, which I analysed, are listed. The 0^{th} cepstral coefficient reflects the excitation frequency. In LPC the filter coefficients of the transfer function are estimated and thus, these coefficients cannot be directly attributed to the excitation frequency.



(a) EmoDB.

(b) eNTERFACE.

Figure 4.10: Mean Unweighted Average accuracy in percent depending on the feature set grouped by the additional term which is either Energy (E) or zeroth cepstral coefficient (0). In all experiments Delta and Acceleration are used which is neglected in the legend (cf. [Böck et al. 2010]).

As it can be seen from Figure 4.10 on the current page, the performance of the feature sets according to the additional term is quite similar, in general. However, the total number varies between the two data sets. Generally, in acted material (cf. Figure 4.10(a) on this page) LPC is working quite good with both terms and has almost the same performance as MFCC. In non-acted material (cf. Figure 4.10(b) on the current page) the 0^{th} cepstral coefficient is the best choice as additional term. Usually, such kind of data sets are not as expressive and again, the distance dependency of the speaker and the microphone influences the energy values. Thus, the energy gives no further information to the classifier to distinguish a specific emotion or disposition. From this, the performance of MFCC are due to their discriminative power itself. On the other hand, the 0^{th} cepstral coefficient provides additional information of the speaker and his disposition which lead to a better performance.

I further investigated the choice of the feature extraction method. Whereas in acted material LPC has comparable performance to MFCC, in non-acted material MFCC cope better with the characteristics. From my point of view, this is due to the cepstral components of MFCC. These characteristics in combination with the mel scale handle emotional and dispositional speech better and help the classifier discriminating those. In contrast, LPC features lack of these cepstral components (cf. Section 4.2.1) and show a lower recognition performance as, for instance, MFCCs (cf. Figure 4.10 on the preceding page). The same observations can be made having realistic data sets as described in Section 3.2.1 and Section 3.2.2 achieving results presented in Section 5.1.

[Vogt & André 2005] presents also a comparison of feature sets for acted and non-acted data sets, namely EmoDB and SmartKom. Though they end up with a larger feature set, having about 90-160 spectral and prosodic features including also functionals like minimum, maximum, etc. They do not achieve better performance because they got 77.4% accuracy on EmoDB which is similar to my results of 77.6% (cf. MFCC in Figure 4.10(a) on the previous page) using only 39 features. Further, from [Vogt & André 2005] I conclude that the experimental results on SmartKom and eNTERFACE can be reasonable compared: As reported in [Böck et al. 2010] with the aforementioned features and parameter settings a recognition performance of 50.0% was obtained on SmartKom. Vogt & André achieved 40.6% (cf. [Vogt & André 2005] Table 3) utilising their feature set given four emotional classes. Their performance is influenced by the feature selection process which was done with WEKA (cf. [Vogt & André 2005]). In my experiments, I used also eNTERFACE which is a near real-life data set in its characteristics related to SmartKom. Hence, a comparison with the achieved results of Vogt & André is possible. I obtained a recognition performance of 44.8% given the six emotional classes defined by eNTERFACE (cf. Section 3.1.2). Considering these results, a respectable improvement in the performance with less number of features was achieved.

Conclusion

From the experiments presented in this Section to evaluate parameter sets for HMMs I conclude and further, suggested the following setup as also published in [Böck et al. 2010].

For features the best choice are MFCC with 0th cepstral coefficient as additional term. Further, to include the feature's temporal characteristics it is reasonable to add the Delta and Acceleration.
In terms of iteration numbers, it is optimal to train HMMs for five iterations. Having more iterations lead to a kind of overfitting and hence, a loss of generalisation.

Finally, to handle the temporal behaviour within an utterance and therefore, the temporal evolvement of the disposition over the corresponding utterance a three state HMM is the best choice. In contrast, if the utterance is quite short (e.g. as in EmoRec in Section 3.2.2) and thus, no temporal change are on hand that is, it is just one dispositional characteristic observed, a one state HMM can cope with it. In this case, as only Gaussian Mixtures are used, the model is called GMM.

4.3.3 Recurrent Neural Networks

In this Section I introduce a specific kind of ANN which is called SMRNN (cf. [Chen & Chaudhari 2009]). As I did the experiments together with Stefan Glüge I refer to his publications for a deeper look into SMRNN, for instance, [Glüge et al. 2011; Glüge et al. 2012]. These references hold also for the detailed mathematical description of SMRNN. An SMRNN is visualised in Figure 4.11.

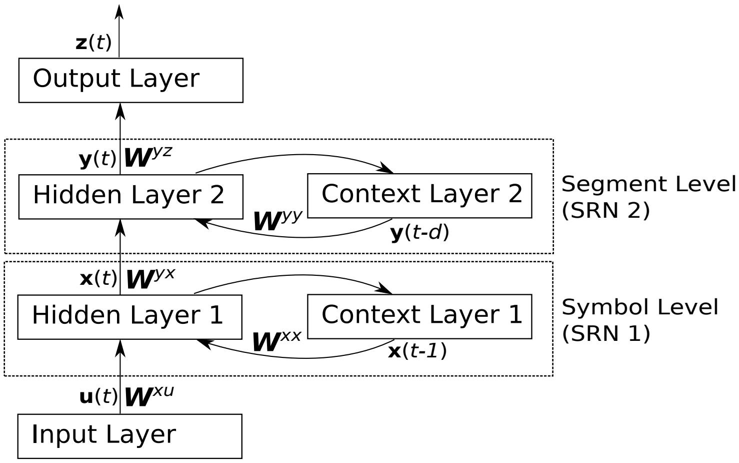


Figure 4.11: SMRNN structure representing a stacked set of SRNs group by the level of process.

The SMRNN consists of two Simple Recurrent Networks which are stacked in a hierarchical way. Each of the subnetworks represents a level of processing. The first SRN is responsible for handling symbols. It gets the input directly from the input layer of the SMRNN. The output of the first SRN is transferred to the second subnetwork and furthermore, to a context layer of the same network. The additional context layer in each subnetwork results from the characteristics of an SRN. The second subnetwork is processing the segment level and is constructed similar to the first SRN. The segment level's output is finally transferred to the output layer where a overall result is generated.

From Figure 4.11 on the preceding page it can be seen that an SMRNN is quite similar to a cascade of several SRNs. The input $\mathbf{u}(t)$ and the output $\mathbf{z}(t)$ are obvious and usually coded by 1-of-N. Further, several weight matrices $\mathbf{W}$ are used to connect the different layers. As it is common, receiver-sender-notation is applied. The matrices towards the context layers are not shown in Figure 4.11 on the previous page since the are equal to the identity matrix $\mathbf{E}$. This means that the output of the hidden layer is directly forwarded to the context layer without any weighting. Furthermore, the content of the context layer is an additional, weighted input to the corresponding hidden layer.

On symbol level the processing of the network is as any standard SRN. The current input is processed with the time-shifted input from the context layer. In contrast, on segment level the context layer's influence is shifted by the parameter d. That means, according to the value of d the context layer sends its content to the hidden layer 2. So, d represents the segment length which is handled on segment level. For this, the output of the segment level is updated only after d time steps. This procedure as well as the additional parameter are the main differences between stacked SRNs and an SMRNN.

SMRNNs can be trained with two training algorithms: i) as proposed in [Chen & Chaudhari 2009] extended Real-Time Recurrent Learning (eRTRL) and ii) extended Backpropagation Through Time (eBTT) by [Glüge et al. 2012]. Both approaches can tackle tasks given to an SMRNN but differ significantly in the computational complexity. According to [Williams & Zipser 1995] the complexity of eRTRL is $\mathcal{O}(n^4)$ in terms of number of adaptable weights in the network. In comparison, eBTT's computational complexity is still $\mathcal{O}(n^2)$ (cf. [Glüge 2013]) which is equal to the original Backpropagation Through Time algorithm (cf. [Williams & Zipser 1995]). Further, as discussed in [Glüge et al. 2013] eBTT benefits from a pre-training of the hidden layers and finally, has a better generalisation performance.

4.3.4 Comparing Hidden Markov Models and Neural Networks

As already introduced, the work on SMRNNs was a collaboration with Stefan Glüge. The following comparison was also done in cooperation with him and is published in [Glüge et al. 2011]. In there, the potential of SMRNNs in emotion and disposition recognition from speech are investigated. The workload for the experiment was distributed as follows: Stefan Glüge did the training and test as well as the tuning of the SMRNN, whereas I did the HMM part and the feature extraction for both classifiers. In this Section I concentrate mainly on the Markov models and not on the SMRNN. Nevertheless, for the sake of comparison the results of both classifiers are given.

Data Set and Feature Sets

From literature, it was so far known that SMRNNs are not used in emotion or disposition recognition from speech. To investigate the capability of SMRNNs in this field, it was decided to look at first on a data set of high quality in the recordings providing significantly distinguishable emotions. Hence, in [Glüge et al. 2011] EmoDB with its clear and expressive emotional utterances was utilised. As introduced in Section 3.1.1 seven emotions are uttered by different actors. The subset of 493 samples was used for the experiments whereas the set was split in training and test sets. For HMMs 90% of the material was randomly selected for training and 10% for testing. In case of SMRNNs the procedure is slightly different. Due to the training algorithm a validation set is necessary defining the end of training. Therefore, the splitting in randomly selected sets is as follows: 80% for training, 10% for validation, and the remaining 10% of the material were used for testing.

As shown in Section 4.3.2, MFCC are extracted as features for the classifiers. Since SMRNNs can handle temporal evolution by design, the temporal components as usually used in speech processing were neglected. Hence, the feature set for SMRNNs is a 13 dimensional feature vector consisting of 12 MFCCs and the 0^{th} cepstral coefficient. To analyse the temporal influence, for the networks the frame rate of feature extraction was varied equally for all samples between 10ms and 25ms applying a Hamming window (cf. Equation 4.1) with 25ms window size. Thus, different numbers of supporting points were generated which directly lead to investigation on the capabilities of SMRNNs to cover temporal characteristics (cf. also [Glüge et al. 2010] for a general discussion of temporal influences in SRNs). The experiments were done on the available material in total. No significant difference in the SMRNNs' performance was recognised. Hence, to reduce the computational effort the feature extraction was done with a frame rate of 25ms. Assuming a mean utterance length of 2.74s, 1430 features per utterance

were extracted; roughly 7.5 times less as for HMMs.
In the experiments with HMMs the 12 MFCCs were extracted and the 0th cepstral coefficient is utilised as additional term. Further, to cover the temporal characteristics of the emotional speech Delta and Acceleration were computed for each feature and thus, 39 features in total were used as HMMs' input. For feature extraction the frame rate was set to 10ms and the window size to 25ms as common in the emotion processing from speech. Having these values and assuming a mean utterance length of 2.74s in total 10686 features per utterance were extracted. To compare HMMs with the neural networks the dynamic features were neglected as well and thus, two feature sets were analysed: i) MFCC and 0th cepstral coefficient only and ii) MFCC and 0th cepstral coefficient with additional Delta. Additionally, to provide the HMMs with full dynamics MFCC with Delta and Acceleration are extracted as a third feature set.

Classifiers

For each emotion a single SMRNN was trained and tuned. Due to the optimisation, each SMRNN has its own configuration as given in Table 4.3 on this page. In common is that all neural networks have two hidden layers with sigmoidal transfer functions (cf. Equation 4.19).

Table 4.3: SMRNN configuration grouped by emotional class in EmoDB (cf. [Glüge et al. 2011]). The differences in the SMRNN settings are due to an optimisation for each emotion.

Emotion	**Number of units**		**Segment length d**
	hidden 1	**hidden 2**	
anger	28	8	17
boredom	19	8	14
disgust	22	14	8
fear	17	17	7
joy	19	29	2
neutral	8	26	19
sadness	13	13	11

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{4.19}$$

Furthermore, all networks were trained for 100 epochs. For the details of the training procedure I refer to [Glüge et al. 2011].

The HMMs are constructed as left-to-right models with three internal states since the task is to model the emotion over a full utterance. The structure as well as the training procedure was according to [Böck et al. 2010]. The training and testing was done using HTK [Young et al. 2009] where a final decision was made on the basis of the log-likelihood; that is, favour the model with the highest log-likelihood which is nearly the winner-take-all principle usually applied in neural networks. As in case of SMRNNs for each emotion a single HMM was generated though with the same configuration.

Results

The performance of both classifiers was evaluated with WA (cf. Section 1.3.1) because the EmoDB is unbalanced in the number of utterances according to each emotion (cf. Section 3.1.1). Further, UA is computed as well to visualise the influence of unbalancing which is given in both classifiers but to a different degree.

As it is given in Table 4.4 on the current page the recognition rates of SMRNNs and HMMs are compared. The Table represents further the results for the HMMs having an enlarged feature set, that is the temporal features of MFCC and 0th cepstral coefficient, namely Delta and Acceleration, are added to the basic features.

Table 4.4: Recognition rates in percent of HMM and SMRNN classifiers during training and testing applying Unweighted Average accuracy (UA) and Weighted Average accuracy (WA) (cf. [Glüge et al. 2011]). For HMMs the two additional sets with temporal features, Delta (Δ) and Acceleration ($\Delta\Delta$), are given as well.

Emotion	**Training**		**Testing**	
	UA	**WA**	**UA**	**WA**
SMRNN	91.6	91.1	73.5	71.0
HMM$\Delta\Delta$	81.8	79.7	77.6	73.8
HMMΔ	81.1	81.2	63.3	60.0
HMM	70.7	71.2	55.1	51.7

The performance of the full set HMM with 39 features is comparable to the results achieved on EmoDB in other experiments, for instance, [Schuller et al.

2009a; Böck et al. 2010]. Further, the differences in the performance between training and testing testify that the models are able to generalise and therefore, cope with unseen material quite well. Reducing the feature set, on the other hand, results in a decrease of the classification performance. Between the full and the minimal set an absolute decrease of $\approx 22\%$ in WA is observed on the test set.

In contrast to the HMMs trained with 13 features only, SMRNNs can handle the task quite well. Their performance is in the range of full set HMMs with 71.0% WA. These results were achieved with 7.5 times less features as used for HMMs. The performance is astonishing considering the HMM with the minimal feature set which is related to the SMRNN's. They have almost the same behaviour regarding the relative difference between training and testing in the accuracies.

Investigating the ability to handle each emotion separately (cf. Table 4.5) it can be noticed that the variation in the results for HMMs is narrower than for SMRNNs. In some cases, for instance, fear and sadness, both classifiers perform equally or at least comparably. On the other hand, for the remaining emotions they provide results which are complementary to each other. This gives indication that both methods can be complementary combined (cf. Section 4.4) to improve the classification of single emotions.

Table 4.5: Correctness in percent grouped by the emotion comparing SMRNNs and HMM$\Delta\Delta$ on the test set. Particular values are taken from [Glüge et al. 2011]

Emotion	SMRNN	HMM$\Delta\Delta$
anger	100.0	84.6
boredom	62.5	87.5
disgust	75.0	50.0
fear	60.0	60.0
joy	57.0	66.7
neutral	62.5	87.5
sadness	80.0	80.0

Discussion

Both methods handle the emotion recognition task quite well. Although, for HMMs this was expected. In contrast, SMRNNs can cope with the issues as well. The main advantage of the neural networks is that they need less features to be trained. This results from the design of the networks itself which provides an ability to learn temporal relations, both short- and long-term dependencies. HMMs lack this ability and thus, dynamic features have to be added. This leads to a feature set per utterance which is 7.5 times larger than for SMRNNs.

On the other hand, the training procedure of HMMs is totally different from SMRNNs. Therefore, the computational complexity of HMMs is in the magnitude of $\mathcal{O}(S^2 T)$ where S is the number of states of an HMM and T is the observed sequence length (cf. [Binder et al. 1997]). For a given setup as in the experiment the complexity reduces to $\mathcal{O}(9T)$ since three state HMMs were used. In contrast, for SMRNNs the computational effort is much higher as discussed in [Glüge et al. 2012], resulting in $\mathcal{O}(4n^3 Nd)$ for the eBTT and $\mathcal{O}(4n^6(Nd)^2)$ for the eRTRL algorithm as in [Glüge 2013] where n is the number of neurons in each layer – to be assume as equal in all layers –, N is the number of segments, and d is the segment's length. Hence, the reduction of the feature set's advantage is compensated by the computational costs in training. Hence, for fast generation of recognisers HMMs are a better choice. Nevertheless, both methods have the potential to be combined to improve the recognition of emotion as well as disposition from speech classifiers, especially, due to their different ways of handling temporal relations (cf. Section 7.2.3).

4.4 Fusion Aspects

Although the main focus of this thesis is not on fusion of classifiers the topic is important to the issue of recognising dispositions and of interest in the context of multimodal classification of non-acted data sets like EmoRec (cf. Section 5.3). In my case, I mainly participated in work which was done by collegues at the Otto von Guericke University Magdeburg (cf. [Siegert et al. 2012d; Siegert et al. 2012c]) and the Ulm University (cf. [Walter et al. 2011; Schels et al. 2012]). Moreover, according to the semi-automatic annotation of data sets (cf. Section 4.1.3), I also proposed and discussed a framework which combines two modalities (cf. [Böck et al. 2012a; Böck et al. 2013a]). The proposed framework is indeed combining modalities but not in the sense of fusion. As introduced in Section 4.1.3 the classification of audio and speech signals provides information for an annotation and further, classification of facial expressions. Therefore, the combination is rather an advanced information flow. Nevertheless, the advantages of one modality are used for another one.

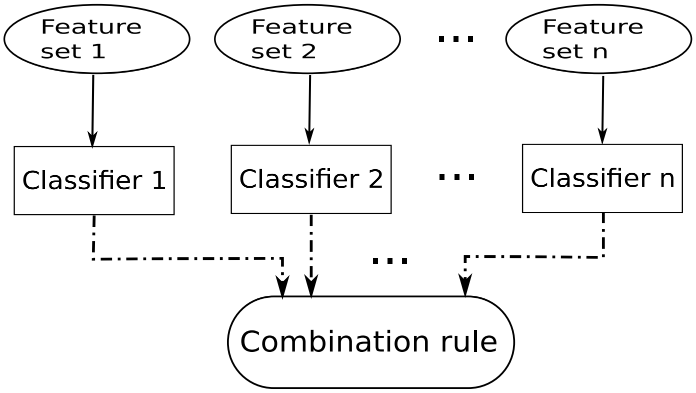


Figure 4.12: Overview of a *decision level* fusion architecture. For each classifier a single set of features is extracted. The final decision is generated by any kind of combination rule previously defined by the system's designer. Just the classifiers are trained from the feature sets (illustrated by solid lined arrows).

Taking into account this small collection of publications, the importance of fusion approaches in disposition recognition can be seen. Again, this thesis is looking at handling non-acted data, but also provides fundamental results which can be used in fusion, especially, in multimodal analyses. As the number of modalities is huge, there are several ways of combining information, too. In general, two points of action can be found where fusion should take place (cf. [Wagner et al. 2011]): i) *feature level* or ii) *decision* and *classifier level* (cf. also [Kuncheva 2004]). The first approach combines extracted features directly by applying techniques like concatenation of vectors or linear combination of feature values. Utilising this approach for Last Minute, results are given in [Panning et al. 2012]. The other method uses the decision of single classifiers which operate on features extracted for their own purpose only. It is called *decision level* fusion visualised in Figure 4.12. The results of the classifiers are fused, for instance, by using combination rules like Bayes' Rule or Dempster's Rule of Combination (cf. Figure 4.12). In *decision level* fusion the designer of the system generates and influences the final decision by constructing the combination function/method. In [Schels et al. 2012] this method is applied to EmoRec.

A third kind of fusion techniques, visualised in Figure 4.13, can be called *mid level* fusion. It is related to the *decision level* fusion approach but the output of the classifiers is fed to another classifier and thus, is the input for the new classifier (cf. in general [Kuncheva 2004]). In fact, this architecture is a cascade of classifiers. The advantage of that approach is that the final decision can be

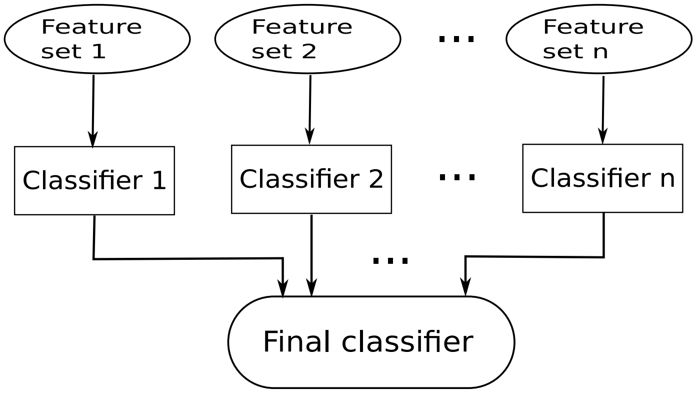


Figure 4.13: Overview of a *mid level* fusion architecture. For each classifier a single set of features is extracted. The final decision is generated by a classifier which uses the outputs of the previous classifiers as input features. In contrast to the *decision level* fusion (cf. Figure 4.12 on the preceding page) the connections between the two levels of classifiers are trained as well.

derived by training from an input which is already classified. Furthermore, this approach overcomes the problem of different sampling times in the input which usually occurs in other kind of fusion techniques, especially, in the feature level fusion. As the classifier outputs are finally fused, differences in time scales for each feature can be neglected. Hence, one could argue that a kind of temporal scale's alignment is established by mid level fusion. The mid level fusion approach was, for instance, used by [Glodek et al. 2011] in the Audi-Visual Emotion Challenge 2011.

These introduced fusion techniques are related to the classification purposes which are discussed in Section 5.3.

4.5 Summary

The development of classifiers for the purpose of emotion and disposition recognition from speech needs several steps and also has requirements. In the beginning, training and test material have to be preprocessed according to the three steps: i) *transcription*, ii) *annotation*, and iii) *labelling*. This was discussed in Section 4.1 where also a framework for semi-automatic annotation was introduced. After this preprocessing features can be extracted from the speech signal. Meaningful features for disposition recognition were presented (cf. Section 4.2).

In general, several kind of classifiers are possible to recognise dispositions from speech. In my thesis I concentrated on two, namely HMMs/GMMs and SMRNNs. The focus was on HMMs/GMMs which were introduced in detail and furthermore, parameters were investigated. From these considerations, I concluded a parameter setting as well as a classifier setup which is used in the experiments presented in Chapter 5. Thereby, audio only, bimodal, and multimodal setups are regarded where also fusion ideas are analysed.

Chapter 5

Non-Acted, Multimodal Experiments

Contents

5.1 Audio Setup Only 102

5.1.1 Results on Last Minute 102

5.1.2 Results on EmoRec I 110

5.1.3 Discussion 116

5.2 Bimodal Setups 117

5.2.1 Comparison of Audio & Facial Expressions 118

5.2.2 Results of Semi-automatic Annotation 122

5.2.3 Comparison of Audio & Biophysiological Features 124

5.2.4 Discussion 125

5.3 Multimodal Setup 127

5.3.1 Experimental Setup 127

5.3.2 Audio Classification Results 129

5.3.3 Discussion in the Context of Multimodality 130

5.4 Summary 131

In Chapter 4 methods and suitable parameter sets were introduced and investigated. Using these, training of classifiers and therefore, related experiments can be done. In the following, I discuss the results of three experiments which are all published – conducted by me as first author. The related references are given at the particular point.

In the first setup, only audio material is used to evaluate the classifier's results. These investigations were also used to refine the parameter sets which were introduced in Section 4.3.2. Based on those results, classifiers were applied in connection with two modalities, namely audio recordings and facial expressions. To cope with the new kind of ground truth and modalities (cf. Section 1.2.3 and 3.2) enhanced feature sets (cf. Section 4.2.2) were investigated, especially, in connection with semi-automatic annotation. Finally, a multimodal setup was investigated which used audio, video, and biophysiological features. For this, my part was the audio classification only that could further be incorporated in a fusion architecture combining all three modalities (cf. [Walter et al. 2011; Schels et al. 2012]).

5.1 Audio Setup Only

For experiments on non-acted material both data sets which were generated in the SFB/TRR 62 were used. In both sets the utterances are quite short (cf. Section 3.2.1 and 3.2.2) whereas the analysis on turn-level means that a dispositional label is given to the whole utterance. Further, no temporal evolution has to be modelled and thus, GMMs can be used for classification (cf. Section 4.3.2). In particular, for each dispositional category a single GMM is generated. In the beginning, HMMs with three internal hidden states were also tested to verify the parameter set on naturalistic corpora suggested in Section 4.3.2, since the internal structure of the classifiers was derived from this, commonly applied in speech recognition. This issue was already discussed in Section 4.3.2. Further, as both data sets are novel material those studies have to be done to explore the material.

5.1.1 Results on Last Minute

I introduced the Last Minute data set (cf. [Frommer et al. 2012b]) in Section 3.2.1. Four barriers (cf. [Rösner et al. 2012]) are included in the experimental setup, namely *baseline*, *challenge*, *listing*, and *waiuku* where the detailed description is given in Section 3.2.1. The sample's distribution according to each class is presented in Table 3.3 on page 53. Given this Table, it is advisable to apply WA as a reference for validation since the number of samples is unbalanced. In contrast to the expectations, almost no difference can be see comparing UA and WA values for Last Minute.

Comparing EmoRec and Last Minute, it is noticeable that both data sets contain quite short utterances with most of the time elliptical characteristic. This means, especially, in Last Minute short sentences are included whereas EmoRec comprises of commands necessary for the game ‘Concentration’. Having this in mind, I investigated the differences in handling dispositional characteristics. As mentioned in Section 4.3.2, HMMs reflect the temporal behaviour or evolution of dispositions where, in contrast, GMMs model the ‘pure’ dispositional characteristic. From this, I compared both modelling approaches to explore the given data set.

Experimental Setup

The experimental setup for the classification of dispositions on Last Minute is twofold concerning the classifier. Nevertheless, the features which I applied for training and testing are the same in both cases. Again, since the corpus and its characteristic are novel it was not clear which features are suitable to cover the dispositions. For this, I concentrated on features that are feasible from the eNTERFACE’s point of view. As eNTERFACE (cf. [Martin et al. 2006]) can be seen as related to a near real-life, naturalistic interaction although it is acted material (cf. Section 3.1), the features can be transferred to non-acted material as well. Can this be done without concerns? Especially, eNTERFACE is on the one hand an acted corpus, but it is a data set where non-actors were recorded. That means, the participants had to be set into a certain dispositional situation which was done with dispositionally coloured texts. From this, I assumed that the disposition was not acted and thus, a near real-life situation is generated. This gives an indication that the same features can be applied for naturalistic material as well where the disposition was induced by a WoZ scenario as in Last Minute (cf. Section 3.2.1). Furthermore, this assumption holds for EmoRec (cf. Section 3.2.2), too.
Taking into account the results from Section 4.3.2, especially Figure 4.10(b) on page 89, I considered MFCC and PLP only, since the LPC shows a significantly low performance on eNTERFACE. Nevertheless, MFCC had only a slightly better performance, particularly with the 0^{th} cepstral coefficient.

At first, I give the setup of the HMMs. Their internal structure is as described in Section 4.3.1 and visualised in Figure 4.7 on page 81. I used models with one Gaussian Mixture in each state and three hidden states. The HMMs are organised as left-to-right models containing self-loops in each state. The transition and production probabilities, a_{ij} and o_i, respectively, are initialised equally at the beginning of the training. The mean μ_i and variance σ_i values of each mixture i

where estimated as a flatstart model (cf. HTK [Young et al. 2009]). That means, the mean and variance are computed as an average of the corresponding values analysing the training material in total. Since I applied only models with one mixture the estimations of μ_i and σ_i are generalised to each state of the HMM. Like in speech recognition, for instance, with phonemes I utilised one HMM per dispositional state. This leads to four HMMs with three internal states modelling the classes baseline, challenge, listing, and waiuku. As proposed in [Böck et al. 2010] each model is trained for five iterations.

Concerning the GMMs, I applied HTK for training and testing as well. This is possible because a GMM can be interpreted as an HMM with one state and multiple Gaussian Mixtures. From this the internal structure of my models is defined as an one state model with an internal self-loop coping with the temporal sequence of a sample. The internal values of μ_i and σ_i are set as flatstart as well.

To show the potential of the GMM approach I trained also HMMs with Gaussian mixtures in the hidden states. Therefore, I varied the number of mixtures in the range of $[1, 10]$, stopping the search when no further gain in performance was determined. In fact, the performance decreased with higher numbers of mixtures (cf. Table 5.1 on the next page).

Results of Hidden Markov Models

From Figure 4.10(b) on page 89, it is to be expected that in the case of non-acted material the 0$^{\text{th}}$ cepstral coefficient would work better than the Energy term as additional parameter. To prove this issue both terms were compared on Last Minute in both cases with HMMs and GMMs and keeping MFCC fixed. Using HMMs the Energy terms show a better performance in terms of accuracy, namely 21% WA with Energy in contrast to 17% applying 0$^{\text{th}}$ cepstral coefficient. Both experiments were conducted on specifications given in the experimental setup applying a cross-validation.
From a general point of view, both additional terms behave quite different in my experiments. With the 0$^{\text{th}}$ cepstral coefficient the HMMs show a worse performance than with the Energy term. Inspecting the trained HMMs, it can be stated that in the particular cases, the partitioning of the 0$^{\text{th}}$ cepstral coefficient's characteristics is quite difficult. For these models, it was not possible to estimate an optimal distribution which might be due to the complexity of the course of this coefficient. In contrast, for the Energy term a proper partitioning of the

characteristics to the different states in the HMM could be established as for this term a temporal evolution can be noticed. Therefore, those models had a better recognition performance, namely 21% WA.

Based on these considerations the observation of the mixture's influence was done utilising the Energy term, only. In Table 5.1 the values of UA and WA for several number of mixtures are given. Increasing the mixtures' number improves the performance of the models significantly.

Table 5.1: Comparison of HMMs with different numbers of Gaussian mixtures measuring Unweighted Average accuracy (UA) and Weighted Average accuracy (WA) in percent.

Number of mixtures	**UA**	**WA**
1	21.0	21.0
2	27.5	27.5
4	33.0	32.0
6	33.5	32.0
8	33.5	32.0
10	28.5	28.5

Utilising more Gaussian mixtures yielded a boost in the performance of the HMM up to a certain threshold. Exceeding this, the advantage is exhausted by the number of internal states. This is related to the phenomenon of overfitting which occurs also in ANNs and means that the model is too specialised or adapted to the given training material. It leads to the loss of generalisation's ability and hence, the performance decreases although the classifiers show a good performance while testing on the training set. Such an effect can be seen with the HMMs in the experiment as well. The internal structure and the number of mixtures are important parameters. With HMMs it is possible that due to the learning of the temporal characteristics of the disposition, the performance gain of the Gaussian mixtures is compensated by the number of hidden states which results in a worse performance. The internal probabilities and the mixtures' parameters reproduce the characteristics of the training samples. Unfortunately, exceeding a certain number of mixtures to more than eight (cf. Table 5.1) this results in the performance loss.

In contrast, using only GMMs this temporal effect does not occur which is related to the design of the model. However, even with GMMs a loss in performance can be see because of the overfitting effect already discussed in Section 4.3.2; that is, using to many mixtures results also in a worse performance since the classifier is highly adapted to the training material. Hence, the number of mixtures is a parameter which has to be investigated. These experiments are discussed in the following Section.

Before I present the results applying GMMs a few more general words considering the results of HMMs have to be said. Regarding the numbers in Table 5.1 on the previous page it can be seen that they are low compared to common results in emotion recognition. From my point of view, there are two main reasons why the results look so bad. First of all, disposition recognition is examined. This is a kind of analysis which is novel in the community. So far, just a few experiences are at hand realising a classification of such an aspect. On the one hand, the topic is somehow 'weak' and hard to capture, especially, in a sense that a commonly excepted definition of disposition is still lacking. On the other hand, the material, which is analysed, is quite naturalistic and up to now, the community just starts to get involved in such studies.
These issues will be partly discussed in Section 5.1.3 in a broader sense.
On the other hand, the values in Table 5.1 on the preceding page show a development to increase the performance of the classifiers as it was discussed already. Improving the parameters yielded a gain in the performance. Of course, 32% WA and 33.5% UA (cf. Table 5.1 on the previous page) appear to be bad results. However, I encourage the reader to an experiment as it is described in Section 3.2 taking into account the performance of a human begin classifying dispositions. Again, with naturalistic, non-acted data sets and more general classes like *challenge* or *waiuku* recognition results are lower than having optimal, acted material (cf. Figure 4.10(a) on page 89). Even in case of eNTERFACE which can be seen as a more realistic corpus recognition accuracies are in the range of 40% (cf. Figure 4.10(b) on page 89) while having a seven classes task. Considering these aspects, an absolute improvement of 7% WA compared to chance level was yielded by the HMMs. With a slight switch in the type of classifier towards GMMs an additional improvement can be achieved.

Results of Gaussian Mixture Models

As I already explained, the utterances contained in Last Minute are quite short compared to other corpora, for instance, the SAL [McKeown et al. 2012] where HMIs are recorded containing also kinds of monologues. From this and based on findings given in Section 4.3.2, I argue that the dispositional information is mainly enclosed in the style of speaking than in the temporal evolution. This means, due to the shortness of the expression no evolution is seen in the disposition since the analyses are on utterance-level. Indeed, considering longer time periods the temporal effect will be observable (cf. Section 1.2.2). But, for such short time spans the temporal characteristic is masked by style effects. Hence, GMMs as classifiers become of interest and will be considered in the following.

Guided by the results of the HMMs' experiments I analysed the performance of MFCC first. As already discussed the eNTERFACE is a good indicator to derive suitable experiment's parameters. Hence, I trained the GMMs also for five iterations (cf. Section 4.3.2). Based on the findings of Schuller et al. the number of Gaussian mixtures was set to 81. In [Schuller et al. 2009a], the authors report on their experiment using 80 additional mixtures and thus, due to the specifications of HTK [Young et al. 2009] the total number of mixtures is 81 as at least one mixture has to be used in each state. Testing other numbers of mixtures in the neighbourhood of 81 showed no significant changes in the average performance, however, in the single runs the accuracies varied. To avoid side-effects of single speakers all experiments were done in a cross-validation manner.

In Table 5.2 the first row shows the results of mean UA and WA for MFCC.

Table 5.2: Results of the cross-validation presenting the Unweighted Average accuracy (UA) and the Weighted Average accuracy (WA) in percent for GMMs comparing MFCC and PLP features. For both experiments the number of mixtures in the GMMs were fixed to 81.

Feature set	**UA**	**WA**
MFCC_0_D_A_Z	43.97	43.97
PLP_0_D_A_Z	43.96	43.96

To handle the variances in the cepstral coefficients' mean values a cepstral mean normalisation was applied to the MFCC which is indicated by _Z. This normalisation is done by estimating the mean "by computing the average of each cepstral parameter across each input speech file" [Young et al. 2009]. From this, long term effects, for instance, from different microphones or channels are compensated. It is not the normalisation of differences in the speakers' characteristics. Hence, it is advisable to do a cepstral mean normalisation as it is not guaranteed – and this is valid for almost all corpora – that the recording conditions kept absolutely stable for the total data collection process.

With GMMs a boost in the performance of more than 10% absolute accuracy

(cf. Table 5.1 on page 105) compared to HMM was achieved. It is conspicuous that there is no difference between UA and WA results. From Table 3.3 on page 53 it is known that the number of samples is not equally distributed as one could assume by regarding those values. In any case, the differences are not as significant, especially, for *baseline* and *waiuku*. Considering each run, as the mean values are estimated over 10 runs with an arbitrary selection of the material to training and test set, separately even so, the two validation measures are equal. Therefore, the values for UA and WA are equal since for each class the same recognition accuracies were achieved by the classifiers. From this, it can be stated that no class was preferably learned by the classifiers even in this case of a not equally distributed data set (cf. Table 3.3 on page 53). Hence, the trained classifiers show a performance that can be assumed to general – not optimal – since the recognition is spread over all classes which is a characteristic of a general recogniser, from my point of view.
Further, as motivated by Figure 4.10(b) on page 89 PLP features were also tested. There are no significant changes in the performance in comparison to MFCC (cf. Table 5.2 on the preceding page). Hence, the same conclusions can be drawn as in the case of MFCC. In contrast, for cross-validation PLP features have a significant better performance in the sense of interindividual validation recognition of dispositions.

The interindividual validation approach, as introduced in Section 1.3.2, reflects the generalisation ability of classifiers to handle material which was not seen in training. This provides the opportunity to generalise dispositional characteristics since the training material is independent from those used in testing.
Again, a cepstral mean normalisation was applied to both MFCC and PLP whereas the feature sets are motivated by Figure 4.10(b) on page 89. The normalisation is already introduced in a previous paragraph.

Table 5.3: Interindividual recognition results of dispositions on the Last Minute corpus measuring the mean Unweighted Average accuracy (UA) and mean Weighted Average accuracy (WA) in percent.

Feature set	**UA**	**WA**
MFCC_0_D_A_Z	40.65	43.33
PLP_0_D_A_Z	43.59	44.96

In Table 5.3 the recognition results of dispositions are presented. Regarding the performance of PLP coefficients an improvement of the recognition accuracy of

≈ 1% absolute WA was achieved compared to MFCC. Usually, interindividual validation results in a worse performance in pure speech recognition. In contrast, for disposition recognition from speech the results show a gain in the performance. This was also seen in the experiment on EmoRec I (cf. Section 5.1.2). The underlying concepts are matter of further interdisciplinary research.

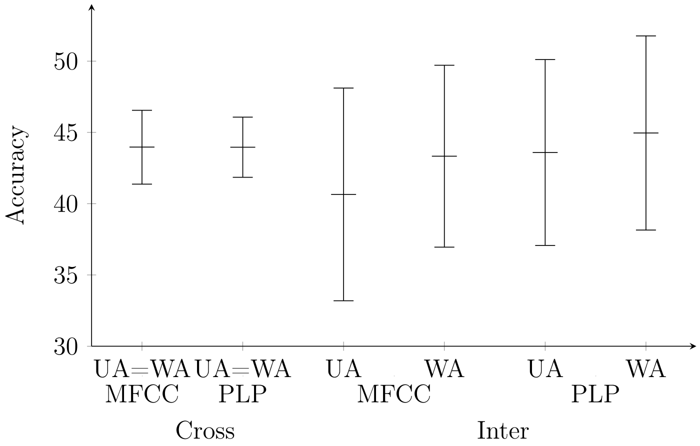


Figure 5.1: Mean values of Unweighted Average accuracies (UA) and Weighted Average accuracies (WA) in percent with corresponding standard deviations on the Last Minute corpus for MFCC and PLP features grouped by cross-validation (Cross) and interindividual validation.

Regarding the values in Table 5.2 on page 107 and Table 5.3 on the preceding page, it seems that the interindividual validation has a better performance compared to cross-validation. Therefore, I computed the standard deviation for each experiment. For the cross-validation it is for both measures equal and thus, resulted in 5.15% for MFCC and 4.22% for PLP. In contrast, for interindividual validation I obtained for MFCC 14.92% in UA and 12.76% in WA, respectively. On the other hand, for PLP the standard deviations are computed to 13.04% in UA and 13.62% in WA. From this, it can be concluded that the cross-validation still works better than the interindividual validation as the standard deviations are smaller. The mentioned values are also visualised in Figure 5.1 on this page. Comparing the feature sets an improvement of accuracy was achieved. Note that the performance is comparable to the findings in eNTERFACE (cf. Figure 4.10(b) on page 89). However, in eNTERFACE MFCC features performed better than PLP features. The analysis of the way how both features are extracted from the speech signal (cf. Section 4.2.1) show two differences that might result in the gain of performance. From my point of view, the scaling which is applying the power law of hearing in PLPs or the mel-scale in MFCC, is relevant and influences the final results significantly. For this corpus, PLP coefficients are more adaptable and seem to be more sensitive to the way dispositions are uttered by the participants of the WoZ scenario.

On the other hand, from Table 5.3 on page 108 it can be observed that the values differ only marginally for both feature sets. Therefore, no general remark for a certain feature set used on non-acted corpora can be derived although, an indicator for both methods is given. In the following the other corpus, namely EmoRec I is analysed with respect to audio features only.

5.1.2 Results on EmoRec I

For the analyses in this thesis I concentrated on the so-called EmoRec I (cf. Figure 3.1(b) on page 54) data set which is a subset of the corpus EmoRec I+II introduced in Section 3.2.2. So far, only the first part of the corpus, namely EmoRec I, is (almost) fully prepared by the colleagues at the Ulm University to be investigated in experiments. Further, it is the set which is also used in multimodal investigations and hence, observed by other groups which are involved in the SFB/TRR 62 (cf. [Walter et al. 2011; Schels et al. 2012; Tan et al. 2012]). In detail, two sequences were distinguished in the whole experiment called ES-2 and ES-5 which are related to positive and negative disposition characteristics of the speaker, respectively, as discussed in Section 3.2.2. The results of audio analyses are mainly published in [Böck et al. 2012b].

Furthermore, I participated on analyses related to two or more modalities and contributed the audio results. The dispositions are induced according to the octants in the PAD space (cf. [Mehrabian 1996] and Figure 5.2 on the facing page). Each octant represents a certain user state and thus, is related to a disposition. For the following experiments I concentrated on ES-2 and ES-5 (cf. Figure 3.1(b) on page 54) with the following coding in PAD space: i) ES-2 with *positive pleasure, low arousal, high dominance* reflecting a positive disposition and ii) ES-5 located at *negative pleasure, high arousal, low dominance* reflecting a negative one.

In addition, in multimodal experiments (cf. Section 5.3) two other sequences were observed which have the following characteristics: ES-4 and ES-6 are based on EmoRec II (cf. Figure 3.1 on page 54). They are connected to *negative valence, high arousal, low dominance* and *positive valence, low arousal, high dominance*, respectively, in the PAD space. For this, they represent similar characteristics of the speaker and thus, are investigated under the same considerations.

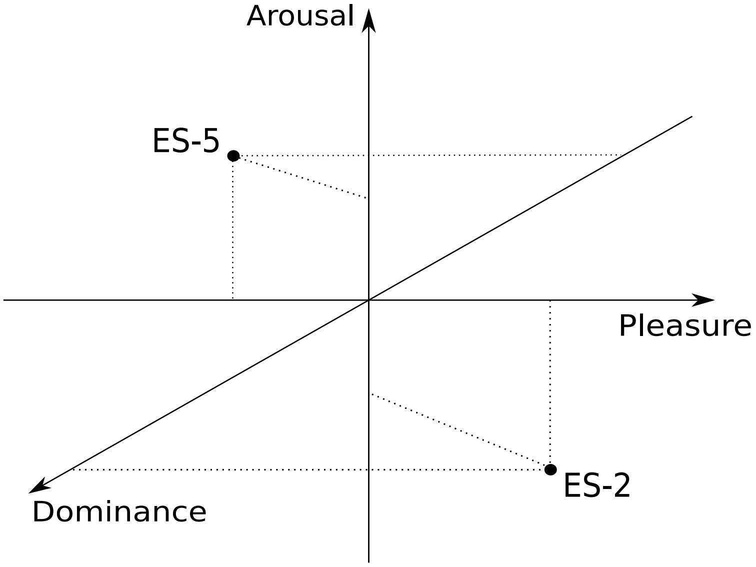


Figure 5.2: Schematic visualisation of the two experimental sequences ES-2 and ES-5 in the PAD space. ES-4 is related to the position of ES-5 whereas ES-6 is in the surrounding of ES-2. Both ESs are neglected for the sake of clarity in presentation.

The results on the experimental sequences are also presented in [Walter et al. 2011; Schels et al. 2012; Böck et al. 2012a; Böck et al. 2013a]. Though, the findings will be discussed in the corresponding Sections (cf. Section 5.2 and 5.3).

Experimental Setup

Since both corpora, namely Last Minute and EmoRec are generated in a WoZ setup, I transferred the experimental setup applied on Last Minute also for the classifiers on EmoRec. This means, GMMs were applied, modelling each dispositional category with a single GMM. In the case of EmoRec this leads to two models tagged with *positive* and *negative*. As GMMs are applied in the classification the speakers' characteristics have to be modelled by using a set of Gaussian mixtures within each classifier. From the results of Last Minute I developed the classifiers with 81 mixtures since both data sets are similar to each other. To achieve comparable results for the inter- as well as intraindividual experiments MFCC are extracted from speech samples using a common setting from speech recognition, that means, a frame rate of 10ms and a Hamming window (cf. Equation 4.1) with a window size of 25ms having an overlap of 15ms. MFCC with 0th cepstral coefficient are selected as in Last Minute they showed significant stable results both in cross-validation and in interindividual validation which both indicate a good performance for EmoRec. In addition to the Delta and Acceleration values, the cepstral mean normalisation indicated by _Z was applied.

The already mentioned two classes, namely *positive* and *negative*, were derived from the experimental design given by the involved psychologists. As already introduced in Section 3.2.2 the dispositions are induced by controlled influence of the system's reactions of the game 'Concentration' played by the participants. In terms of the psychologists they call them still emotions, but from considerations in Section 1.2.2 the states of the participants different from pure emotions. Namely dependent on the different situations which are generated by the wizard and the game itself, more complex interconnections have to be regarded. These are heavily related to situatedness and hence, from my point of view, the term disposition is justified.
All audio samples from the corresponding sequences were cut from the audio stream and manually assigned. The aforementioned features were extracted utilising HTK (cf. [Young et al. 2009]).

Interindividual Results

The two class issue mentioned in the experimental setup of EmoRec was analysed using an interindividual validation approach. The averaged results for each speaker while combining ES-2 and ES-5 are presented in Table 5.4 on page 114. Comparing the UA and WA values, the difference between both is small. Especially, the mean values are not as different since UA is 52.3% and WA 55.1%. Again, the EmoRec corpus is a data set which provides naturalistic speech samples with non-acted dispositions. As already discussed in Section 3.2, this leads to accuracy values that are respectable but not as high as those achieved on acted corpora. Even for human listeners and annotators the classification task is quite difficult. From this, the automatic classification based on GMMs shows respectable results. Moreover, the results can be ranked in more detail if they are compared to other modalities as it will be done in Section 5.2 and Section 5.3. Further, in Table 5.9 on page 129 the classification accuracies on EmoRec II which is the second cycle in the scenario (cf. Figure 3.1(a) on page 54), are presented.

Regarding these values, two aspects can be stated: i) the inducing of the dispositions worked on both cycles similarly well and ii) the results of EmoRec I can be reproduced. This mean, the methods show similar recognition accuracies on both subsets of EmoRec, for instance 52.3% UA on EmoRec I and 52.2% UA on EmoRec II (cf. Table 5.4 on the following page and Table 5.9 on page 129). However, the improvement of the classification performance on such naturalistic material is the matter of current research and thus, aspects of how to deal with this issue are discussed in several Subsections of Section 7.2.

Investigating the single results for each participant in most cases they are narrow and concentrated between 51% and 57% for WA. For an interindividual classification this is quite noticeable even as the variation is small. Furthermore, regarding the single results in Table 5.4 on the following page, it can be seen that for some participants UA has a greater value as WA. This aspect derives from the characteristics of the confidence measures and is discussed in Section 1.3.1. Nevertheless, two conclusions can be emphasised from the experiment as follows. At first, the features which are derived from the LAST MINUTE corpus are transferable to EmoRec. One step further, they seem therefore also generally applicable for i) disposition recognition as such and ii) for the handling of non-acted material. Even more, MFCC_0_D_A_Z features cope with both aspects in combination, that means, they can be used for a disposition recognition from non-acted speech. This is reflected by the performance of the classifiers, namely GMMs, given the feature set. Thus, it is shown – as in emotion recognition from speech – that MFCC enriched with the zeroth cepstral coefficient are suitable for disposition recognition from speech.

On the other hand, and this is more a kind of a meta-interpretation, the hypothesis arises that the dispositional characteristics for ES-2 and ES-5 are speaker independent. Of course, this issue is an aspect which is more to psychologists to discuss but it is also of interest from a technical and classification point's of view. As the disposition was induced (cf. Section 3.2.2) the experimental conditions can be assumed as being fixed. Therefore, the personal characteristics of each participant are in the focus. From the results in Table 5.4 on the next page the small variations in the classification give hint that the hypothesis is true and especially, the distinction between positive and negative dispositions is independent from a speaker. For emotion recognition from speech the question was also observed in, for instance, in [Kostoulas et al. 2008; Kotti et al. 2010]. The investigation for dispositions is so far just done on as technical issue. It lacks the verification under the focus of psychological analysis related and combined with technical aspects (cf. Section 7.2.3). According to emotions such a combination was investigated, for instance, by [Lugger & Yang 2008].

Table 5.4: Classification results of dispositions for EmoRec I based on ES-2 and ES-5 measuring Unweighted Average accuracy (UA) and Weighted Average accuracy (WA) in percent for each participant of the experiment. Further, the mean accuracies are given. In the Table the accuracy values are grouped according to interindividual and intraindividual validation.

Participant	**Interindividual**		**Intraindividual**	
	UA	**WA**	**UA**	**WA**
112	65.0	58.7	55.6	57.5
114	47.6	51.6	62.5	63.5
118	32.1	46.2	65.3	56.5
125	49.4	54.1	62.5	60.5
127	43.2	52.4	71.4	70.0
129	51.5	55.3	68.0	67.0
208	41.8	49.8	79.7	75.0
212	81.1	79.6	86.7	90.0
215	39.7	52.4	100.0	100.0
219	57.1	56.6	68.5	68.5
225	56.9	52.6	96.0	95.5
226	59.7	56.6	84.3	86.5
308	59.7	57.8	65.8	67.5
423	60.3	56.0	49.2	58.5
427	60.7	61.0	68.8	69.0
506	56.8	51.3	70.5	72.5
510	42.5	53.2	52.5	52.5
511	38.6	48.0	62.9	63.5
518	69.2	67.3	68.6	66.0
602	39.2	42.3	60.0	60.0
mean	52.3	55.1	69.9	70.0

Intraindividual Results

In Table 5.4 on this page the classification results for the intraindividual validation experiments are given, grouped by UA and WA. As introduced in Section 1.3.3, for this kind of validation I utilised only material from one speaker participating in the WoZ scenario EmoRec I. To get significant results each experiment was repeated ten times splitting the material arbitrarily and averaging the accuracies afterwards. From this, it is possible to show the classification results compared to the accuracies achieved in interindividual experiments.

In general, the behaviour of the results is similar to those achieved in interindividual validation, whereas the differences between UA and WA are not as large. As supported by the values in Table 5.4 on the preceding page an improvement of the recognition power is attained by observing each participant separately. This aspect is also known in speech recognition and is, for instance, used in several dictation systems (cf. [Nguyen 2009]). As it can be seen, such approach is suitable for disposition recognition from speech as well (cf. Section 5.1.3). Doing so, a gain of 14.9% absolute was achieved for mean WA. Considering the single speakers the improvement is different. Except three participants, the switch in the validation method results in an increase of performance. As the same features were applied for all experiments and also the same setup was used, it was first unclear why the performance for these three speakers decreased (cf. in Table 5.4 on the facing page these are participants 112, 510, and 518). Watching the video and listening to the audio material of those users, I realised that they are ‘switching’ between the two dispositional states quite often even in the same ES. From this, it was challenging for the classifier to generalise. In contrast, having different user characteristics in the training helped to extract a common characteristic in the behaviour. This is the case for the three participants.
Another aspect of detailed observations is that the described effect (cf. interindividual analyses) of larger WA values than UA ones occurs only with nine users whereas 11 are effected in interindividual validation. This is due to the more balanced distribution of the samples, that means, the total number of audio samples which are assigned to the two classes is more similar. Nevertheless, even in intraindividual validation the counterbalancing effect is given. These characteristics of the confidence measures are already discussed in Section 1.3.1. Further, the classifiers were able to learn both classes which indicates that also the utilised features, namely MFCC_0_D_A_Z, reflected the characteristics of both classes in a better way. Moreover, the difference of the means is almost negligible (cf. Table 5.4 on the preceding page with 69.9% UA and 70.0% WA).

In Section 5.2.3 as well as [Böck et al. 2012b] the intraindividual results are compared to these achieved with classifiers trained and tested on biophysiological features. As mentioned in Section 3.2.2, EmoRec provides the possibility to compare user reactions on audio as well as on biophysiological characteristics. Of course, internal features like heart rate, blood volume pulse, etc. are more user specific and therefore, are able to deal with his characteristics, nevertheless, in comparison also with audio features significant results were achieved. For detailed analyses I refer to Section 5.2.3.

5.1.3 Discussion

Disposition recognition from speech is so far a novel kind of analysis. From this, the community has just few experiences how to handle such a complex task, especially, in the sense of classification. In Section 5.1.1 I started the analysis of a non-acted data set which can be regarded as material enriched with dispositional features. In contrast to SAL [McKeown et al. 2012], for instance, which is still dealing with emotions, Last Minute and EmoRec are designed to provide material related to dispositions; in case of Last Minute even dispositions are assigned. As I already said, I started the analyses and from my point of view both corpora are rich in aspects worth investigating, in the sense of disposition recognition. Therefore, this thesis could only provide an incubator for several aspects of research.

Especially, when regarding the values given in Table 5.1 on page 105 as well as Table 5.2 on page 107 and Table 5.3 on page 108 it can be seen that they are low compared to common results in emotion recognition. In this case, common emotion research means applying acted material. As I already discussed in Section 3.2, I encourage the reader trying to classify dispositions by himself. From this, it will be more obvious that this is a challenging task. This is even more true, as the community at all is still lacking a common definition of disposition combining psychological and technical aspects. The interdisciplinary character of this research field entails potency for discussions, whereas several ideas are already published covering details of each involved discipline. In this thesis I provide a definition which is influenced by the psychological view, but is related to a technical sense, see Definition 1.6 on page 7 in Section 1.2.2.

Aiming on the exploration of the two non-acted data sets, I compared the features, which are well-known from speech recognition and already proved to be suitable for emotion recognition from speech, under the issue of availability for disposition recognition from speech samples. Shown in the corresponding Sections (cf. Section 5.1.1 and 5.1.2) it has been figured out that especially, MFCCs are suitable and provide a significantly good opportunity for generalisation. This

is also valid for different kinds of validation methods. In particular, a gain in performance was achieved by an intraindividual validation.
This aspect is also discussed in [Böck et al. 2012b] were I suggest to provide a technical system with the opportunity to switch between interindividual and intraindividual validation if necessary. Such procedure is just important if the technical system is intended to be a companion (cf. [Wendemuth & Biundo 2012]). Therefore, the system has the possibility to adapt towards a certain user. In the beginning, a common model for disposition recognition is necessary which can be derived by an interindividual approach. While interacting with a certain user the system is collecting data and uses the classification result to refine the internal models, namely GMMs. After a while the switch to an intraindividual validation can happen. So far, it is still an open question when to switch between the strategies. A sketch proposal is to use a decrease of WA as an indicator as inspired by the work with ANNs.
Furthermore, the investigated validation approaches helped to rank the classification results incorporating other modalities than the audio channel. This will be discussed in more details in Section 5.2. In Section 5.3 I present further considerations of multimodal investigations that lead directly to fusion aspects. As fusion in general is not in the focus of this thesis, I refer also to open issues that are given in Section 7.2.3.

5.2 Bimodal Setups

The following settings provide another view of results gained in disposition recognition from speech. They are linked with other modalities, especially, facial expression derived and annotated with FACS as well as biophysiological parameters of the speaker. Since just EmoRec provides such an enriched and detailed feature set all experiments of this Section are done on this corpus only. Further, the material is analysed under several aspects and the corresponding results are published. However, I concentrate on those findings I achieved by myself or participated on; that means, for the purpose of publication I was the first author or a co-author. The references are mentioned at the corresponding point.

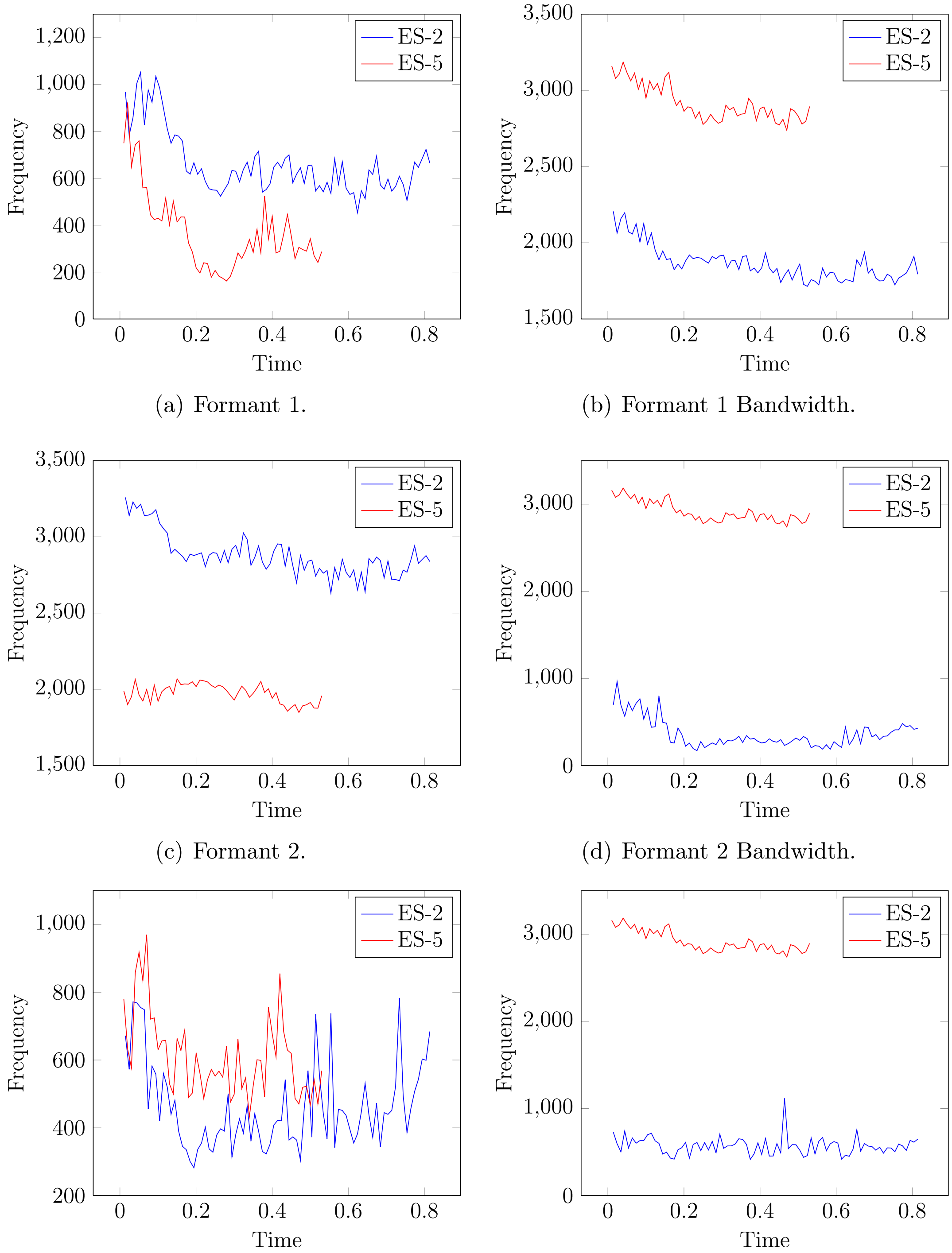
5.2.1 Comparison of Audio & Facial Expressions

As presented in [Böck et al. 2012a] a connection between facial expressions and spoken utterances can be built. Especially, disposition recognition from speech provides the possibility to analyse parts of the experimental sequences in EmoRec where facial expressions are not evaluable since the participant is speaking. Moreover, the current Section and Section 5.2.2 are linked as both are working on the same data set, namely EmoRec. The difference of the Sections is that the current one is presenting the basic results used afterwards in the semi-automatic annotation.

Using the aforementioned methods and features for disposition recognition from speech, I achieved the results already discussed in Section 5.1.2. They are entirely based on the audio material without any relation to other modalities. This means, the audio samples were processed and classified as they appear in the data set and no link to information from other modalities is established.

In addition to MFCC features I also investigated prosodic features as introduced in Section 4.2.2, namely formants and their bandwidth, intensity, pitch, and jitter. According to [Scherer 2005], for instance, those are meaningful in emotion recognition from speech. Evaluating this in the context of disposition recognition and especially, aiming for a semi-automatic annotation (cf. Section 4.1.3 and 5.2.2) I analysed the features on EmoRec I. To derive an additional feature set the analyses (published in [Böck et al. 2012a]) were done manually, extracting the features by applying Praat (cf. [Boersma 2001]). Finally, with these additional feature set classifiers were trained. The results are presented in Section 5.2.2.

In addition to the results presented in [Böck et al. 2012b], the full set of participants was analysed. In Figure 5.3 on the facing page the three formants with their corresponding bandwidths are given. The plot shows the mean values averaged over all 20 participants who are selected as a common set having all modalities in both data sets – EmoRec and Last Minute – as discussed in Section 3.2.1 and Section 3.2.2. For this, the features are extracted from the speech of each participant using Praat (cf. [Boersma 2001]) and are averaged over all speakers afterwards. Hence, all audio samples of all speakers are used to calculate the mean values. As it is known from literature, for instance [Scherer 2001; Vlasenko et al. 2011b; Vlasenko 2011], formants are discriminative for emotions. Figure 5.3(a) on the next page and Figure 5.3(c) on the facing page support the hypothesis that this is also valid for disposition recognition from speech. Especially, formant 2 is highly discriminative for the two dispositions – namely ES-2



(a) Formant 1.

(b) Formant 1 Bandwidth.

(c) Formant 2.

(d) Formant 2 Bandwidth.

(e) Formant 3.

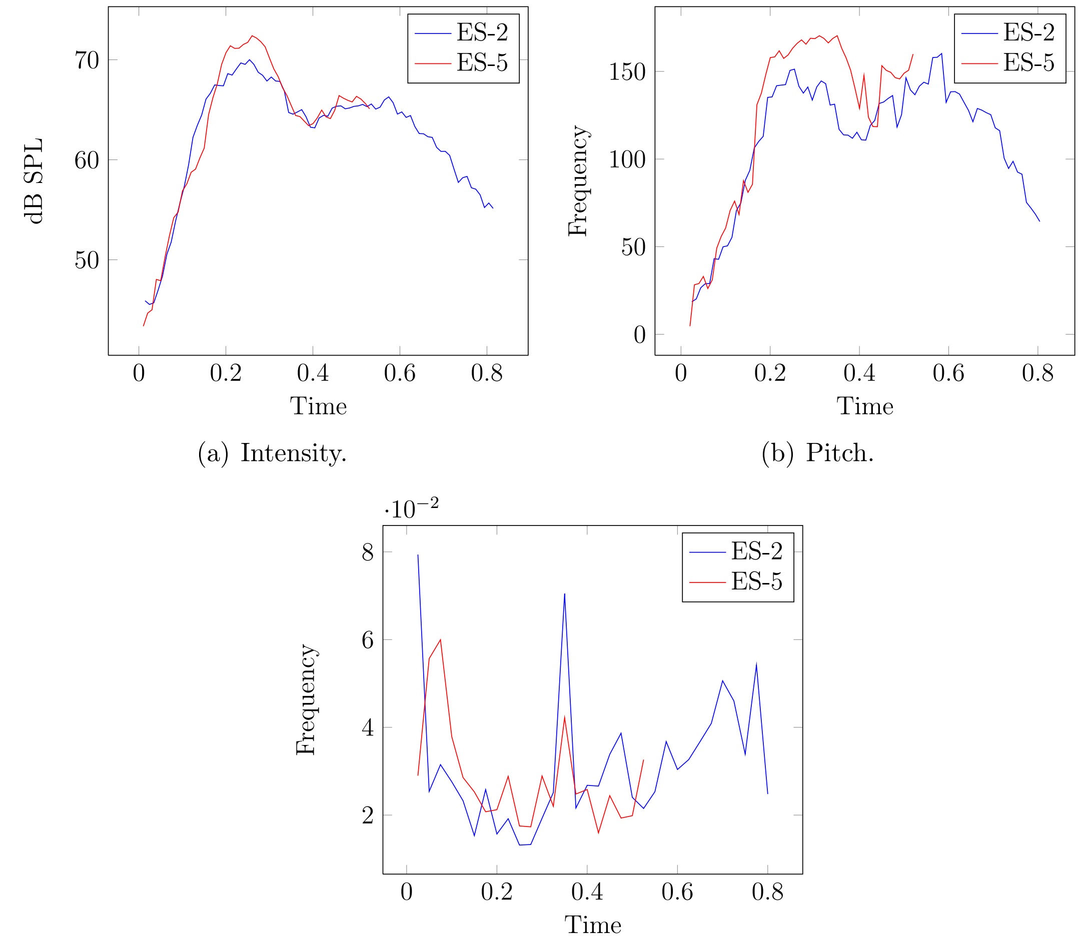
(f) Formant 3 Bandwidth.

Figure 5.3: Global mean values in Hz for the formants 1 to 3 (cf. Figure 5.3(a), Figure 5.3(c), and Figure 5.3(e)) and their corresponding bandwidths (cf. Figure 5.3(b), Figure 5.3(d), and Figure 5.3(f)).

and ES-5 – as the average frequencies of both are a long way apart from each other; approximately 1000Hz (cf. Figure 5.3(c) on the previous page). It is of interest that in such naturalistic HMI the shift of frequencies is contrary to the expectations derived in [Scherer 2001], for example. From my point of view, this characteristic is due to the non-acted material. Scherer achieves the findings by regarding HHIs which usually supply slightly different characteristics. Investigating formant 3 (cf. Figure 5.3(e) on the preceding page) the course is as expected. However, especially, in several parts of the diagram overlaps are given. Therefore, this formant is not as discriminative as the others.
The formants' bandwidths (cf. Figure 5.3(b) on the previous page, Figure 5.3(d) on the preceding page, and Figure 5.3(f) on the previous page) show the characteristic as predicted by Scherer. Moreover, all three are significant features to distinguish ES-2 and ES-5; the two dispositions which are investigated.
The mean values of the features presented in Figure 5.3 on the preceding page and Figure 5.4 on the next page are extracted from each participant's audio samples and afterwards averaged over all speaker; this means, that for all speakers also all samples were used to get these mean values. From this, the results can be seen as a kind of interindividual observations. For classification purpose they work quite good for interindividual experiments but also for intraindividual ones.

In contrast to the formants and their bandwidths, intensity, pitch, and jitter are not as discriminative. In general, these features show the same characteristic in the course either in ES-2 or ES-5. The differences are marginal (cf. Figure 5.4 on the facing page). Nevertheless, for the issue of semi-automatic annotation (cf. Section 4.1) they give a slight improvement, unfortunately, it is not significantly high. Considering jitter, for example, this is related to fear, anger, and anxiety (cf. Section 4.2.2), but in case of ES-5 it is not significantly discriminative compared to the course of ES-2 because none of the aforementioned dispositions is observed. Again, the naturalistic, naïve HMI shows different characteristics as an acted one. This circumstance is already discussed in [Batliner et al. 2000] and can be visualised in the results of jitter.

The analysed features are applied in the semi-automatic annotation experiments where the process as such was introduced in Section 4.1. They were used to improve the classifiers' performance that was achieved with MFCC.



(a) Intensity.

(b) Pitch.

(c) Jitter.

Figure 5.4: Global mean values for the intensity (cf. Figure 5.4(a)), pitch (cf. Figure 5.4(b)), and jitter (cf. Figure 5.4(c)). Frequency values are given in Hz. The plots visualise ES-2 and ES-5 results.

5.2.2 Results of Semi-automatic Annotation

The experimental setup of the semi-automatic annotation, this is the framework, is already discussed in Section 4.1. Especially, Figure 4.1 on page 65 explains the main idea of using audio classification to select relevant sequences in video material for facial expression annotation. As given in Figure 4.3 on page 67, only speech samples which ended not more than two seconds before the facial expression are used as training and test material. For the experiment the interindividual validation method (cf. Section 1.3.2) was applied since the aim is to annotate material that was not processed so far. Hence, the classifier has to be trained on given material which is usually not from the participant processed at the moment. Again, GMMs are the method of classification on hand as the characteristics of the speaker is modelled (cf. Section 4.3.2). In general, I distinguished three classes: i) *FACS in ES-2*, ii) *FACS in ES-5*, and iii) *no-FACS*. The latter class represents the circumstance that no relevant facial expression was shown by the participant. In this study only facial expressions are considered that are related to negative dispositions; in terms of emotions that means anger, fear, sadness, etc. This restriction is reasonable due to the small number of samples representing positive dispositions. In fact, for most of the participants no samples for positive facial expressions are at hand which can be used for training.

Table 5.5: Classification results according to the number of mixtures used in a GMM in percent (cf. [Böck et al. 2013a]). As validation methods Unweighted Average accuracy (UA) and Weighted Average accuracy (WA) are applied.

Number of mixtures	UA	WA
6	69.9	72.6
9	75.6	73.9
12	84.2	73.4
15	84.2	70.4

In the first step, I determined the number of mixtures to model the speakers' characteristics by utilising MFCCs and prosodic features as introduced in Section 4.2.2. The results are presented in Table 5.5. From this, the increase of performance with higher number of mixtures can be seen. On the other hand, having more than nine mixtures results in a loss of performance. In fact, the UA is still gaining from it, but as the data set is unbalanced in the common sense WA is the measure to be observed. With WA the effect of overfitting (cf. Section 5.1.1) is given. Furthermore, as in semi-automatic annotation the aim is to extract as many relevant sequences as possible a high performance is necessary. Taking the values of Table 5.5 on the preceding page into consideration semi-automatic annotation experiments were done with the following parameters: for each class label a single GMM with 9 mixture models was trained for five iterations (cf. Section 4.3.2) on extracted spectral and prosodic features.

Table 5.6: Classification results (cf. [Böck et al. 2013a]) in percent as Weighted Average accuracy over all interindividual experiments combining ES-2 and ES-5 (FACS in ES). The other class, no-FACS, represents the sequences where no facial expressions are shown by the participant. The secondary diagonal (lightgray cells) represent the false acceptance and false rejection rates.

		Classifier output	
		FACS in ES	**no-FACS**
Target	**FACS in ES**	61.9	38.1
	no-FACS	18.8	81.2

As I said, for the purpose of annotation, on the one hand, the performance of the classifiers is of interest; this is given in Table 5.5 on the facing page. On the other hand, the rates of acceptance are also important. In fact, both values are linked as misclassifications are influencing the accuracies and therefore, the performance. To rank a given system false acceptance and false rejection rates are calculated for a certain task. In the current issue this means, whether a relevant sequence was classified as *no-FACS* (false reject) or the other way around (false accept). Both values are shown in Table 5.6 and in addition, the average performance of the classifier itself. I merged the *FACS in ES-2* and *FACS in ES-5* classes as in the first step of semi-automatic annotation only the relevant sequence is presented to the annotator without any influence given by an optional preclassification (cf. Figure 4.2 on page 65).

The rate of false acceptance with 18.8% (cf. Table 5.6) is relatively high but in this case, a more conservative classification can be accepted. Of course, roughly 20% of the sequences are marked as relevant which increases the annotation time. Further, the loss of important sequences is measured by the false rejection rate. Due to the restricted setup where only the successive seconds of an event (i.e. *FACS in ES-2* or *FACS in ES-5*) have to be watched the additional effort is limited. However, the false rejection rate of 38.1% (cf. Table 5.6 on this page) is even worse. This means, a relatively high number of relevant sequences is discarded by the classifier. Especially, this number has to be reduced in the further research (cf. Section 7.2.1). A reduction of the false rejected sequences resulted in a even higher number of false accepted snippets. The hope is that with more material which will be available as soon as the EmoRec II data set is started to be analysed in total, the performance increases and thus, the '*false rates*' are reduced.

5.2.3 Comparison of Audio & Biophysiological Features

The data set EmoRec provides three modalities: audio, video, and biophysiological recordings which are synchronised in time. Analyses of the audio material were introduced in Section 5.1.2 and further, audio and video material was compared in Section 5.2.1. Finally, I investigated also the links between audio and biophysiological features.

As introduced in Section 3.2.2 the following biophysiological features were recorded by applying a Nexus-32 and the Biobserve software: skin conductance level, respiration, blood volume pulse, heart rate, and electromyogram. Technical details of the recording conditions are given in [Walter et al. 2011]. The sampling rate of the features was 20ms. For the purpose of classification ANNs are utilised, mainly MLPs where the parameters are adapted to each feature separately. I participated on such kind of bimodal analyses in the case of audio processing. The work on biophysiological material was done by colleagues at the Ulm University. However, the results of the cooperations are published in [Walter et al. 2011; Böck et al. 2012b].

The classification results for the audio material are given in Table 5.4 on page 114. I refer to this Table and give in this Section only the results of biophysiological characteristics.
Also, for biophysiological classification the samples of ES-2 and ES-5 are investigated. Since these are the same sequences as used in audio analyses a comparison between audio and biophysiological results is directly possible. In Table 5.7 on the facing page the recognition results of the MLPs are given. Both kinds of validation measures, namely inter- and intraindividual validation (cf. Section 1.3) are utilised. It is to be noticed that only UA is calculated for the experiments by the colleagues.
Taking into account the audio results (cf. Table 5.4 on page 114), it can be stated that in intraindividual experiments the biophysiological features outperform the classification based on audio features. For 65% of the participants the MLPs achieved recognition rates of more than 80%. This is not reached by the GMMs

Table 5.7: Recognition rates for intraindividual (intra) and interindividual (inter) classification of ES-2 vs. ES-5 in percent utilising biophysiological features applying ANNs (cf. [Böck et al. 2012b]). The results are grouped by each participant. The classification was done by colleagues at the Ulm University.

Participant	**112**	**114**	**118**	**125**	**127**	**129**	**208**	**212**	**215**	**219**
ES-2 (intra)	99.5	86.7	96.8	63.9	81.3	93.9	90.5	59.8	61.4	84.8
ES-5 (intra)	98.8	94.7	98.6	81.1	95.6	98.5	95.6	75.5	77.1	83.7
ES-2 (inter)	3.7	0.0	6.5	6.0	0.0	2.6	3.6	92.2[1]	0.0	0.0
ES-5 (inter)	6.1	1.8	34.7	30.0	17.4	20.1	2.9	18.0	6.4	12.6
Participant	**225**	**226**	**308**	**423**	**427**	**506**	**510**	**511**	**518**	**602**
ES-2 (intra)	95.5	77.5	80.9	63.4	76.4	81.9	91.6	99.1	50.2	77.0
ES-5 (intra)	92.1	79.1	87.4	65.5	68.0	90.3	96.9	98.4	67.2	84.0
ES-2 (inter)	0.0	27.5	0.0	34.6	0.0	25.4	1.8	13.9	1.0	7.9
ES-5 (inter)	51.4	9.7	4.3	1.7	15.8	3.4	4.4	54.8	36.7	7.0

based on audio samples. In contrast, interindividual validation has a quite bad performance. In particular, for ES-2 where the disposition is expressed quite weakly, the MLPs lose performance. With ES-5 the situation is slightly better. In this case of interindividual validation, audio feature based classifiers are able to generalise from the given training material. Especially, biophysiological parameters and thus, biophysiological characteristics are significantly participant dependent. Any speaker reacts differently in several dispositions in terms of body characteristics. From this, it can be seen that audio features are in such a way more universal. This issue is further considered in Section 5.2.4.

5.2.4 Discussion

The bimodal investigation of the EmoRec data set introduced the facial expressions derived from video material and further, the biophysiological features. Both sets have their own advantages which can improve the understanding of the user in HMI. So far, the two modalities are only analysed and compared. In fact, no fusion in the true sense is established, yet. Nevertheless, important information can be concluded.

[1] The participant has shown quite high reactions on the induced dispositions. This effect is also reflected by his SAM selfrating (cf. Table 3.4 on page 56). Due to this extreme expressiveness of the dispositions in the biophysiological features, a high performance in the classification could be achieved (cf. [Walter et al. 2013]).

Comparing facial expressions and audio features, a significant improvement in the annotation could be achieved, in the sense of time effort reduction. Audio analysis of the given material identified relevant sequences which only have to be considered in manual annotation. Hence, audio classification overcame disadvantages of other tools which provide labelling of AUs (cf. Section 4.1). These are based on video analysis only and thus, utilising another modality helped to free from restrictions like no fringes, no sensors in the face, etc. Of course, in cases where the user shows facial expression but no acoustic samples are given, my framework is not able to find relevant sequences. To cover such events biophysiological methods have to be considered. This is not in the focus of my research and consequently not discussed here.

In relation of facial expressions I also regarded prosodic features and especially, the formants and their bandwidths (cf. Figure 5.3 on page 119) demonstrated a significant potency to distinguish positive and negative dispositions. As it is known from literature (cf. e.g. [Vlasenko et al. 2011b]), in acted as well as in scripted non-acted material, for instance, eNTERFACE (cf. [Martin et al. 2006]) or Vera am Mittag (VAM) (cf. [Grimm et al. 2008]) the discriminating ability of these features is shown. With the analyses in Section 5.2.1 I demonstrated their discriminating ability for non-acted, naturalistic data sets as well. From my point of view: that the intensity is not as discriminative, results from the low expressiveness in naturalistic material. While listening to the samples the assumption is strengthened and investigating the spectrogram which can be extracted with Praat (cf. [Boersma 2001]) this is affirmed. Such considerations are also valid for pitch. On the other hand, jitter lacks on performance since the users are not induced to be afraid. In the case of real-life scenarios, it is hypothesised that jitter might become an indicator (cf. Section 7.2.2).
Moreover, using these features in the classification results in an improvement as it can be seen in comparing WA values for interindividual classification in Table 5.4 on page 114 and Table 5.6 on page 123. Especially, in semi-automatic annotation the mean WA is 71.6% whereas without this additional features 55.1% were achieved.

On the other hand, comparing audio classification results to these gained with biophysiological features, the already discussed advantage of audio analyses in interindividual validation is obvious (cf. Section 5.2.3). Though in intraindividual validation biophysiological analysis show its high potential. From this, I argue to utilise such analyses to collect user-dependent material automatically marked with a kind of ground truth. Such samples can be used to adapt user-independent classifiers towards a certain participant. In the mean time, classifiers based on interindividual validation can handle the HMI. Afterwards the system like a companion (cf. [Wendemuth & Biundo 2012]) switches to an intraindividual and user-dependent classification. The idea was also discussed in [Böck et al. 2012b]. Finally, an overall view on the participant can be just achieved while using all supplied modalities.

5.3 Multimodal Setup

As I already mentioned the analyses of the two data sets, Last Minute (cf. Section 3.2.1) and EmoRec (cf. Section 3.2.2), are still ongoing or are recently started. Therefore, the results which are presented in this Section are first experimental findings while exploring the data sets, though these reflect noticeable results in the domain of multimodal disposition recognition. The multimodal analysis is a subproject of the SFB/TRR 62 in which I am participating. The main work is done at the University Ulm and this results in publications where I just co-author (cf. [Walter et al. 2011; Schels et al. 2012]) because I concentrated only on the audio investigations. Therefore, I present in this Section results which are so far not published in a greater context but intended to be combined with other modalities or even with the full set of modalities which are on hand supplied by the corpora.

5.3.1 Experimental Setup

As EmoRec in total supplies several modalities, namely audio, video, and biophysiological features, this data set is selected to be analysed considering the aforementioned modalities. In addition, Last Minute also has full recordings of all modalities but just for 20 participants.

Considering the two multimodal publications two different setups can be distinguished according to the material applied in sense of utilised ESs.
In [Walter et al. 2011] the already well-known ES-2 and ES-5 were analysed regarding all modalities. For this, bimodal investigations were already presented. In Section 5.2, audio based classifiers are compared against facial expression based analyses where it was shown that classifiers on audio material can support the annotation and further, the alignment of AUs to corresponding classes. Furthermore, comparing audio based classifiers against biophysiological based classifiers

showed that both can support each other in different phases of the classification process.
More interesting are the ESs as utilised in [Schels et al. 2012]. In addition to ES-2 and ES-5 two further ESs, namely ES-4 and ES-6, are investigated. As it can be seen from Figure 3.1 on page 54, for instance, both ESs reflect other sections of the experimental scenario but can be again regarded as complementary in the case of dispositions. In this setting ES-4, *low pleasure, high arousal, low dominance*, is considered as the user is in a 'negative' disposition, and on the other hand, ES-6, *high pleasure, low arousal, high dominance*, represents the 'positive' one (cf. Figure 5.2 on page 111). Both reflect crucial sequences in the scenario. ES-4 is the breakpoint in the scenario where the participant is influenced to switch from positive to negative disposition (cf. Figure 3.1 on page 54). It is essential because the user should get the feeling that he looses the control of the system and hence, a negative disposition is induced. In contrast, ES-6 tries to evoke a positive disposition to leave the user with a kind of satisfaction from the experiment. Furthermore, the material is taken from the EmoRec II which is the second round in the scenario (cf. Figure 3.1 on page 54). So far, just a subset of participants' material is prepared to be multimodally analysed, that means, annotated and labelled. Hence, the results are based on the samples of eight participants only. The distribution of the samples in total is given in Table 5.8.

Table 5.8: Sample's distribution in EmoRec II.

Disposition	**Number of samples**
positive	373
negative	320

As features I applied the already well examined MFCCs with the 0^{th} cepstral coefficient as additional parameter by using HTK to extract these. The extraction of parameters is as follows: the frame rate is 10ms and a Hamming window (cf. Equation 4.1) with a window size of 25ms having an overlap of 15ms is applied. Further, since the characteristic of speech is to be analysed GMMs are acting as classifiers (cf. Section 4.3.2). The setup was an interindividual validation by decision of the colleagues from the Ulm University.

5.3.2 Audio Classification Results

As in the bimodal analyses (cf. Section 5.2), I cooperated with colleagues from the Ulm University. Therefore, I concentrate on the results I achieved and shared with the other colleagues.

To explore the data set, I run at first cross-validation experiments as the number of provided samples is relatively high and also almost equally distributed (cf. Table 5.8 on the facing page). Applying GMMs, I achieved 64.7% in UA and 64.1% in WA accuracy. Comparing these results to equivalent experiments on EmoRec I an improvement of approximately 10% was gained. The underlying concepts why this improvement was achieved are so far matter of further investigations of psychologists and are, from my point of view, related to the design of the WoZ scenario. For this, it can be concluded that up to now the validation method does not matter.

Table 5.9: Interindividual classification results on EmoRec II in terms of Unweighted Average accuracy (UA) and Weighted Average accuracy (WA) in percent.

Participant	UA	WA
1201	51.2	50.2
0112	45.1	43.7
0131	48.8	49.3
0202	53.4	51.9
0207	55.3	53.5
0211	61.3	61.3
0221	56.1	57.8
0224	46.4	55.2
mean	52.2	52.9

Of interest is the interindividual validation which was also investigated in [Schels et al. 2012]. The results are given in Table 5.9. The distribution of samples is again unbalanced in the common sense as it is already the case in EmoRec I. In general, the results on EmoRec II are comparable to these achieved on EmoRec I. From my point of view, this shows three aspects. At first, even in a second cycle of the scenario (cf. Figure 3.1(a) on page 54) the dispositions can be steadily induced which is also shown with biophysiological features that directly reflect the dispositional characteristics of the participants (cf. [Schels et al. 2012; Walter et al. 2013]). Further, the features which were extracted from the data are robust to reflect the dispositions. The results on EmoRec I are presented in Section 5.1.2

and Table 5.4 on page 114. For the mean values of accuracies no huge differences in the two cycles are present, for instance 52.3% UA on EmoRec I and 52.2% UA on EmoRec II (cf. Table 5.4 on page 114 and Table 5.9 on the preceding page). Finally, the values which are achieved in the analyses sound reasonable as they can be reproduced not only on different data sets (e.g. EmoRec I and Last Minute) but also within one setup or scenario (e.g. EmoRec I+II). Especially the effect of reproducibility underlines the universal characteristics of the utilised features and applied classifiers for the handling of naturalistic data sets.

5.3.3 Discussion in the Context of Multimodality

As I already stated, the achieved values are reasonable and represent a significant foundation for the purpose of fusion. In [Schels et al. 2012] a mid-level fusion architecture (cf. Section 4.4) is introduced that can cope with the three modalities on hand.

In contrast to the results in Table 5.9 on the previous page, the accuracy values in [Schels et al. 2012] are reached with a slightly different setting. The main difference is the sample time; I used for my results 25ms whereas in the paper 40ms and even 200ms are applied. This is due to the sampling rates of the other modalities but it is arguable. I argue for a smaller sampling rate as otherwise, in my opinion, the stability assumption from speech recognition is violated considering emotions. For a disposition recognition from speech larger time scales need to be investigated. Therefore, in the first step of exploring such naturalistic data sets the sampling with a higher sampling rate might be possible. Doing so, a recognition accuracy of $\approx 58\%$ was achieved for audio classifiers (cf. [Schels et al. 2012]). Fusing audio and video or audio, video, and biophysiological classifications for a final decision the recognition could be improved to $\approx 61\%$ (cf. [Schels et al. 2012]). This means, the multimodal recognition obtained a roughly 10% better performance as the classifier which is based on audio only (cf. Table 5.9 on the preceding page).

From these results the importance of a multimodal observation of the user becomes apparent. Of course, this causes that more multimodal data sets have to be generated but also that in HMIs the investigation of the user has to be driven multimodally. On the other hand, such multimodal observations create also problems. The main aspect is the synchrony of the data collection. The more modalities are involved the more complex is this issue. Furthermore, every fusion architecture provides advantages and disadvantages which have to be traded for each classification task. In particular, even in one modality several classifiers can be fused as already discussed in Section 4.3.4. In general, the issue of multimodal fusion is a task for further research and especially, with the focus on disposition in HMI (cf. Section 7.2.3).

5.4 Summary

Starting with an audio only setup I analysed the two corpora recorded in the SFB/TRR 62, namely Last Minute and EmoRec. Since both data sets are naturalistic HMI generated in a WoZ setting the accuracy values are not as high as achieved with acted material. Nevertheless, they are comparable to those gained on other naturalistic data sets (cf. [Schuller et al. 2011a]).

In the bi- and multimodal parts I related the audio based recognition to the other modalities of the data sets, in particular, facial expressions and biophysiological features. With these additional features a more general analysis of the participant could be realised. Particularly, in the multimodal Section I presented novel results on the EmoRec II data set which are so far not published. In general, multimodality increases the ability to recognise the user's dispositions and further, to react on these in a proper way. An additional aspect of disposition is the involvement which is an integral part of an interaction (cf. Section 1.4). So far, this has not been considered but will be discussed in the following Chapter.

Chapter 6

Group Involvement

Contents

6.1 Why Consider Involvement in Group Conversations? . . 134

6.2 Annotating Changes in Group Involvement 135

6.2.1 Annotation of Day1 - Group 136

6.2.2 Annotation of Day1 - Each Participant 143

6.3 Detecting Changes in Group Involvement 147

6.3.1 Automatic Detection of Changes 147

6.3.2 Systems Reaction towards a User 151

6.4 Summary 152

SO far, in this thesis I considered the process of disposition recognition from speech in the context of near real-life, which is, for instance, dispositional speech by non-actors in a laboratory environment (cf. eNTERFACE), and naturalistic speech (cf. Last Minute and EmoRec). For this, the focus was on the classification and recognition of dispositions which usually occur in HMI, for example, positive or negative reactions towards the system, emotional behaviour of the user, etc. On the other hand, dispositions can be seen much broader, in the sense of user behaviour, situation, and intention (cf. also Definition 1.6 on page 7). Such an aspect is further the involvement in a conversation whereas this is either an HHI or an HMI. As already introduced in Section 1.4, a conversation is usually a multi-party interaction. Therefore, a data set that supplies an interaction of several parties is to be used, which is the case with the TableTalk corpus (cf. Section 3.3).

In this Chapter, I further motivate the detection of involvement as this is, from my point of view, complementary to disposition and yet another source of information to interpret the user's characteristics. As this topic is quite novel the preprocessing of the TableTalk data set, in the sense of group involvement annotation, is introduced and the labelling results are given. Furthermore, the reliability of the annotation is discussed. Finally, first results in the detection of changes in group involvement are presented. In connection with those the focus shifts from the recognition of the user's disposition towards a system-centred view in a sense how such information can be used to control a technical system (cf. Section 1.5).

6.1 Why Consider Involvement in Group Conversations?

As stated in Definition 1.7 on page 18, there is a narrowed and at once a more general meaning of involvement in the case of speech processing; namely that a user is participating in a conversation. I already discussed (cf. Section 1.4) that involvement reflects a kind of situation and thus, is also subsumed by situatedness (cf. Definition 1.5 on page 7). Therefore, it can be seen as a part of the user's behaviour which is influenced by his disposition. Involvement in a conversation is something I called meta-information or meta-analysis. There is not a direct connection from a signal to a reaction like in emotions, for instance. Of course, sensors can measure movements or voice activity, but the more interesting effect is when a user is not involved. In this case, the interaction might be in a crucial situation focused on a certain participant. Especially, in a dyadic conversation or in an HMI this effect can indicate a breakdown in the dialogue. So far, it is not under research as it should. Hence, I argue for a deeper investigation of situations related to the absence of involvement. For this, meta-analyses by psychologists and linguists are necessary and thus, cannot be accomplished in this thesis.

So far, usually, a kind of dyadic situation in the HMI is considered, namely one user interacts with one system. For such an interaction the aforementioned statements and the following ones are valid. Additionally, communicative situations, especially, in group conversations are more complex. This leads to constellations like a group of users and one system. Hence, the detection and classification of changes in involvement culminates in a multi-party HMI. Up to now, I conducted my analyses regarding only one user, however, I will go further considering also multiple user settings. For this, in the following several aspects of the involvement and its detection are mentioned – keeping in mind that these are relevant for single user and multi-party HMIs.

In a purely technical sense, detecting that a member of the group starts getting involved should be recognised by the system, for instance, by controlling a certain microphone or focussing a camera to this participant as it is done (manually) by conference equipment. Moreover, the other way around means detecting whether some kind of technical equipment should be shut-off because the user is not interacting with the system any more but with another person or device.

These ideas and considerations indicate the necessity to detect involvement in a group conversation. And again, I do not distinguish between a multi-party HHI or a multi-party HMI. The only assumption so far is that only one technical system or companion-like system (cf. [Wendemuth & Biundo 2012]) is comprised. In general, the remarks I gave in this Section are valid in the analysis of HMI but further in HHI, too. In the TableTalk data set as introduced in Section 3.3 the interaction is between humans in total. Nevertheless, investigating those conversations may help to improve real-life HMI. Furthermore, these kind of analyses may provide insights towards Machine-Machine Interactions (MMIs) (cf. Section 6.3), which is also covered by the ideas of technical companion-like systems (cf. [Wendemuth & Biundo 2012]). However, data sets with reliable annotations considering involvement and its change are necessary. In Section 6.2 I present the effort that was taken to preprocess TableTalk to fit conditions for detection of involvement. By its nature such an analysis is technically more difficult and is thus, matter of future research (cf. Section 2.5), too.

6.2 Annotating Changes in Group Involvement

As stated in Section 6.1, involvement is an important additional information to assess the behaviour of a user. Especially, in a multi-party interaction the fact of being involved is influencing further analyses of a participant. Before any kind of automatic system can at all detect changes in the involvement a preprocessing for training and thus, a preparation of the data set is necessary. Foremost the data has to be collected and afterwards be annotated (for an annotation process as such cf. Section 4.1.1). The TableTalk corpus was collected by Nick Campbell [Campbell 2009] and in the case of video analyses postprocessed (cf. [Campbell 2009]) according to [Campbell & Douxchamps 2007]. No preparations of the corpus with respect to audio processing and involvement detection are available. In the following, I present and discuss the results of the data preparation, especially, with the focus on the annotation of changes in involvement. Besides the data set was not collected by myself, the annotation and the analyses according to reliability are my work done with co-authors listed in the references of [Bonin et al. 2012] and [Böck et al. 2013b][1].

6.2.1 Annotation of Day1 - Group

As mentioned in Section 3.3, *day1* is already postprocessed as described in [Campbell 2009], namely providing features like head and body movement, head activity, etc. in the case of video material and further, a literal transcription exists. Unfortunately, the whole data set lacked of an annotation according to dispositions as well as involvement. To characterise the group, especially in detecting changes (cf. Section 6.3), the latter is the interesting parameter. Annotation for the issue of disposition is not considered, yet.

In [Oertel et al. 2011b] a scaled annotation scheme is presented which handles the involvement according to a fixed scaling in the range of $[0, 10]$. From this annotation procedure, changes in the involvement result in a switching between the levels given by the scaling. Thus, changes are only realised in the sense of discrete, rather large level differences. In the following, I introduce the annotation process which I adopted and which is focused on the changes of involvement in an interaction.

In cooperation with the Trinity College Dublin I headed the annotation process for group involvement. As published in [Bonin et al. 2012; Böck et al. 2013b] six advanced psychology students[2] of the Otto von Guericke University Magdeburg annotated changes in the involvement for both, group in total and each participant separately, watching the audio-visual material. Thus, an increase of involvement is marked with $+$, whereas a decrease is annotated with $-$. Moreover, I assumed that the intervals without any label are considered as stable (i.e. no change) represented by *0* (cf. [Böck et al. 2013b]). This assumption is suitable as the annotation task was to mark only changes. Hence, non labelled parts of *day1* includes no changes, otherwise those were marked. Based on these considerations the implicit label *0* can be seen as meaningful. To evaluate and visualise

[1] The remaining part of this Chapter is based on the paper [Böck et al. 2013b] and thus, parts of the text are rephrased or taken verbatim from the article.

[2] It has to be noticed that in the context of annotation the students got credit points which are necessary for their study. This was the reward for annotating the material. Beyond that, no further reward or payment was given.

the annotation I applied the following assignment: $+1$ increasing involvement, -1 decreasing, and 0 no change.

While annotating a corpus the question of reliability arises. To ensure qualitative labels, as aforementioned, this aspect has to be discussed as well. To handle the enormous amount of data (cf. Table 6.2 on page 141) and, which is more important, the labelling of the six annotators, I decided to apply Krippendorff's α [Krippendorff 2012] assessing the quality of annotation. For this, the ratio of the observed and expected disagreement between annotators is calculated as stated in Equation 6.6. With most of the other reliability measures no distance measure can be applied which leads to less comparability of different label paradigms. A common implementation of α is given by [Hayes & Krippendorff 2007]. Further, I implemented the computation of the observed agreement A_o and the expected agreement A_e in Matlab. A_e represents the expected value of an agreement if all annotators would guess based on the distribution given by the data itself. According to [Artstein & Poesio 2008] A_e is computed as

$$A_\mathrm{e} = \sum_k \left(\frac{1}{ic} \sum_i n_k^{(i)} \right)^2 \tag{6.1}$$

where k is the number of categories or labels, i is the total number of samples, c is the number of raters, n_k is the number of ratings according to a category, and ic represents the total number of assignments. From this, A_e is the weighted absolute frequency of the labels given by the annotators.

In contrast, A_o is the agreement which is given by inspecting the annotations; that is the percentage of pairwise agreements over all annotators and all samples. Given the variables as aforementioned in Equation 6.1, the observed agreement is calculated as follows

$$A_\mathrm{o} = \frac{1}{ic(c-1)} \left(\sum_i \sum_k n_k^{(i)} (n_k^{(i)} - 1) \right), \tag{6.2}$$

where the normalisation is done by $\frac{1}{ic(c-1)}$ which is the total number of assignments grouping the number of coders pairwise.

Given the values of A_o and A_e the most general reliability κ_g is computed as a ratio of both

$$\kappa_\mathrm{g} = \frac{A_\mathrm{o} - A_\mathrm{e}}{1 - A_\mathrm{e}} \tag{6.3}$$

Using Equation 6.3 and taking into account that the disagreement values which are substantial for the computation of Krippendorff's α are defined as

$$\begin{aligned} D_{\mathrm{o}} &= 1 - A_{\mathrm{o}} && (6.4) \\ D_{\mathrm{e}} &= 1 - A_{\mathrm{e}}, && (6.5) \end{aligned}$$

where D_{o} is the observed disagreement and D_{e} is the expected one, α is generally calculated as

$$\alpha = 1 - \frac{D_{\mathrm{o}}}{D_{\mathrm{e}}} \tag{6.6}$$

As I introduced the equations to compute the reliability values, I give the current values for the group annotation on *day1*. The observed agreement $A_{\mathrm{o}} = 0.4348$ and the expected agreement $A_{\mathrm{e}} = 0.4107$. The expected agreement is in the context of reliability considered as the lower bound of agreement which could be achieved if the annotators guess the labelling. Therefore, the observed agreement A_{o} has to be larger than A_{e} because, in general, the annotators do not perform a pure guessing but provide labels which are related to the given material. Hence, they annotate the material in a similar way that leads to a larger observed agreement. Taking the six annotators and the 20541 samples (cf. Table 6.2 on page 141) into account $\alpha_{\mathrm{n}} = 0.0408$ which is the *nominal* version of Krippendorff's α. Krippendorff's α, in detail, applies a distance measure to weight the differences in the ratings. The nominal version α_{n} does not distinguish between different labels; that is, uses an equal distance measure and thus is computed as in Equation 6.6.

As discussed in [Artstein & Poesio 2008] this is not feasible because it is quite a difference whether two annotators label, for instance, either -1 and 0 or -1 and $+1$. For this, a distance measure which do not assume equal distances is introduced. Therefore, according to [Artstein & Poesio 2008] the expected disagreement D_{e} and observed disagreement D_{o} can be written as follows:

$$D_{\mathrm{e}} = \frac{1}{ic(ic-1)} \sum_{i} \sum_{l} n_k^{(i)} n_k^{(l)} d_{il}^{(k)} \tag{6.7}$$

$$D_{\mathrm{o}} = \frac{1}{ic(c-1)} \sum_{k} \sum_{i} \sum_{l} n_k^{(i)} n_k^{(l)} d_{il}^{(k)} \tag{6.8}$$

where $d_{il}^{(k)}$ is the distance measure which weights the differences in the labels. The other symbols are defined as already mentioned in the agreements. Using the disagreements as given in Equation 6.7 and Equation 6.8 the *ordinal* Krippendorff's α_o is calculated according to Equation 6.6. Considering this, I concentrated on α_o, again with all annotators and all samples. Thus, I obtained $\alpha_\mathrm{o} = 0.1562$.

Table 6.1: Values of the observed agreement A_o, the expected agreement A_e, and Krippendorff's α (nominal α_n and ordinal α_o) on the TableTalk corpus.

A_o	A_e	α_n	α_o
0.4348	0.4107	0.0408	0.1562

Let us consider the reliability values (cf. Table 6.1) in more detail as also partly done in [Böck et al. 2013b]. In fact, there are two ways to analyse and to assess annotation quality, namely i) based in the classes and ii) based on the annotators' agreement. For class level, the chance level can be estimated under consideration of the given distribution of labels. In the current annotation task three labels can be assigned by the annotator, namely +, −, and *0*. For each class the probability to get an assignment for it can be estimated as follows (the calculation is done exemplarily for the class 0):

$$Pr(\text{class } 0) = \binom{6}{4}\cdot\left(\frac{1}{3}\right)^4\cdot\left(\frac{2}{3}\right)^2+\binom{6}{5}\cdot\left(\frac{1}{3}\right)^5\cdot\left(\frac{2}{3}\right)^1+\binom{6}{6}\cdot\left(\frac{1}{3}\right)^6\cdot\left(\frac{2}{3}\right)^0 . \quad (6.9)$$

Taking into account all three classes the chance level is calculated as

$$Pr(\text{all classes}) = \left(\frac{1}{3}\right)^5\left[4\cdot\binom{6}{4}+2\cdot\binom{6}{5}+\binom{6}{6}\right] = 0.3004. \quad (6.10)$$

Thus, the chance level based on the distribution of classes is 30%. From Table 6.2 on page 141 it can seen that the percentage of labels assigned to the three classes is 67.7% which is more than twice the change level. On the other hand, the class based chance level estimation do not consider any weighting of the assignments. Hence, in the agreement or more specific the disagreement this kind of weighting is incorporated, especially since the assignments are assessed pairwise. This means, if two annotators just slightly differ in their assessments the expected disagreement (cf. 6.7) is lower than in cases of larger differences. Consequently the agreement between the annotators is increasing. From these considerations, the expected agreement as well as the observed agreement provide a more experienced ranking of the annotation's reliability.

On the one hand, there is no huge difference between A_o and A_e. On the other hand, the main disadvantage of the reliability measures is that they assume usually a bias within the annotators, only. But in case of TableTalk, it is also a bias in the material itself as there is a trend to an increase of involvement over time. This aspect is neither reflected by Krippendorff's α nor other reliability measures as shown in [Devillers et al. 2005], where an HHI is observed as in the particular case. However, as it can be seen from Table 6.2 on the next page, a large amount of samples is positively assigned to the three classes; namely 67.7% of all samples are assigned to a distinct class applying a majority vote to get class labels; that is, at least four annotators assigned a sample to the same class.

For the sake of comparison, I rank the reliability also to those given in [Oertel et al. 2011a] reflecting the reliability on the D64 data set (cf. [Oertel et al. 2010]). Unfortunately, the authors give only a κ value. Nevertheless, κ and Krippendorff's α are slightly comparable. Oertel et al. obtained $\kappa = 0.56$. This result is better than the reliability on TableTalk, but the annotation process of [Oertel et al. 2011a] is much more restricted and thus, the raters are more guided as in my study. Hence, this is reflected by a larger κ value.

Based on these considerations, I set up a classification task using MLPs (cf. Section 6.3). Therefore, the labels of the six annotators had to be combined. The merging is done due to an 1-of-N coding of the three classes. To get the combination I summarised all assessments of all annotators and divided the sum by the number of raters for weighting purpose. Hence, I got ratings for three distinct classes (i.e. -1, 0, and $+1$; cf. Figure 6.1 on page 142), but also ratings which are undecidable (there is no majority for any label by the annotators). This especially occurs because a majority vote (more than four annotators has to agree to a certain class) is applied to get the classification labels. To handle the cases where no majority vote could be established, the additional class *undecidable '?'* was introduced. Thus, in total the classification task is a four class problem – three classes provided by the annotators and the additional class handling the undecided samples.

In Table 6.2 on the next page the distribution of the annotator's labels in *day1* is given. Details of the experiment can be found in Section 6.3.

Furthermore, the labelling provides also statements which I will call meta-analyses. In Figure 6.1 on page 142 the distribution of the samples for each weighting value is given. The ordinate values are crisp in $\{0, \frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}, \frac{6}{6}\}$ representing the values of weighted assignments to each class. Each bar in Figure 6.1 on page 142 shows the distribution of those assignments to a class where the num-

Table 6.2: Distribution of samples per label in total numbers and percent, whereas -1 represents decreasing and $+1$ increasing involvement, respectively. Further, 0 indicates no changes and '?' denotes of the undecided class. Moreover, the total time for all samples per label is given.

Label	Number of samples	Percentage	Overall time
-1	1999	9.7	03:20.21
0	28	0.13	00:02.68
$+1$	11879	57.8	19:52.99
?	6635	32.3	11:06.67
total	20541	100.0	34:24.00

ber of samples in a certain distribution sums up to the total number of samples (cf. Table 6.2). Thus, the following remarks can be given from the plot: i) the bars on $\frac{3}{6}$ – which reflects that in each case two annotators vote for -1, 0, and $+1$ – are always related to the undecidable class because no majority vote can be done. Fortunately, the amount of samples which are given there is quite small and further, the disagreement originates from the distinct classes $+$ and $-$; that is, either one of these classes was selected by the annotators. On the other hand, the amount of 0 samples is negligible. ii) Moreover, the plot has two further areas corresponding to the certainty of the annotator: a) being certain that a current class cannot be assigned to a sample (these are the values $\{0, \frac{1}{6}, \frac{2}{6}\}$), and b) being certain that an assignment is applicable (these are the values $\{\frac{4}{6}, \frac{5}{6}, \frac{6}{6}\}$). From this, one can see that especially for class $-$ and class 0 the annotators tend to be certain that this is not the correct class for a sample. But, it cannot be instantly said which class is favourably selected for a sample. iii) The other way around, a high number of samples are certainly assigned to $+$, however, with a much broader distribution. It varies over the full range of $\{\frac{4}{6}, \frac{5}{6}, \frac{6}{6}\}$. Additionally, by the design of the annotation task the following can also be seen from Figure 6.1 on the following page: To achieve an assignment on $\frac{4}{6}$ for class $+$ three assignments to the two remaining classes are possible, namely a) 0 annotations for $-$, 2 annotations for 0, b) 1 annotation for $-$, 1 annotation for 0, and c) 2 annotations for $-$, 0 annotations for 0. Similar analyses can done for $\{\frac{5}{6}, \frac{6}{6}\}$. These considerations explain the high number of labels in $\{0, \frac{1}{6}, \frac{2}{6}\}$ for class $-$ and 0 (cf. Figure 6.1 on the next page).

Another meta-analysis considers the combinations of assignments reflecting the 'degree of non-agreement'. In fact, the two investigations (cf. the distribution of labels) are related to each other. In Figure 6.2 on page 143, the combinations

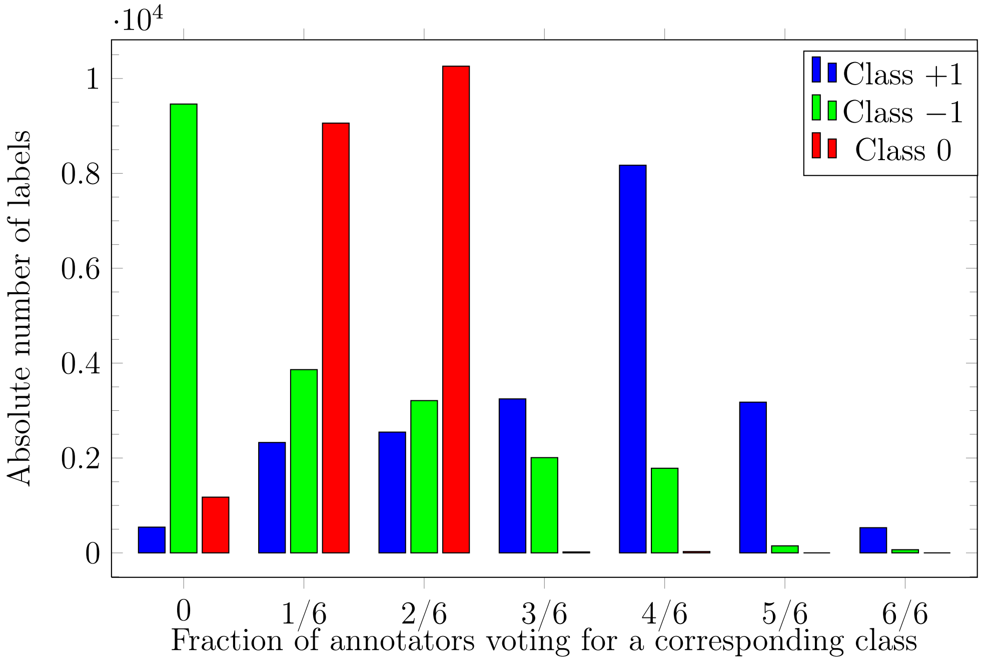


Figure 6.1: Distribution of the labels for the annotator's voting for the group in total.

of numbers of assignments to each of the three classes – class +, class −, and class *0* – are visualised. In addition to Figure 6.1, this representation provides the possibility to distinguish the assignments that generated the fractions shown in Figure 6.1, especially for $\{\frac{3}{6}, \frac{4}{6}\}$. The consideration of combinations starting with 1 or 0 is not necessary since these are counted for the other classes in the specific combination. Both analyses are similar in their results. But, especially the combination (2 2 2) indicates if the annotators do not tend to a clear decision. Because this particular combination represents the case where no majority vote could be established. The number of assignments resulting in (2 2 2) is relatively low; namely just 525 samples were rated according to this combination. The number of samples is equal for all three classes since to obtain this combination the assignments have to be equal for the three classes. In contrast to $\{\frac{2}{6}\}$ (cf. Figure 6.1), the combination (2 2 2) does not count any side effects which might are given by the annotation task (cf. discussion above) but only reflects that the annotators could not tend to or decide for any class in total. For the other combinations the results are comparable to those already discussed in the context of Figure 6.1. In general, as both analyses are similar and moreover, also complement each other, I would suggest to consider the distribution of labels as well as the combinations of assignments to assess the quality of an annotation and the corresponding class assignments.

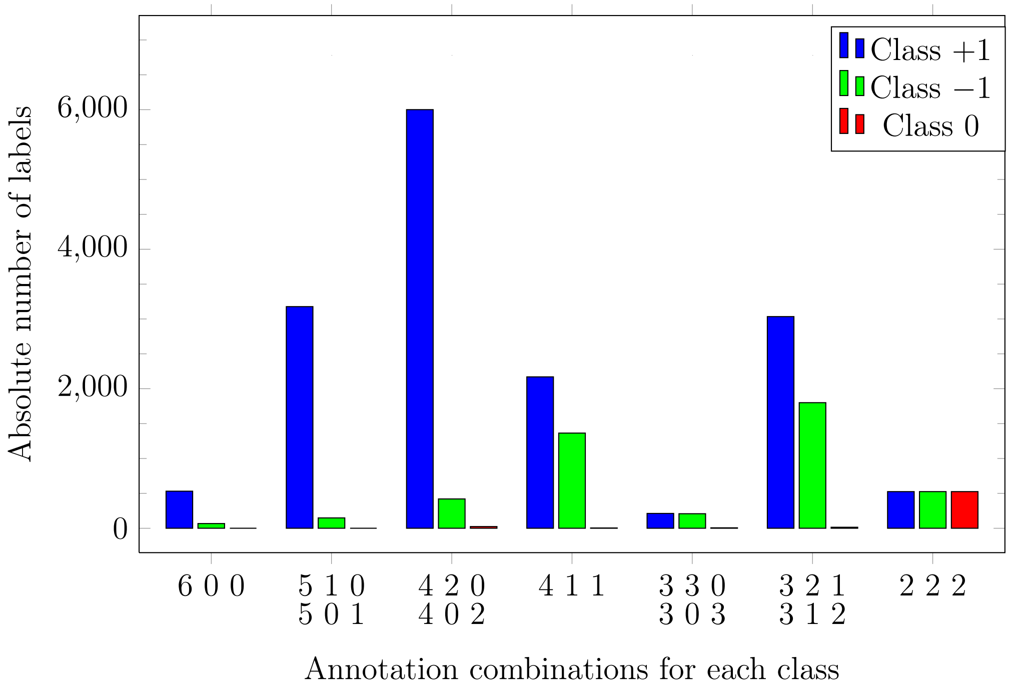


Figure 6.2: Combinations of numbers of assignments for each class. The first number represents the number of assignments to the corresponding class whereas the remaining two numbers determine the assignments' number to the other classes, for instance, (4 1 1) means 4 assignments for class +1 and 1 assignment for class −1 and class 0, respectively. For four combinations the the two alternatives in the combinations have been considered as equal and thus, are counted together.

It is obvious that such meta-analyses are specific to the corpus and the annotation method of the material. Nevertheless, given the data the interpretation is feasible. However, it is a matter of future research (cf. Section 7.2.4) to derive clear indications and detailed analyses of involvement changes in group interaction, whereas it does not matter whether they are related to HHI or HMI.

6.2.2 Annotation of Day1 - Each Participant

As one part of this thesis deals with the detection of changes in the group's involvement in multi-party interactions I discussed the annotation of TableTalk according to this purpose in detail. On the other hand, the data set was also labelled for each participant separately, in particular, his involvement in the conversation.

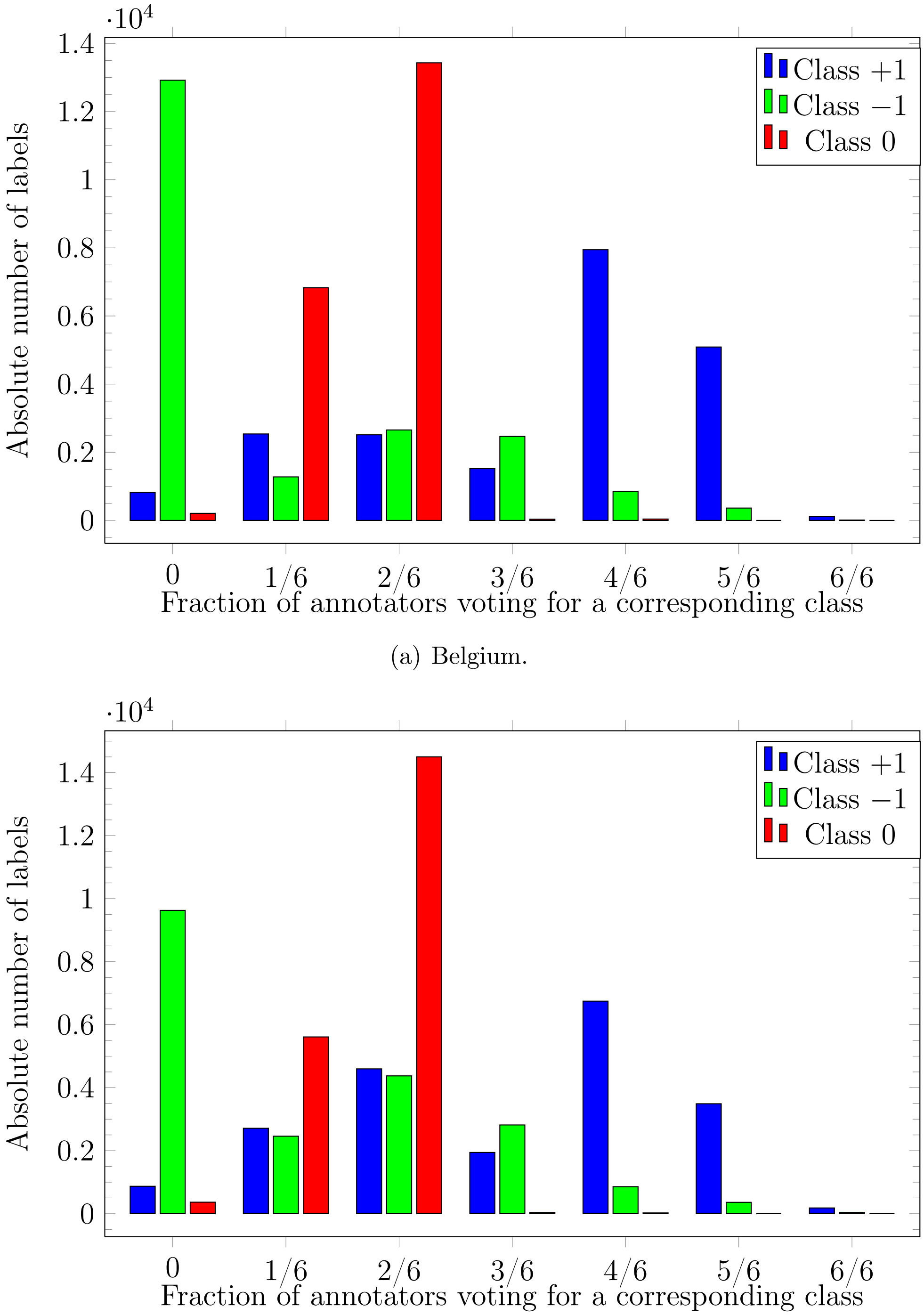
The annotation setup was the same as for the group's labelling whereas in each annotation iteration only one participant was regarded. I asked the annotators insistently to ignore the group's behaviour and the reactions of the other participants. As the single involvement is not issue of this thesis I will just briefly introduce the outcome of the labelling as it is related to the group annotation and also supports the interpretation of this labelling.

The reliability values as introduced in Equation 6.1, Equation 6.2, and Equation 6.6 are given in Table 6.3. As it can be seen they are in the same range as for the group in total. On the other hand, except for the Belgian participant, the annotation task seems to be harder observing only one participant since the group has to be neglected. Therefore, the estimation of involvement has to rely on individual characteristics of the participant. Hence, the annotators are much more discordant than in the group labelling. In contrast, the Belgian group member is, by watching the material, a quite relaxed and predictable participant in terms of his behaviour. This results in a better agreement of the annotators.

Table 6.3: Reliability values according to Krippendorff's α (nominal and ordinal), the expected agreement A_e, and the observed agreement A_o for each participant in the group conversation of TableTalk.

Participant	α_n	α_o	$\mathbf{A}_\mathrm{e}$	$\mathbf{A}_\mathrm{o}$
Belgium	0.0542	0.2241	0.4232	0.4545
Finland	0.0301	0.1631	0.3956	0.4138
Great Britain	0.0364	0.2005	0.3850	0.4075
Japan	0.0309	0.1822	0.3788	0.3980

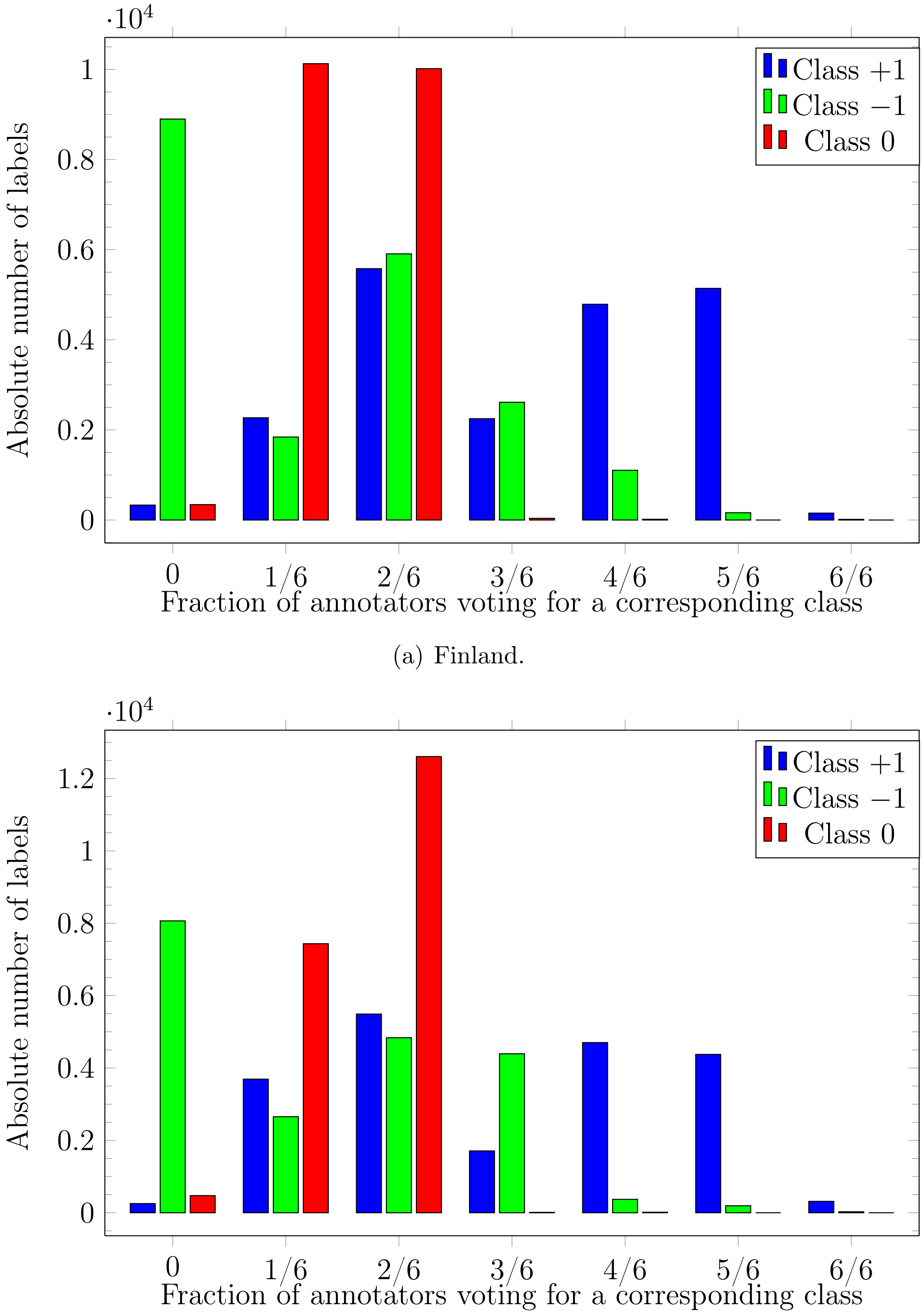
As for the group (cf. Figure 6.1 on page 142) I also present the distribution of the samples according to a 1-of-N coding. For that purpose, I grouped the distributions regarding the gender of the participants which leads to Figure 6.3 on the facing page with male participants and Figure 6.4 on page 146 showing the female participants. The distributions are quite similar to each other and to the group's one (cf. Figure 6.1 on page 142). Nevertheless, slight differences are on hand, especially, in the distribution of the + labels. For the female participants it tends towards an equal distribution whereas the male ones are more fixed to



(a) Belgium.

(b) Great Britain.

Figure 6.3: Distribution of the labels for the annotators' voting for each participant separately. The two diagrams show the results for the male participants.



(a) Finland.

(b) Japan.

Figure 6.4: Distribution of the labels for the annotators' voting for each participant separately. The two diagrams show the results for the female participants.

the classes $\{\frac{4}{6}\}$ and $\{\frac{5}{6}\}$. For the labels $-$ and *0* similar characteristics as in the group rating can be found.

Again, from these analyses no general remarks can be given, so far. This is just a starting point for such kind of analyses. Therefore, it needs more naturalistic multi-party corpora and thus, further research to extract general rules or conditions to assess changes in the involvement. Here, I laid the basics and I will discuss open issues and further questions in Section 7.2.4. From this, a kind of roadmap for future research can be established.

6.3 Detecting Changes in Group Involvement

Analysing and detecting changes of involvement is so far an upcoming part of research and thus, to the best of my knowledge, just a few investigations were done in this field, yet (cf. Section 2.5). Especially, considering involvement as complementary to disposition is a quite novel approach and the circumstances as well as the conditions to retrieve it were discussed in Section 6.1. From this, in the current Section a first study on automatic detection of changes in group involvement is presented and finally, discussed. My investigations are based on the TableTalk corpus. Furthermore, a roadmap towards detailed investigations of involvement is given in Section 7.2.4.

On the other hand, automatic detection of involvement is effecting the way a system is influencing its user. In particular, in companion-like technologies as intended by [Wendemuth & Biundo 2012] the system's reactions are an important part of the interaction. Of course, this is influenced by the user's involvement as discussed in Section 6.3.2. However, before a system might interact in such an intended way it is necessary to detect changes in the involvement properly. First experiences are presented in the following.

6.3.1 Automatic Detection of Changes

Up to my knowledge, as also discussed in Section 2.5, an automatic detection of changes in involvement is a novel approach. Hence, no comparable results and already appraised methods are available. Only a general remark from other corpora like D64 (cf. [Oertel et al. 2010]) can be given, namely that involvement can be grouped by levels, and thus, transferred to TableTalk. However, the involvement as such was not considered as complementary to disposition, yet, and further, no changes of involvement. For this, in the current Section novel and therefore, first results on automatic detection of involvement changes are presented. The work as done in cooperation with Stefan Glüge and published in [Böck et al. 2013b].

Experimental Setup

As introduced in Section 3.3 the TableTalk corpus is already postprocessed according to seven different video features (cf. [Campbell 2009]), in particular, these are (x,y)-coordinates of head's position, head's area, (x,y)-coordinates of head's motion considering only the changes of the head's coordinates in the two dimensional picture, body activity, and head activity applying a Viola-Jones face tracker (cf. [Campbell & Douxchamps 2007]). The output and thus, the features are provided by the distributor of TableTalk. For the current experiments the features for each participant of the conversation were individually handled, that means, from the given data which framewise contains the full information of all participants I extracted the relevant values and compiled those for each speaker separately. It results in $4 \cdot 7 = 28$ video features as four participants are interacting in *day1* of TableTalk. As the video features are extracted every 100ms (cf. [Campbell 2009]) a total number of 20541 samples is achieved. So far, only *day1* can be analysed since this is the part which is labelled according to involvement (cf. Section 6.2), yet. The classes are $+1$, -1, 0, and '?' which represent the labels $+$, $-$, *0*, and *?*, respectively, (cf. Section 6.2.1) and the distribution of samples to each class is given in Table 6.2 on page 141.
Remarks according to the usage of audio features for the detection are given in the discussion part of this Section.

For the task, detecting changes in involvement, so far MLPs are utilised. The setting of the ANN is as follows: The MLP had 28 input units which refer to the 28 features extracted from the video signal. Further, there is one hidden layer whose number of units was varied in the range of $[5, 55]$ with a step size of 5. An optimum was achieved on the test set with 30 hidden units. For each class a single output unit is provided.
In the hidden layer the hyperbolic tangent function tanh was applied as the transfer function. In contrast to the input units which had a linear transfer function, the output units' transfer function was the logistic sigmoid (cf. Equation 4.19). The whole network was trained with the Levenberg-Marquard al-

gorithm (cf. [Marquardt 1963]) which is an optimisation approach for non-linear problems utilising the least-squares method. In the experiments the corresponding implementation of this method in Matlab was applied. For training an intraindividual validation strategy was applied according to Section 1.3.3. Though, not the intraindividual material of one speaker was used, but the material of the whole group as such. This is possible since the detection of involvement changes for the group in total was intended. The training and optimisation was done mainly by Stefan Glüge whereas I provided the annotation and feature handling.

Results

In Table 6.4 on the following page the accuracies (WA and UA) of the detection task using an MLP are shown. With 67.8% WA a result was achieved which is high above chance level assuming an equally distributed decision; that is, 25% in a four class task. Unfortunately, so far the result has to be compared to chance level as, to my knowledge, no other values for this issue, especially, on TableTalk, are available. Of course, multimodal analyses of meetings were done but usually in the sense of group actions like pointing gestures, presentation styles, etc. (cf. [Carletta et al. 2005]), especially, on the AMI meeting corpus (cf. [McCowan et al. 2005]). On the other hand, such results are not feasible for a comparison as the classification and detection task is quite different and thus, results given in [Carletta et al. 2005], for example, cannot be used to rank the achievements related to involvement.
On the other hand, the D64 corpus (cf. [Oertel et al. 2010]) deals with involvement. In [Oertel et al. 2011a] the authors present prediction results for involvement and achieved a mean accuracy of 66.4% on a two class task and 60.6% for a three class application (cf. Table 3 in [Oertel et al. 2011a]) using video features only. As given in Table 6.4 on the next page, on TableTalk an accuracy of 84.2% UA was achieved considering all four classes. To be specific, Oertel et al. utilised less features, namely gaze and blinking rate, to detect involvement. Further, in the current experiment the changing of involvement was regarded. Therefore, both tasks are slightly comparable and due to the differences in the settings both results are somehow equal in ranking. However, the change detection task is slightly harder since no fixed levels as in [Oertel et al. 2011a] are applied in annotation.

For a detailed investigation, in Table 6.5, the average confusion matrix of the ten folds is given. This shows that, in particular, the classes are not as confused

Table 6.4: Classification results in percent applying a 10-fold cross validation. As measures the Weighted Average accuracy (WA) and Unweighted Average accuracy (UA) on the test set is presented according to [Böck et al. 2013b].

Feature	**UA**	**WA**
Video	84.2	67.8

with the 0 one, but on the other hand, '?' is proned to be confused with +1. Furthermore, it is quite obvious that classes with a small number of samples resulted in smaller accuracy; namely 62.4% for −1 and 50.0% for 0. Again, as this kind of research is still starting, an improvement of accuracy in detection of involvement changes is to be expected.

Table 6.5: Average confusion matrix and class-wise accuracy (acc.) in percent on the test set for the 10 folds of the experiment. It is to be noticed that the values of the matrix are rounded to the corresponding nearest integer value. Thus, the average overall accuracy (WA) slightly differs from the one given in Table 6.4 on the current page. The Table is taken from [Böck et al. 2013b].

		MLP output				
		+1	**−1**	**0**	**?**	**acc.**
Target	**+1**	1068	10	0	104	90.4
	−1	43	132	0	30	64.5
	0	1	0	1	1	33.3
	?	126	12	0	529	79.4

Discussion

The presented experiment of detecting changes of involvement in group interaction is a first presentation of such analysis done on the TableTalk corpus. The classification is based on the annotation provided by myself (cf. Section 6.2.1) and the video features prepared by the TableTalk's distributor (cf. [Campbell 2009]). A postprocessing of the features was done by myself.

Up to now, only video features are working well on the data sets though audio samples were also utilised for a detection. Due to several reasons, which I discuss in the following, the detection performance was less than 30% WA. First of all, by design of TableTalk only recordings of one microphone are available covering the speech of all four participants simultaneously (cf. Section 3.3) which will be called in the following ‘combined audio signal’. In the processing of the group in total this should not influence the performance, however, for each participant separately no high quality audio samples are on hand. As the speech is highly mixed from all participants it results in a scenario where changes in the group’s involvement are depending on single participants. Unfortunately, due to the combined audio signal it is not possible to handle these slight differences; that means, they are hidden in the overall discussion. On the other hand, the features, namely MFCC, are not able to cope with the nuances of involvement changes that are hidden in the mixed signal. From this, alternative feature sets or enhancements of existing features for speech processing have to be found or developed. Better speech features may also be used in combination with video features in a multimodal involvement classification which is expected to reach better results since it is able to make use of the best available features at each point in time, overcoming less meaningful features in one modality by extracting information from the other. These issues are also reflected in the roadmap for future research on detection of changes in involvement in group conversations (cf. Section 7.2.4).

Furthermore, a robust involvement detection for single participants in an HHI might be helpful, but this issue is more important in HMI, especially, if multiple users are interacting with a system. Moreover, this information can be used by a technical system in the sense of a companion (cf. [Wendemuth & Biundo 2012]) to react in a proper way to its user and further, influence him. Both aspects are discussed in Section 6.3.2.

6.3.2 Systems Reaction towards a User

In the systems reaction towards a certain user two aspects have to be considered: i) the systems proper reaction itself and ii) the influence of the system towards the user.

The first aspect is quite on hand because an HMI as such is oriented towards a reaction on a user. For this, detection of involvement can be utilised as an additional information, especially, to evaluate if a user is interacting with the system or not. The latter is important if the participant is situated in an environment which allows multiple interactions, for instance, the user is part of a group or handles several devices, etc. As it can be seen, the environment contributes an influence to an interaction even if it is passive; that means, no interaction of the environment is headed to the technical system. This is also reflected in the definition of situatedness (cf. Definition 1.5 on page 7) and thus, is included in the term disposition (cf. Definition 1.6 on page 7). Incorporating information retrieved from the involvement enables the system to decide whether a user's action requires system reaction or can be ignored. Moreover, it can adapt its response to the situation, for instance, selecting a proper output modality.

On the other hand – and this is the second aspect –, a companion-like technical system (cf. [Wendemuth & Biundo 2012]) should be able to influence its user. This means, guiding the intention and involvement of the user towards a certain task and issue, for instance. It is not that the system patronises the user but keep him tracked on the interaction, usually to complete a task. For this, it is important to find a well-defined balance between guiding system's reactions and annoying system outputs. In general, such part of HMI switches the system's part from a passive analyser and task solver to a pro-active partner in an interaction, including the opportunity to affect the user as well as keep him involved and on track in an interaction and conversation. This issue is definitely an interdisciplinary research task incorporating computer scientists, engineers, and psychologists which will be in the focus in the future. Open issues regarding this aspect are presented in Section 7.2.5.

6.4 Summary

In this Chapter I discussed the issue of involvement in both senses: i) annotation of it (cf. Section 6.2) and ii) detecting changes of involvement in group conversations (cf. Section 6.3). As I pointed out, involvement is complementary to disposition which has also a meaning in a meta-analysis of given material and thus, given interactions. From my point of view, meta means that this is not an analysis which influences an utterance but reflects the course of a conversation or an interaction.

The group involvement was studied on the TableTalk corpus (cf. [Campbell 2009]) that supplies an HHI of four participants. The annotation process as such was presented and the reliability of the labelling was discussed whereas the focus was on the group's annotation (cf. Section 6.2.1). Comparing the values of Krippendorff's α to another data set, namely VAM, it can be seen that these are in a similar range. From this, even if the numbers themselves are low, the

annotation can be regarded as reliable.
Based on this annotation, three classes reflecting the changes in involvement were defined and afterwards, applied to detect and classify changes of group involvement. For this, video based features were utilised so far to train an MLP. In this context, I also discussed the disadvantages of centrally recorded audio data and hence, audio features for the purpose of involvement detection, especially on TableTalk.

Finally, the aspect of systems reactions on user involvement were elaborated. From both issues, namely detection of involvement and proper reaction on it, I derive a roadmap for further research which will be presented in the outlook, in particular, in Section 7.2.4.

Chapter 7

Conclusion and Future Work

Contents

7.1 Conclusion 155

7.2 Open Issues for Future Work 163

7.2.1 Advanced Semi-automatic Annotation 163

7.2.2 Naïve, Multiparty Disposition Recognition 165

7.2.3 Fusion Methods in Naïve Disposition Recognition 167

7.2.4 Involvement in Conversations 168

7.2.5 Disposition Recognition for Interaction Control 170

IN the preceding Chapters of this thesis I presented the utilised methods and results which were achieved in disposition recognition from speech. In this Chapter overall conclusions are given in Section 7.1 and finally, open issues are presented. Especially, in Section 7.2.4 I discuss a kind of roadmap for the analyses of involvement in conversations.

7.1 Conclusion

This thesis regarded the aspect of disposition recognition from speech. Therefore, the methods which are known from speech recognition are at first investigated for the purpose of emotion recognition from speech and afterwards, transferred to the issue of disposition recognition.

In the context of disposition recognition, the community is faced with the problem that so far no general valid definitions or interpretations of *dispositions* are available, especially in the sense of technical perception. In Section 1.2.2, in particular, in Definition 1.6 on page 7 I presented a definition focusing on a technical interpretation. In fact, this is influenced by a psychological point of view but goes beyond that because the suitability to technically detect and classify dispositions is incorporated. In addition, in Section 1.2.1 the terms *emotion* and *mood* are defined as well. Moreover, the connections between the three terms, *emotion*, *mood*, and *disposition*, are discussed. From my point of view, disposition is the most general one. As it covers besides aspects of emotions and moods also the issue of the situation and intention of a user. My definitions of the terms are technically inspired and I am aware that these are differently understood in other disciplines like psychology, linguistics, etc. However, I argued to regard dispositions in HMI to achieve a global analysis of a system's user, especially, if the system is considered to be companion-like (cf. [Wendemuth & Biundo 2012]).

Beside the definitions for the terms *emotion*, *mood*, and *disposition*, in Chapter 1 the main idea of the disposition recognition from speech was introduced. For this, I presented the labelling methods and the derived systems (cf. Section 1.2.3) which are usually applied in emotion and disposition recognition and further, are based on psychological foundations. In my terms they are orded as follows: i) categorical labelling systems like Basic Emotions according to Ekman (cf. [Ekman 1992]) or Plutchik (cf. [Plutchik 2001]), ii) quasi-continuous labelling as proposed in GEW (cf. [Scherer 2005]), and iii) continuous labelling systems, for instance, SAM (cf. [Bradley & Lang 1994]) which is based on the PAD space by Mehrabian. Furthermore, in Section 1.3 the inter- and intraindividual validation approach for the validation of classification are introduced and also put into the context of disposition recognition. Both methods are compared in Section 1.3.4 and the usage in companion-like technical systems is discussed. For this, I proposed the following workflow: In the beginning of an HMI the system is usually not adapted towards classifying the disposition of a certain user properly. Therefore, a general validation approach has to be chosen (cf. interindividual validation) and simultaneously material from the current speaker has to be collected to adapt the system. After the adaptation of the recogniser an intraindividual validation can be performed. Usually, this leads to an improvement of the recognition performance (cf. Table 5.4 on page 114). This issue affects also the way how a system is influenced by the conversation, in particular, the course of the interaction. However, it is connected to two aspect of controlling an HMI. On the one hand, the user controls and influences the conversation by his actions. Moreover, the interaction is influenced by the modalities utilised by the user, for example, gestures, touch actions, or speech, are affecting the interaction (cf. Section 1.5 and Section 4.4). Further, the system can influence its user and thus, the interaction in total (cf. Section 6.3 and Section 7.2.5).

In Chapter 2 the current state of the art in emotion and disposition recognition from speech as well as in the detection of changes in involvement is presented. Therefore, I reviewed the results and approaches which are used in emotion recognition from speech (cf. Section 2.2) and further, presented the ongoing research in this field. In Section 2.2 the achievements and methods for disposition recognition from speech are listed. The obtained results in both kinds of recognition are highly related to the classification methods applied on certain data sets. An overview of classification approaches is given and further, the classifiers used in my experiments are filed accordingly. Finally, in Section 2.5 the state of the art in automatic analysis of involvement in a conversation is considered. So far, such kind of research is not as widespread in a technical sense. Thus, it indicated the necessity to foster the work in this direction.

As already stated, the classification results depend on the material which is used to train and validate the recogniser. Hence, the corpora that are used in the experiments have to be described. This is done in Chapter 3 distinguishing two types of data sets, namely acted and non-acted ones. Furthermore, a data set supplying an HHI, namely TableTalk (cf. Section 3.3) was introduced which was applied in the analyses regarding involvement in conversations.
In the acted material, usually actors are asked to produce the expected emotion or disposition (cf. Section 3.1.1). For some corpora, for instance, eNTERFACE (cf. Section 3.1.2), the experimental setup supplies emotion induction to provoke or strengthen the emotional reaction. Therefore, also informed non-actors are suitable to generate an appropriate data set.
In contrast, non-acted corpora rely on subjects who are mostly not informed about the main concept of the experiment; namely, recording emotional or dispositional actions. Both kinds of those data sets introduced in Section 3.2 were recorded as WoZ scenarios in the SFB/TRR 62 either at the Otto von Guericke University Magdeburg or at the Ulm University. They represent data sets which are naturalistic HMIs since the participants are no actors and further, not informed about the underlying concepts of the experiments. Moreover, the interactions reflect tasks that occur in daily life, namely gaming (cf. EmoRec in Section 3.2.2) and planning (cf. Last Minute in Section 3.2.1). In both corpora more abstract events that can be seen as dispositions are supplied, for instance, the barriers *baseline*, *challenge*, *listing*, and *waiuku* in Last Minute (cf. [Rösner et al. 2012]) and, for example, positive valence, low arousal, and high dominance in EmoRec. Besides the induced dispositions, it can be assured that the process of evoking is working. Especially, for EmoRec the participants gave a self-rating

in terms of SAM. As it can be seen from Table 3.4 on page 56 the values reflect the expected disposition. However, such analysis is depending on the data set and the participant. Hence, it has to be evaluated for each recording separately since no generally valid method as for emotion recognition is established for disposition recognition, yet.

With the corpora described in Chapter 3 experiments were done to detect and classify dispositions from speech. Therefore, in Section 4.1 the preprocessing of a corpus is described; that is, *transcribing*, *annotating*, and *labelling*. Preprocessing also includes aspects of signal handling, for example, which is not considered in the context of this thesis. To focus the preprocessing – that means, to handle the process in total by just one tool – ikannotate (cf. Section 4.1.2) was developed. Ingo Siegert and I realised the three preprocessing steps based on utterance-level. In the tool, the utterances can be transcribed and afterwards, annotated according to GAT (cf. [Selting et al. 2011]). One important advantage of the tool is that the annotator does not have to care about the symbols of annotation because those are inserted by the system automatically while selecting the corresponding characteristics. This enables also non-trained annotators to generate a properly annotated document. Further, for the purpose of labelling the common schemes, namely Basic Emotions (similar to [Ekman 1992]), GEW (cf. [Scherer 2005]), and SAM (cf. [Bradley & Lang 1994]) are implemented. Hence, it is possible to process a corpus according to these schemes with one tool and therefore, comparisons are quite easy. The tool was presented at the conferences ICME 2011 (cf. [Böck et al. 2011a]) and ACII 2011 (cf. [Böck et al. 2011b]). Furthermore, it was used for several other experiments which were published, for instance, in [Siegert et al. 2012d; Siegert et al. 2012c].
Because the process of annotation and labelling is quite time consuming as discussed in Section 4.1.3, I introduced a framework which allows a semi-automatic annotation. So far, it is used to ease the annotation of facial expressions. In this framework, audio analyses based on spectral and prosodic features identify relevant sequences that afterwards, have to be rated by a FACS coder. Significant classification results (cf. Table 5.6 on page 123) were achieved which reduce the manual effort.

Based on the processed data sets and utilising the features introduced in Section 4.2, I investigated classifiers that are known from speech recognition. For my work, I mainly concentrated on HMMs (cf. Section 4.3.1) and GMMs. As discussed, HMMs can model and thus, handle a temporal evolution in a speech utterance. In contrast, GMMs are applicable to cope with the characteristics of a

sentence; that means in my context, with disposition. Furthermore, in short utterances as occurring in EmoRec almost no evolution of a disposition is observed. Thus, in most of the experiments I applied GMMs as classifiers.
On EmoDB I derived a parameter set for HMMs/GMMs (cf. [Böck et al. 2010]) that was also tested on eNTERFACE, which is a data set in the borderland of acted and non-acted material. The parameter setting which was further used in the experiments reported in Chapter 5 is as follows: For the number of hidden states in an HMM three states achieved the best performance with roughly 43% accuracy on eNTERFACE (cf. Figure 4.8 on page 85) whereas for GMMs by design one hidden state is used. As can be derived from Figure 4.9 on page 87 five iterations are sufficient to train an HMM as well as a GMM. More iterations result in a kind of overfitting, that means, the classifier is highly adapted to the training material and hence, the generalisation performance is decreasing. This is also valid for all kinds of feature sets. MFCC_D_A_0 which is MFCC with Delta, Acceleration, and 0^{th} cepstral coefficient achieved the best performance on eNTERFACE with 44.8% accuracy (cf. Figure 4.10(b) on page 89). The achieved results were also reflected taking results of [Vogt & André 2005] into consideration. They obtained similar achievements on comparable data sets, but utilising more features (90-160 features). On the other hand, Vogt & André do not discuss how to transfer the features to naturalistic corpora and disposition recognition from speech. This is done in my work.

In terms of classifiers, additionally SRNs, more specific SMRNNs, are observed. This work was done in cooperation with Stefan Glüge. Compared with HMMs they achieved roughly the same performance. The advantage is that by design the networks learn temporal characteristics which results in a reduced feature set of 13 features. For HMMs 39 features are necessary to obtain the same performance (cf. Table 4.5 on page 96). Unfortunately, SMRNNs are time consuming in the training and thus, are outperformed by HMMs/GMMs in this sense (cf. Section 4.3.4).

In Section 4.4 I also introduced the main aspects of fusion architectures. Since this issue is not a main topic of my thesis I kept it short. Nevertheless, my results can be utilised in the architectures which benefit from fusion. Especially, in the sense of multimodality the aspects are important and thus, I contributed to corresponding experiments (cf. Section 5.3) and also respond to it in Section 7.2.3 focusing on the audio analysis part, only.

Relying on the features and classifiers which the corresponding parameter sets from Chapter 4, I investigated the two data sets, Last Minute and EmoRec under three setups: i) a setting which uses only audio features, ii) a bimodal setup, and iii) a multimodal setting. My experiments which are presented in this thesis are the first audio analyses conducted on these corpora.
Analysing disposition on Last Minute means that the four situations which occur after a barrier (cf. [Rösner et al. 2012] and Section 3.2.1) have to be considered. As the material supplies full sentences I decided to validate and compare both types of classifiers: i) HMMs and ii) GMMs. The experimental setup was kept fixed for both classifiers and therefore, comparable results could be achieved. Though full sentences were observed the GMMs outperformed the HMMs. With HMMs 32.0% WA (cf. Table 5.1 on page 105) were obtained where several parameter sets were tested even if the system would benefit from more than one Gaussian mixtures in each state of the HMM. In contrast, GMMs yielded 43.97% WA for a cross-validation (cf. Table 5.2 on page 107) and 43.96% WA in an interindividual validation (cf. Table 5.3 on page 108). Regarding various feature sets the differences in the performance are not significant.
As the EmoRec data set is similar to Last Minute and further, only elliptical utterances – utterance in command style as it is necessary for the 'Concentration' game – occur I decided to use just GMMs for further investigations. It was also investigated whether a combination of HMMs and higher numbers of Gaussian mixtures per HMM' state boost the performance which was not the case. In particular, with the EmoRec corpus inter- and intraindividual validation experiments are possible because the corpus is designed to be suitable for those investigations, especially, in the sense of biophysiological analyses. In Table 5.4 on page 114 the recognition results for a subset of participants are listed. Utilising the complete set of participants is still under investigation since not all participants' material has been preprocessed, yet. The results on EmoRec were achieved while classifying the two ESs, namely ES-2 and ES-5 which are assumed to be representative for positive and negative dispositions, respectively. Comparing the achievements, it can be seen that for interindividual validation 55.1% WA, and for intraindividual validation 70.0% WA were obtained. Given these results I derived the aforementioned validation approach that a system starts with an interindividual validation and adapts itself towards a user. After this, the validation method is switched to an intraindividual analysis (cf. Section 5.1.3).

The classification results which were gained in the audio-only setup were also considered in a bimodal context, incorporating facial expressions and biophysiolo-

gical features.
Comparing recognition results for audio and biophysiological features (cf. Table 5.4 on page 114, Table 5.7 on page 125, and [Böck et al. 2012b]) it can be seen that in an intraindividual validation, classifiers based on biophysiological features outperform those trained on audio features. In contrast, for interindividual validation it is the other way around. Hence, both kinds of features and thus, classifiers can support each other. Especially, in the idea of switching validation methods, biophysiologically trained classifiers provide a kind of ground truth which is necessary to adapt audio-based classifiers towards a user (cf. Section 5.2.4). In the meantime interindividual audio-based recognisers are applied in a system, in particular, in a system with companion-like characteristics (cf. [Wendemuth & Biundo 2012]).
The system setup can also be used in semi-automatic annotation. Here, I concentrated on facial expressions as features. A pure comparison of both audio and facial modalities has already been published in [Böck et al. 2012b]. As described in Section 4.1.3 audio-based classifiers, specifically GMMs, that apply spectral and prosodic features, identify relevant affective sequences which afterwards are annotated manually by FACS coders. With this approach the manual effort in annotation of facial expressions is reduced significantly. As stated in [Böck et al. 2013a], to annotate the two sequences ES-2 and ES-5 for 20 participants, that means, $20 \cdot 7\text{min} = 140\text{min}$ of video material, roughly 13-19 hours are necessary. Moreover, the 140 minutes contain only a few minutes of relevant facial expressions. If this relevant material can be preselected the effort can be reduced drastically. Of course, therefore, the classification process has to be robust in the sense of marking relevant sequences. With GMMs I achieved recognition results of 61.9% for *FACS in ES* and 81.2% for *no-FACS* sequences – the two classes that mark relevant information – which corresponds to a false acceptance rate of 18.8% and a false rejection rate of 38.1%. Especially, the results for *FACS in ES* sequences have to be improved which is the issue of further research (cf. Section 7.2.1). Furthermore, the idea of semi-automatic annotation could be also transferred to other modalities like biophysiological measures or gestures, which culminates in a multimodal support of annotation.

In Chapter 5, I also presented first results of disposition recognition on the EmoRec II corpus. Unfortunately, so far, the material of only eight participants is prepared to be used for analyses. The whole experiment was included in the multimodal analyses which were conducted in cooperation with the colleagues at the Ulm University. Since this thesis is not dealing mainly with aspects of fusion I concentrated on the audio results. Nevertheless, I compared those to the results achieved by a system applying fusion techniques (cf. [Schels et al. 2012]). With GMMs an recognition performance of 52.9% WA based on an interindividual validation (cf. Table 5.9 on page 129) was achieved where the fused system yielded 62.0% at its best. This is comparable to the values in EmoRec I (cf. Table 5.4 on page 114).

In particular, investigating the results on the EmoRec I+II data sets (cf. Table 5.4 on page 114 and Table 5.9 on page 129) it is noticeable that in several cases the WA results contain higher values than UA. For this, I deeply analysed the settings in terms of the samples' distribution. For these cases, I found that the samples' distribution is unbalanced in the frequency of occurrence of the disposition classes. Usually, it is expected that the classifier tends towards the class which occurs most frequently in the material to optimise its performance where the UA values are then too optimistic. Therefore, WA was invented to rank the results properly. In the particular cases, it is the other way around and the class having the fewest samples is recognised almost perfectly, resulting in a too optimistic WA $>$ UA. I phrased this phenomenon 'counterbalanced'. Therefore, both UA and WA measures have been given in this thesis whenever counterbalanced effects occur.

Finally, a more abstract concept is introduced, namely the involvement in a conversation. From my point of view, which is supported by the definition of disposition (cf. Definition 1.6 on page 7), involvement is complimentary to disposition. Therefore, it is worthwhile to study involvement. The results are given in Chapter 6. Since this type of analysis is quite novel, at first a corpus which is suitable has to be identified. This was attained with the TableTalk data set (cf. Section 3.3 and [Campbell 2009]) where this kind of analysis was triggered by discussions with Nick Campbell which I had during my internship at the Trinity College Dublin.

TableTalk has, so far, not been annotated according to either involvement as such or changes of involvement. Hence, I conducted the annotation of changes in involvement and derived three classes which are of interest: + (increase of involvement), $-$ (decrease of involvement), *0* (no change), and additionally *?* which represents the undecided cases. The whole process is explained in Section 6.2 and [Böck et al. 2013b], relating the method also to other approaches in the community. As it is usual in annotation the reliability of the assignments is analysed. For this, Krippendorff's α is calculated. The ordinal Krippendorff's α on the data set, in particular *day1*, is $\alpha_o = 0.1562$. In contrast, the reliability on the D64 interaction corpus (cf. [Oertel et al. 2011a]) is higher which results from a more fixed and restricted annotation process. Furthermore, first results of automatic detection of changes in involvement are presented in Section 6.3.1. In comparison to classification of involvement on the D64 data set (cf. [Oertel et al. 2011a]) remarkable results were achieved. On the TableTalk corpus for the classification of changes in involvement a WA of 67.8% were obtained, considering the four class classification task discussed in 6.3. Lines for further improvements are discussed in (cf. Section 7.2.4).

From the observation of involvement in conversations, finally, I discussed the way a system can guide and influence its user (cf. Section 6.3.2). This aspect is quite important. Especially, if a technical system reacts companion-like (cf. [Wendemuth & Biundo 2012]), it can be assumed that it is rather treated as a partner in the conversation. By investigating this issue one gets aware of the open questions that will be discussed in the following Sections.

7.2 Open Issues for Future Work

In this thesis, I considered the automatic recognition of dispositions from speech regarding suitable features sets and classifiers in relation to non-acted, naturalistic corpora. Since the work on this topic is still not finished and this thesis is not able to cover all aspects, open issues are outstanding and have to be investigated in future research. In the current Section, I will sketch open issues that are building a bridge between the semi-automatic annotation and preprocessing of corpora, the disposition recognition under various aspects, and the influence of disposition recognition on interaction control.

7.2.1 Advanced Semi-automatic Annotation

Concerning the semi-automatic annotation of facial expressions, so far, only the EmoRec I data set was considered (cf. Section 5.2.2). Therefore, obviously the approach should be used for the preprocessing of EmoRec II to reduce the effort of annotation for this corpus, too. In this context, the classifiers which select the relevant sequences can be trained on more samples which leads to an expected improvement in the classification accuracy. Moreover, the framework should be tested on other corpora that supply also multimodal recordings. For this, the method should be integrated in an annotation tool like ikannotate to support the labellers in their work.

The aspect of the framework's dissemination affects the performance of the system itself. Up to now, the recognition performance is reasonable good, but the false rejection rate has to be decreased. On the one hand, it is assumed that this can be achieved by training with more data samples. Having more material provides the possibility to further tune the classifiers' parameters, and so usually, an improvement can be achieved. On the other hand, the more disjoint the classes are, the better is the classification and thus, the false rejection rate is decreasing. Hence, an investigation of the given classes according to their disjoint characteristics result in gain an improvement in the framework's performance as well. Though, this is a matter of psychologists since they introduced the categories and this is far beyond my expertise.

Generally, the idea of the annotation framework can be carried to other characteristics of the user. So far, I suggested to use this setting to find relevant video sequences for facial expression. This approach can also be applied to sequences containing dispositional gestures occurring in video recordings. In general, audio-based preselection of video sequences that contain relevant affective characteristics is possible with the framework. In future investigations this approach should be considered.
Moreover, it is to be checked whether this idea can support the labelling of biophysiological characteristics, too. Up to now, it is unclear how both modalities, namely acoustic utterances and biophysiological measures, are influencing each other. From my point of view, biophysiological reactions of the user will occur before acoustic ones. However, to fit the framework, these aspects have to be analysed, first.

In general, semi-automatic annotation of corpora, especially, with the aid of various modalities is an important issue in the future. By this way, the annotation of corpora which have to be generated to cover dispositional reactions of users in HMI can be handled and the manual effort can be reduced. This holds even more if naturalistic multiparty disposition recognition from speech is analysed as therefore, the given material has to be prepared for i) all participants and ii) all combinations of possible interactions; that is, dyadic and group interactions. This will result in a faster availability of novel labelled data sets because semi-automatic preprocessing is applied.

7.2.2 Naïve, Multiparty Disposition Recognition

Based on annotated material the recognition of dispositions is, so far, analysed in a sense of naturalistic ‘dyadic’ HMI. In connection with the term *naïve* not only a naturalistic characteristic of the data set is meant, further, it includes that the participant behaves naturally and is not concerned that the whole interaction might be a mock-up in any sense. The HMI becomes more and more a conversation between equal partners due to the system’s capabilities. Therefore, the participant can behave as a naïve user. Such HMIs are challenging tasks for classifiers which should assess the user’s dispositions. Hence, at first, corpora reflecting such naïve interactions are to be generated and afterwards, processed to be suitable for classification purposes. With data sets like Last Minute, EmoRec, or SAL (cf. [McKeown et al. 2012]) foundations are laid that have to be extended.

Furthermore, it is not exhaustively analysed, yet, if the dispositional ‘categories’ which are used today to describe the characteristics of a person are covering the whole set of dispositions. In close interdisciplinary cooperation with psychologists this aspect has to be evaluated and if necessary further dispositions should be defined. Especially in HMI, it is to be expected that certain dispositions do not occur or, the other way around, are only realised in such interactions. Such aspects are even more important if the context of an HMI is extended to a multiparty interaction.

There are two ways of multiparty interactions: i) multiple users interact with one system or ii) multiple technical systems interact with one user. The case where just systems are interacting is not in the scope of this thesis and is not considered as well as the most complex task where multiple users interact with multiple systems. The interesting issue is the interaction of multiple users with one system. For this, the system has to distinguish between reactions of the participants related to the system or to the group. Therefore, additional aspects like group dynamics, group structures, etc. are on hand which are briefly introduced in Section 7.2.4. Nevertheless, suitable data sets are needed which allow any kind of investigation regarding those multiparty interactions.

In a more technical sense, a system has to differentiate between various user reactions according to their meaning. In addition, it has to track the dispositions of several users, too. Both issues are complex by themselves, especially, under the circumstances of robustness and ubiquitous availability. Still, systems are highly influenced by the surrounding that means, the environment in total including noise, echo, lighting conditions, etc. For this, the classifiers and thus, the system in total have to be improved and thus, they are still under the process of optimisation.

Another important issue which is still under research is the question whether dispositions are independent of a user. The other way around, it has to be analysed what are the dependencies of a disposition. From a technical point of view, engineers can contribute to this discussion, but it is mainly the psychologists' task to come up with ideas which afterwards can be validated by applying technical approaches and methods.

From these general considerations I will, finally, introduce some open issues which are more technically inspired.
At first, the data recording needs to be tuned to handle multiple users, in particular, for audio processing. As known from approaches like beamforming and blind source separation, methods to locate and distinguish several sources, that means, different users, are available. So far, they are usually not used in HMI due to the handling of such equipment – it is typically not handy due to the microphone array and the necessary distances between microphones. With Kinect™a system is on hand that provides a microphone array which is adequate for limited multiparty scenarios. Unfortunately, even this system is still improper for distributed systems or companion-like technologies as intended by [Wendemuth & Biundo 2012]. Therefore, the technical equipment is an issue for multiparty recognitions – it does not matter if this is a disposition recognition –, especially, in the sense of mobile, distributed systems.
Furthermore, up to now, single user HMIs are considered. For these, feature sets have been adapted and optimised. On the other hand, it is an open issues whether the features are suitable for a multiparty interaction. In the analyses, for instance, the number of participants have to be considered which is usually not done, yet. Thus, non-affected overlapping or cross-talk of several users might have the same appearance as affected speech uttered by just one user. Thus, I encourage the research community to look at upcoming multiparty corpora and derive or develop proper features that with cope dispositional characteristics. Of course, this does not solve the problem of regarding several participants but multiparty corpora at least provide the option to conduct suitable analyses. Further, they have to be suitable for recognisers which handle and incorporate fusion methods.

7.2.3 Fusion Methods in Naïve Disposition Recognition

Related to aspects discussed in Section 7.2.2 the recognition of dispositions is an issue of multimodality. A general view on the user's behaviour can be only achieved if several modalities are considered simultaneously. Related to naïve disposition recognition first results could be presented on, for instance, the EmoRec data set. Unfortunately, actually a multimodal observation of a user in total cannot be generally supplied. This is due to the analysis of multimodal settings for which the time alignment of all modalities is necessary. Usually, this can only be guaranteed during the recording using a universal clock which assures synchrony in the collection process. The problems are more due to the alignment of results; that means, for different modalities the user's reaction can be detected on different time scales only. For example, biophysiological signals visualise reactions on longer periods while audio classifications reacts on utterance-level; that is, in milliseconds. Another problem is manifested since the technical standard of recording in many cases does not support real-time ability. For example, data may be recorded at a rate as it is piped into a communication channel which frequently results in small data losses which cumulate to large latencies of up to tenths of a second. Further, internal nonlinear data compression is often applied in recording devices which is not retractable or invertible from the output data, for instance, if stored in mpeg file format. Hence, specialised real-time recording equipment and corresponding data formats have to be employed.

In Section 4.4 approaches for fusion techniques are presented. Nevertheless, the proper handling of multiple modalities is still a matter of research, especially, to incorporate the time alignment of different classifier results. From my point of view, hybrid architectures that combine classifier results and additional features extracted from the data are worthwhile to look at.

In the future, for example, the combination of HMMs and SMRNNs should be investigated, especially, regarding how both architectures can support each other in terms of temporal relations. Both classifiers are able to handle temporal relations, but on different levels. HMMs represent a straightforward time evolution whereas SMRNNs by design distinguish different temporal layers, namely symbol-level and segment-level. Coupling both methods could improve the straightforward handling of context.

The considered temporal alignment is already a complex issue in a naturalistic HMI. It will be even more difficult in a naïve one. Furthermore, the alignment depends also on the disposition itself. The more complex and thus, longer lasting it is, the more the extracted features or classifier results are apart from each other in a temporal sense. For this problem appropriate fusion techniques have to be developed which can be utilised in detection of changes in involvement as well, since this is complementary to disposition as I already explained.

7.2.4 Involvement in Conversations

As I previously discussed, involvement can be seen as complementary to disposition und thus, can be analysed by methods analogous to disposition recognition from speech. So far, the analysis of changes in involvement as well as the detection of involvement as such are quite novel issues in the context of speech recognition and especially, disposition recognition. Therefore, I briefly sketch a kind of roadmap for future research on this issue.

In Section 6.2 I presented the annotation process for the *day1* subset of the TableTalk corpus related to changes in involvement. Based on this procedure other data sets, which provide options to detect involvement, can be processed and thus, a reliable labelling can be achieved. Of course, for each annotation task the reliability of the labelling has to be investigated separately. Three classes as introduced in Section 6.2, namely *increase of involvement*, *decrease of involvement*, and *no change in involvement* reflect the characteristic of involvement quite well. However, a more fine scaled granularity in the annotation can be established. Though, enough material in terms of training and testing is necessary and thus, it highly depends on the analysed data set. The preprocessing of the data is quite important. I encourage the researchers to do the annotation for both, the group in total as well as each participant, providing possibilities for various comprehensive analyses.

I suggest that at first the involvement of the group is analysed, if possible. This founds the basics for the whole conversation. In musical term this would be the beat of the conversation. The participants are embedded into this beat and add the rhythm and notes to it. For this, each event can be related to the overall structure of the conversation. Furthermore, it is necessary to derive the organisational structure of the group. In psychology several concepts can be found which reflect the constellation within a group. From these considerations three basics steps in the analysis can be derived: i) the group's constellation, ii) the involvement of the group in total, and iii) the role of each participant including his involvement.

As I pointed out in Section 6.3, the detection of changes in involvement is a complex task. From my point of view, this issue can be handled only if the analyses are based on multiple modalities and further, also on specialised classifiers. So far, only limited technical studies are available whereas in psychology the issue of involvement and its change is already being discussed for a longer period. Hence, from a technological point of view, suitable classifiers have to be found. It is an open issue which architectures are feasible at all. In medium time range, a decision level fusion approach (cf. Section 4.4 and Figure 4.12 on page 98) might be worthwhile to investigate.
On the other hand, even in terms of features the community lacks of suitable sets. In video analysis promising features have already been found (discussed in Section 6.2 and further cf. [Wrede & Shriberg 2003; Yu et al. 2004; Campbell & Douxchamps 2007]), but for audio processing the investigation is still at the beginning in terms of detecting changes in involvement. Especially, during cross-talk of several participants the common spectral and prosodic features seem to be inapplicable. For this, other features or feature sets – inspired by the handling of music – have to be developed that cope with such complex interactions.

In general, assuming that the technical issues for involvement detection are handled, the question arises how to exploit this information. For technical systems at least two 'applications' are possible.
Analysing an HHI, the technical system could provide support in handling the equipment. That means, for instance, which microphone has to be switched on or which camera should focus on a certain participant. Such ideas are in parts already realised in conference tools but usually, the automatic handling is quite basic. Most of the time, an operator or even the conference's participant is controlling the equipment.
On the other hand, analysing HMIs is the more interesting issue. For this, the system interacts with a group and thus, has to analyse the whole group as well as each participant. Several aspects have to be highlighted: Who is, in particular, interacting with the system? Is the interaction focused on the members of the group or towards the system? Who is getting to start an interaction with the system or is joining into an existing interaction? Who is no longer involved? Besides these questions it is also necessary to develop strategies how to react to the information supplied by detecting changes in involvement. This leads directly to aspects of interaction control which will be considered in the following Section.

7.2.5 Disposition Recognition for Interaction Control

In the previous Sections I discussed open issues that are related to the classification process of dispositions or are dealing with the system's reaction to the user's disposition. Now, I consider the interaction the other way around; that means, how the system could use the recognised dispositions to control the interaction. Of course, this can be just a brief sketch of ideas.

From the detected changes in involvement the system can conclude whether the user is still active in an interaction. Otherwise, the further system's reactions have to be aimed at keeping the user in the interaction or at getting him back to it. For this, the current situation has to be analysed and proper actions have to be provided. For example, an explanation of the current task should be given to the user, the output modality may have to be changed, or even the general interaction strategy may have to be evaluated. Moreover, the same aspects are valid for any reaction on another disposition like being in a positive/negative attitude, being stressed, etc. On the other hand, if the user could not be motivated and left the interaction, the course should be analysed and rules for a proper reaction in the future need to be derived.

Besides a robust recognition of disposition and involvement, these considerations require also a model of the user including interaction patterns and background knowledge and a world model containing information about the current situation, plans, and goals. In the sense of a technical realisation, various disciplines like artificial intelligence, interface design, computer science, and engineering science have to work together to generate systems that are enabled with the corresponding characteristics. On the other hand, psychologists have to analyse the human strategies of communication with technical cognitive systems and furthermore, the human's way of presenting dispositions towards such a system. All these information culminate in a system which is adapted to a certain user and reflects as well as knows his needs. Therefore, it can support its counterpart and further, influence the user's situated interactions.

I am confident that in the future almost everybody has its own companion-like system that supports its user in daily life and is recognised as a kind of 'companion'.

Glossary

Basic Emotions
: Set of emotional categories which is seen as universal; that means, expressions are shown by humans and animals.

cross-validation
: The recorded material of all speakers, contained in a corpus, is split into two sets, namely a training and test set. The material of the test set is not used in the training, but in testing only. This method is related to intraindividual validation.

Human-Human Interaction
: Human-Human Interaction describes the communication and interaction of multiple human beings with each other.

Human-Machine Interaction
: Human-Machine Interaction describes the communication and interaction of single or multiple human users with any kind of technical system.

interindividual validation
: For the training of a classifier the material in total is used except for one speaker or a speaker group. This remaining data is applied in the testing only.

intraindividual validation
: The recorded material of one speaker is split into two sets, namely a training and test set. The material of the test set is not used in the training, but in testing only.

Leave-One-Speaker-Group-Out
: For the training of a classifier the material of all speakers grouped by a certain characteristic is used except for one particular group. The remaining data of this group is applied in the testing only.

Leave-One-Speaker-Out
: For the training of a classifier the material of all speakers contained in a data set is used except for one particular speaker. The remaining data of this speaker is applied in the testing only.

Machine-Machine Interaction

Machine-Machine Interaction considers the communication and interaction of multiple technical systems whereas the type of the system and the way of interaction does not matter for this thesis.

phoneme-level

Any kind of recognition or processing which is based on phonemes as the unit of interest.

Unweighted Average accuracy

The accuracy is calculated based on all samples whereas this is the ratio of correctly classified samples over all samples in the data set. Hence, the distribution of the samples is not considered.

utterance-level

Any kind of recognition or processing which is based on utterances as the unit of interest.

Weighted Average accuracy

This measure reflects the class-wise accuracy. Therefore, the accuracy is calculated for each class separately and afterwards averaged over all classes.

Wizard-of-Oz scenario

Such a scenario simulates a system in a way that the system's functionality is substituted by an operator. The participants of the scenario are not informed about the simulation.

word-level

Any kind of recognition or processing which is based on words as the unit of interest.

Abbreviations

ANN	Artificial Neural Network
AU	Action Unit
AVEC	Audio/Visual Emotion Challenge
BoW	Bag-of-Words
DES	Danish Emotional Speech database
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DST	Dempster-Schafer Theory
eBTT	extended Backpropagation Through Time
EEG	electroencephalogram
EM	Expectation-Maximisation
EmoDB	Berlin Emotional Speech Database
EmoRec	EmoRec I+II
eNTERFACE	eNTERFACE'05
eRTRL	extended Real-Time Recurrent Learning
ES	Experimental Sequence
ESN	Echo State Network
FACS	Facial Action Coding System
FFT	Fast Fourier Transform
GAT	Gesprächsanalytisches Transkriptionssystem (dialogue analytic transcription system)
GEW	Geneva Emotion Wheel
GMM	Gaussian Mixture Model
HHI	Human-Human Interaction
HMI	Human-Machine Interaction
HMM	Hidden Markov Model
HTK	Hidden Markov Toolkit

IDFT	Inverse Discrete Fourier Transform
ikannotate	interdisciplinary knowledge-based annotation tool for aided transcription of emotions
LLD	Low-Level Descriptors
LOSGO	Leave-One-Speaker-Group-Out
LOSO	Leave-One-Speaker-Out
LPC	Linear Predictive Coding
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
MMI	Machine-Machine Interaction
PAD	Pleasure-Arousal-Dominance
PC	Personal Computer
PLP	Perceptual Linear Predictive Coefficients
SAL	Sensitive Artificial Listener
SAM	Self-Assessment Manikin
SFB/TRR 62	Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems”
SmartKom	SmartKom Database
SMRNN	Segmented-Memory Recurrent Neural Network
SRN	Simple Recurrent Network
SVM	Support Vector Machine
UA	Unweighted Average accuracy
VAM	Vera am Mittag
VTN	Vocal Tract Normalisation
WA	Weighted Average accuracy
WoZ	Wizard-of-Oz

Symbols

0^{th}	Zeroth cepstral coefficient
A	Set of state transition probabilities
a_{ij}	State transition probability
α	Krippendorff's α
B	Set of production probabilities
$b_i(o_i)$	Production probability
$\mathbf{E}$	Identity matrix
E	Energy term
F0	Fundamental frequency
K	Output alphabet of a Hidden Markov Model
κ_{g}	General reliability
O	Observation sequence
o_i	Observation of a state in a Hidden Markov Model
S	State sequence
s_i	State of a Hidden Markov Model
$s(n)$	Speech signal
s_{XX}	Autocorrelation
s_{XY}	Cross-correlation
$T(n)$	Transfer function
$u(n)$	Excitement

References

References related to the Author

Böck, R.; K. Limbrecht-Ecklundt; I. Siegert; S. Walter & A. Wendemuth (2013a). 'Audio-based Pre-classification for Semi-automatic Facial Expression Coding'. In: *Human-Computer Interaction.* Ed. by Kurosu, M. Kurosu, M. Vol. 8008. Lecture Notes in Computer Science. Springer, pp. 301–309.

Böck, R.; D. Hübner & A. Wendemuth (2010). 'Determining Optimal Signal Features and Parameters for HMM-Based Emotion Classification'. In: *Proceedings of the 15th IEEE Mediterranean Electrotechnical Conference.* Valletta, Malta: IEEE, pp. 1586–1590.

Böck, R.; I. Siegert; B. Vlasenko; A. Wendemuth; M. Haase & J. Lange (2011a). 'A Processing Tool for Emotionally Coloured Speech.' In: *Proceedings of the 2011 IEEE International Conference on Multimedia & Expo*. Barcelona, Spain: IEEE, s.p.

Böck, R.; I. Siegert; M. Haase; J. Lange & A. Wendemuth (2011b). 'ikannotate - A Tool for Labelling, Transcription, and Annotation of Emotionally Coloured Speech.' In: *Affective Computing and Intelligent Interaction.* Ed. by D'Mello, S. D'Mello, S.; Graesser, A.; Graesser, A.; Schuller, B.; Schuller, B. & Martin, J.-C. & Martin, J.-C. Vol. 6974. Lecture Notes in Computer Science. Springer, pp. 25–34.

Böck, R.; K. Limbrecht; I. Siegert; S. Glüge; S. Walter & A. Wendemuth (2012a). 'Combining Mimic and Prosodic Analyses for User Disposition Classification'. In: *Proceedings of the 23. Konferenz Elektronische Sprachsignalverarbeitung.* Cottbus, Germany: TUD Press, pp. 220–227.

Böck, R.; K. Limbrecht; S. Walter; D. Hrabal; H. C. Traue; S. Glüge & A. Wendemuth (2012b). 'Intraindividual and Interindividual Multimodal Emotion Analyses in Human-Machine-Interaction'. In: *2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support.* New Orleans, USA: IEEE, pp. 59–64.

Böck, R.; S. Glüge; I. Siegert & A. Wendemuth (2013b). 'Annotation and Classification of Changes of Involvement in Group Conversation'. In: *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction.* Geneva, Switzerland: IEEE, pp. 803–808.

Bonin, F.; R. Böck & N. Campbell (2012). 'How Do We React to Context? Annotation of Individual and Group Engagement in a Video Corpus'. In: *2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust.* Amsterdam, The Netherlands: Odysci, pp. 899–903.

Glüge, S.; R. Böck & A. Wendemuth (2011). 'Segmented-Memory Recurrent Neural Networks versus Hidden Markov Models in Emotion Recognition from Speech'. In: *Proceedings of the 3rd International Joint Conference on Computational Intelligence.* Paris, France: INSTICC, pp. 308–315.

— (2012). 'Extension of Backpropagation Through Time for Segmented-Memory Recurrent Neural Networks'. In: *Proceedings of the 4th International Joint Conference on Computational Intelligence.* Barcelona, Spain: INSTICC, pp. 451–456.

— (2013). 'Auto-Encoder Pre-Training of Segmented-Memory Recurrent Neural Networks'. In: *Proceedings of the European Symposium on Artificial Neural Networks.* Bruges, Belgium, pp. 29–34.

Schels, M.; M. Glodek; S. Meudt; M. Schmidt; D. Hrabal; R. Böck; S. Walter & F. Schwenker (2012). 'Advances in Affective and Pleasurable Design'. In: ed. by Ji, Y. G. Ji, Y. G. Advances in Human Factors and Ergonomics Series. CRC Press. Chap. Multi-modal classifier-fusion for the classification of emotional states in WOZ scenarios, pp. 644–653.

Siegert, I.; R. Böck & A. Wendemuth (2012a). 'Modeling users' mood state to improve human-machine-interaction'. In: *Cognitive Behavioural Systems. COST 2102.* Ed. by Esposito, A. Esposito, A.; Esposito, A. M.; Esposito, A. M.; Vinciarelli, A.; Vinciarelli, A.; Hoffmann, R.; Hoffmann, R. & Müller, V. C. & Müller, V. C. Vol. 7403 LNCS. Dresden, Germany: Springer, pp. 273–279.

Siegert, I.; R. Böck & A. Wendemuth (2012c). 'The Influence of Context Knowledge for Multimodal Annotation on natural Material'. In: *Joint Proceedings of the IVA 2012 Workshops*. Santa Cruz, USA: Otto von Guericke University Magdeburg, pp. 25–32.

Siegert, I.; R. Böck; D. Philippou-Hübner; B. Vlasenko & A. Wendemuth (2011). 'Appropriate Emotional Labelling of Non-Aacted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins'. In: *Proceedings of the 2011 IEEE International Conference on Multimedia & Expo*. Barcelona, Spain: IEEE, s.p.

Siegert, I.; R. Böck; D. Philippou-Hübner & A. Wendemuth (2012d). 'Investigation of Hierarchical Classification for Simultaneous Gender and Age Recognitions'. In: *Proceedings of the 23. Konferenz Elektronische Sprachsignalverarbeitung*. Cottbus, Germany: TUD Press, pp. 58–64.

Vlasenko, B.; D. Philippou-Hübner; D. Prylipko; R. Böck; I. Siegert & A. Wendemuth (2011b). 'Vowels Formants Analysis Allows Straightforward Detection of high Arousal Emotions'. In: *Proceedings of the 2011 IEEE International Conference on Multimedia & Expo*. Barcelona, Spain: IEEE, s.p.

Vlasenko, B.; D. Prylipko; R. Böck & A. Wendemuth (2012). 'Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications'. In: *Computer Speech & Language*. in press.

Walter, S.; S. Scherer; M. Schels; M. Glodek; D. Hrabal; M. Schmidt; R. Böck; K. Limbrecht; H. Traue & F. Schwenker (2011). 'Multimodal Emotion Classification in Naturalistic User Behavior'. In: *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*. Ed. by Jacko, J. Jacko, J. Vol. 6763. Lecture Notes in Computer Science. Springer, pp. 603–611.

References

Albornoz, E. M.; D. H. Milone & H. L. Rufiner (2010). ‘Multiple feature extraction and hierarchical classifiers for emotions recognition’. In: *Development of Multimodal Interfaces: Active Listening and Synchrony*. Ed. by Esposito, A. Esposito, A.; Campbell, N.; Campbell, N.; Vogel, C.; Vogel, C.; Hussain, A.; Hussain, A. & Nijholt, A. & Nijholt, A. Vol. 5967 LNCS. Lecture Notes in Computer Science. Springer, pp. 242–254.

Anagnostopoulos, C.-N.; T. Iliou & I. Giannoukos (2012). ‘Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011’. In: *Artificial Intelligence Review*. without issue assignment, pp. 1–23.

Antil, J. (1984). ‘Conceptualization and Operationalization of Involvement’. In: *Advances in Consumer Research* 11.1, pp. 203–209.

Anusuya, M. & S. Katti (2009). ‘Speech Recognition by Machine: A Review’. In: *International Journal of Computer Science and Information Security* 6.3, pp. 181–205.

Appenrodt, J.; A. Al-Hamadi & B. Michaelis (2010). ‘Data Gathering for Gesture Recognition Systems Based on Single Color-, Stereo Color- and Thermal Cameras’. In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 3.1, pp. 37–50.

Artstein, R. & M. Poesio (2008). ‘Inter-Coder Agreement for Computational Linguistics’. In: *Computational Linguistics* 34.4, pp. 555–596.

Banister, E. & D. Begoray (2004). ‘BEYOND TALKING GROUPS: STRATEGIES FOR IMPROVING ADOLESCENT HEALTH EDUCATION’. In: *Health Care for Women International* 25.5, pp. 481–488.

Batliner, A.; K. Fischer; R. Huber; J. Spiker & E. Nöth (2000). ‘Desperately Seeking Emotions: Actors, wizards and human beings’. In: *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*. Belfast, UK: Textflow, pp. 195–200.

Batliner, A.; S. Steidl & E. Nöth (2008). ‘Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus’.

In: *Proceedings of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect*. Marrakesh, Marocco: Elsevier, pp. 28–31.

Batliner, A.; E. Nöth; J. Buckow; R. Huber; V. Warnke & H. Niemann (2001). 'Whence and Whither Prosody in Automatic Speech Understanding: A Case Study'. In: *Proceedings of the Workshop on Prosody and Speech Recognition 2001*. Red Bank, USA: ISCA, pp. 3–12.

Bencherif, M. A.; M. Alsulaiman; G. Muhammad; Z. Ali; A. Mahmood & M. Faisal (2012). 'Gender Effect in Trait Recognition'. In: *Proceedings of the World Congress on Engineering and Computer Science*. San Francisco, USA: IAENG, pp. 676–679.

Bertelsmann (1993). Bertelsmann Universal Lexikon. 4 & 17. Bertelsmann.

Bezooijen, R. van (1984). Characteristics and recognizability of vocal expressions of emotion. Netherlands phonetic archives. Dordrecht, The Netherlands: Foris Publications.

Binder, J.; K. Murphy & S. Russell (1997). 'Space-efficient inference in dynamic probabilistic networks'. In: *Proceedings of the Fifteenth International Joint Conference on Artifical Intelligence*. Nagoya, Japan: Morgan Kaufmann Publishers Inc., pp. 1292–1296.

Boersma, P. (2001). 'Praat, a system for doing phonetics by computer.' In: *Glot International* 5.9-10, pp. 341–345.

Borasca, B.; L. Bruzzone; L. Carlin & M. Zusi (2006). 'A fuzzy-input fuzzy-output SVM technique for classification of hyperspectral remote sensing images'. In: *Proceedings of the 7th Nordic Signal Processing Symposium*. Reykjavik, Iceland: IEEE, pp. 2–5.

Bousmalis, K.; M. Mehu & M. Pantic (2013). 'Towards the automatic detection of spontaneous agreement and disagreement based on non-verbal behaviour: A Survey of related cues, databases, and tools'. In: *Image and Vision Computing* 31.2, pp. 203–221.

Bradley, M. M. & P. J. Lang (1994). 'Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential'. In: *Journal of Behavioral Therapy and Experimental Psychiatry* 25.1, pp. 49–59.

Bross, F. (2010). 'Grundzüge der Akustischen Phonetik'. In: *A Multidisciplinary Online Journal* 1.1, pp. 89–104.

Burg, J. (1975). Maximum entropy spectral analysis. Stanford Exploration project. Stanford University.

Burger, S.; V. MacLaren & H. Yu (2002). 'The ISL meeting corpus: the impact of meeting type on speech style'. In: *INTERSPEECH-2002*. s.p. Denver,USA: ISCA.

Burkhardt, F.; A. Paeschke; M. Rolfes; W. F. Sendlmeier & B. Weiss (2005). 'A Database of German Emotional Speech'. In: *INTERSPEECH-2005*, pp. 1517–1520.

Campbell, J. P. & D. A. Reynolds (1999). 'Corpora for the Evaluation of Speaker Recognition Systems'. In: Phoenix, USA: IEEE, pp. 829–832.

Campbell, N. (2009). 'An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data'. In: *INTERSPEECH-2009*. Brighton, England: ISCA, pp. 2159–2162.

Campbell, N. & D. Douxchamps (2007). 'Processing Image and Audio Information for Recognising Discourse Participation Status'. In: *INTERSPEECH-2007*. Antwerp, Belgium: ISCA, pp. 730–733.

Campbell, N.; T. Sadanobu; M. Imura; N. Iwahashi; S. Noriko & D. Douxchamps (2006). 'A Multimedia Database of Meetings and Informal Interactions for Tracking Participant Involvement and Discourse Flow'. In: *Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy: ELRA, pp. 391–394.

Carletta, J. et al. (2005). 'The AMI meeting corpus: a pre-announcement'. In: *Proceedings of the Second international conference on Machine Learning for Multimodal Interaction*. Edinburgh, UK: Springer, pp. 28–39.

Carroll, J. M. (2013). 'Human Computer Interaction - brief intro'. In: *The Encyclopedia of Human-Computer Interaction*. Ed. by Soegaard, M. Soegaard, M. & Dam, R. F. & Dam, R. F. 2nd ed. Aarhus, Denmark: The Interaction Design Foundation, s.p.

Cave, R. L. & L. P. Neuwirth (1980). 'Hidden Markov Models for English'. In: *Hidden Markov Models for Speech*. Ed. by Ferguson, J. Ferguson, J. IDA, Princeton, s.p.

Chen, J. & N. Chaudhari (2009). 'Segmented-memory recurrent neural networks'. In: *IEEE Transactions on Neural Networks* 20.8, pp. 1267–1280.

Chen, L.; R. Rose; Y. Qiao; I. Kimbara; F. Parrill; H. Welji; T. X. Han; J. Tu; Z. Huang; M. P. Harper; F. K. H. Quek; Y. Xiong; D. McNeill; R. Tuttle & T. S. Huang (2005). 'VACE Multimodal Meeting Corpus'. In: *Proceedings of the Second international conference on Machine Learning for Multimodal Interaction*. Edinburgh, UK: Springer, pp. 40–51.

Cheveigné, A. de & H. Kawahara (2002). 'YIN, a fundamental frequency estimator for speech and music'. In: *The Journal of the Acoustical Society of America* 111.4, pp. 1917–1930.

Cowie, R.; E. Douglas-Cowie; N. Tsapatsoulis; G. Votsis; S. Kollias; W. Fellenz & J. Taylor (2001). 'Emotion recognition in human-computer interaction'. In: *Signal Processing Magazine* 18.1, pp. 32–80.

Cowie, R.; E. Douglas-Cowie; S. Savvidou; E. McMahon; M. Sawey & M. Schröder (2000). ''FEELTRACE': An Instrument For Recording Perceived Emotion In Real Time'. In: *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*. Belfast, UK: Textflow, pp. 19–24.

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

Davis, S. & P. Mermelstein (1980). 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences'. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* 28.4, pp. 357–366.

Dellaert, F.; T. Polzin & A. Waibel (1996). 'Recognizing emotion in speech'. In: *Proceedings of the Fourth International Conference on Spoken Language*. Philadelphia, USA: IEEE, pp. 1970–1973.

Dennett, D. (1987). The Intentional Stance. Cambridge, USA: MIT Press.

Devillers, L.; L. Vidrascu & L. Lamel (2005). 'Challenges in real-life emotion annotation and machine learning based detection'. In: *Neural Networks* 18.4, pp. 407–422.

Diamantidis, N. A.; D. Karlis & E. A. Giakoumakis (2000). 'Unsupervised stratification of cross-validation for accuracy estimation'. In: *Artifical Intelligence* 116.1-2, pp. 1–16.

Dietterich, T. G. (1998). 'Approximate statistical tests for comparing supervised classification learning algorithms'. In: *Neural Computation* 10.7, pp. 1895–1923.

Dillon, R. (2005). 'A Possible Model for Predicting Listeners' Emotional Engagement'. In: *Third International Symposium on Computer Music Modeling and Retrieval*. Vol. 3902. Lecture Notes in Computer Science. Pisa, Italy: Springer, pp. 60–75.

Ehlich, K. & J. Rehbein (1979). 'Erweiterte halbinterpretative Arbeitstranskriptionen (HIAT2): Intonation'. In: *Linguistische Berichte* 59, pp. 51–75.

Ekman, P. (1992). 'Are there basic emotions?' In: *Psychological Review* 99, pp. 550–553.

Ekman, P. & W. V. Friesen (1978). Facial Action Coding System: Manual. Bd. 1-2. Consulting Psychologists Press.

Elsholz, J.-P.; G. de Melo; M. Hermann & M. Weber (2009). 'Designing an extensible architecture for Personalized Ambient Information'. In: *Pervasive and Mobile Computing* 5.5, pp. 592–605.

Engbert, I. & A. Hansen (1996). Documentation of the Danish Emotional Speech Database DES. Tech. rep. Aalborg, Denmark: Center for Person Kommunikation, Aalborg University.

Eyben, F.; M. Wöllmer & B. Schuller (2009). 'OpenEAR - Introducing the munich open-source emotion and affect recognition toolkit'. In: *Proceedings of 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, The Netherlands: IEEE, pp. 576–581.

Fant, G. (1960). Acoustic Theory of Speech Production. The Hague, The Netherlands: Mouton.

Farrús, M.; J. Hernando & P. Ejarque (2007). 'Jitter and shimmer measurements for speaker recognition'. In: *INTERSPEECH-2007*. Antwerp, Belgium, pp. 778–781.

Fragopanagos, N. & J. Taylor (2005). 'Emotion recognition in human–computer interaction'. In: *Neural Networks* 18.4, pp. 389–405.

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

Frommer, J.; B. Michaelis; D. Rösner; A. Wendemuth; R. Friesen; M. Haase; M. Kunze; R. Andrich; J. Lange; A. Panning & I. Siegert (2012b). 'Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus'. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation*. Istanbul, Turkey: ELRA.

Fry, D. B. (1992). The Physics of Speech. Cambridge, UK: Cambridge University Press.

Gerhard, D. (2003). Pitch Extraction and Fundamental Frequency: History and Current Techniques. Tech. rep. TR-CS 2003-06. Regina, Saskatchewan, Canada: Department of Computer Science, University of Regina.

Glodek, M.; S. Tschechne; G. Layher; M. Schels; T. Brosch; S. Scherer; M. Kächele; M. Schmidt; H. Neumann; G. Palm & F. Schwenker (2011). 'Multiple classifier systems for the classification of audio-visual emotional states'. In: *Affective Computing and Intelligent Interaction*. Ed. by D'Mello, S. D'Mello, S.; Graesser, A.; Graesser, A.; Schuller, B.; Schuller, B. & Martin, J.-C. & Martin, J.-C. Vol. 6975. Lecture Notes in Computer Science. Springer, pp. 359–368.

Glüge, S. (2013). 'Implicite Sequence Learning in Recurrent Neural Networks'. PhD thesis. Otto von Guericke University Magdeburg.

Glüge, S.; O. Hamid & A. Wendemuth (2010). 'A Simple Recurrent Network for Implicit Learning of Temporal Sequences'. In: *Cognitive Computation* 2 (4), pp. 265–271.

Gnjatovic, M. & D. Rösner (2008). 'On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System'. In: *Proceedings of the International Conference on Language Resources and Evaluation*. Marrakech, Marocco: ELRA, pp. 573–580.

Gold, B. & N. Morgan (2000). Speech and audio signal processing: processing and perception of speech and music. Hoboken, USA: John Wiley.

Goodwin, M. & C. Goodwin (2000). 'Emotion within Situated Activity'. In: *Linguistic Anthropology: A Reader*, pp. 239–257.

Goudbeek, M.; J. P. Goldman & K. R. Scherer (2009). 'Emotion dimensions and formant position'. In: *INTERSPEECH-2009*. Brighton, UK: ISCA, pp. 1575–1578.

Grimm, M.; K. Kroschel; E. Mower & S. Narayanan (2007). 'Primitives-based evaluation and estimation of emotions in speech'. In: *Speech Communication* 49.10-11, pp. 787–800.

Grimm, M.; K. Kroschel & S. Narayanan (2008). 'The Vera am Mittag German audio-visual emotional speech database'. In: *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo*. Hannover, Germany: IEEE, pp. 865–868.

Gustafson, J. & D. Neiberg (2010). 'Prosodic cues to engagement in non-lexical response tokens in Swedish'. In: *Proceedings of DiSS-LPSS Joint Workshop 2010*. Tokyo, Japan: ISCA, s.p.

Hansen, J. & S. Bou-Ghazale (1997). 'Getting started with SUSAS: a speech under simulated and actual stress database'. In: *EUROSPEECH-1997*. Rhodes, Greece: ISCA, pp. 1743–1746.

Harris, T. K. & A. I. Rudnicky (2007). 'TeamTalk: A platform for multi-human-robot dialog research in coherent real and virtual spaces'. In: *Proceedings of the National Conference on Artificial Intelligence.* Vancouver, Canada: Association for the Advancement of Artificial Intelligence, pp. 1864–1865.

Hartmann, K.; I. Siegert; S. Glüge; A. Wendemuth; M. Kotzyba & B. Deml (2012). 'Describing Human Emotions Through Mathematical Modelling'. In: *Proceedings of the MATHMOD 2012 - 7th Vienna International Conference on Mathematical Modelling.* Vienna, Austria: University Vienna, s.p.

Hartwig, M.; P. Granhag; L. Strömwall & A. Vrij (2002). 'Deception detection: Effects of conversational involvement and probing'. In: *Göteborg Psychological Reports* 32.2, pp. 1–12.

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

He, L.; M. Lech; N. Maddage; S. Memon & N. Allen (2009). 'Emotion Recognition in Spontaneous Speech within Work and Family Environments'. In: *3rd International Conference on Bioinformatics and Biomedical Engineering.* Beijing, China: IEEE, pp. 1–4.

Hermansky, H.; N. Morgan; A. Bayya & P. Kohn (1991). 'Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP).' In: *EUROSPEECH-1991.* Genova, Italy: ISCA, pp. 1367–1370.

Hiroshige, S.; H. Asano; M. Uchida & H. Ide (2009). 'Group behavior of agents with an emotional model'. English. In: *Artificial Life and Robotics* 14.1, pp. 67–70.

Hoffmann, H.; A. Scheck; T. Schuster; S. Walter; K. Limbrecht; H. C. Traue & H. Kessler (2012). 'Mapping discrete emotions into the dimensional space: An empirical approach'. In: *2012 IEEE International Conference on Systems, Man, and Cybernetics.* Seoul, Korea: IEEE, pp. 3316–3320.

Höök, K. (2013). 'Affective Computing'. In: *The Encyclopedia of Human-Computer Interaction, 2nd Ed.* Ed. by Soegaard, M. Soegaard, M. & Dam,

R. F. & Dam, R. F. Aarhus, Denmark: The Interaction Design Foundation, s.p.

Hübner, D.; B. Vlasenko; T. Grosser & A. Wendemuth (2010). 'Determining optimal features for emotion recognition from speech by applying an evolutionary algorithm'. In: *INTERSPEECH-2010.* Makuhari, Japan: ISCA, pp. 2358–2361.

Iliou, T. & C.-N. Anagnostopoulos (2010). 'Classification on Speech Emotion Recognition - A Comparative Study'. In: *International Journal On Advances in Life Sciences* 2, pp. 18–28.

Jaeger, H. (2001). The Echo State Approach to Analysing and Training Recurrent Neural Networks. Tech. rep. GMD - Forschungszentrum Informationstechnik GmbH.

Jaimes, A. & N. Sebe (2007). 'Multimodal human–computer interaction: A survey'. In: *Computer Vision and Image Understanding* 108.1–2, pp. 116 –134.

Janin, A.; D. Baron; J. Edwards; D. Ellis; D. Gelbart; N. Morgan; B. Peskin; T. Pfau; E. Shriberg; A. Stolcke & C. Wooters (2003). 'The ICSI Meeting Corpus'. In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing.* Hong Kong, China: IEEE, pp. 364–367.

Jiang, H. (2005). 'Confidence measures for speech recognition: A survey'. In: *Speech Communication* 45.4, pp. 455–470.

Kameas, A. D.; C. Goumopoulos; H. Hagras; V. Callaghan; T. Heinroth & M. Weber (2009). 'An Architecture That Supports Task-Centered Adaptation In Intelligent Environments'. In: *Advanced Intelligent Environments.* Ed. by Kameas, A. D. Kameas, A. D.; Callagan, V.; Callagan, V.; Hagras, H.; Hagras, H.; Weber, M.; Weber, M. & Minker, W. & Minker, W. Springer, pp. 41–66.

Kang, H. Y. (2012). Large group meetings in the preschool classroom: Co-constructing meaning making through group interaction. Ann Arbor, USA: ProQuest Information & Learning.

Karray, F.; M. Alemzadeh; J. A. Saleh & M. N. Arab (2008). 'Human-Computer Interaction: Overview on State of the Art'. In: *International Journal on Smart Sensing and Intelligent Systems* 1.1, pp. 137–159.

Keyton, J. & S. J. Beck (2009). 'The Influential Role of Relational Messages in Group Interaction'. In: *Group Dynamics* 13.1, pp. 14–30.

Kim, J. & E. André (2008). 'Emotion Recognition Based on Physiological Changes in Listening Music'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30.12, pp. 2067–2083.

Koelstra, S.; C. Mühl & I. Patras (2009). 'EEG analysis for implicit tagging of video data'. In: *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, The Netherlands: IEEE, pp. 27–32.

Kostoulas, T.; T. Ganchev & N. Fakotakis (2008). 'Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data'. In: *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*. Ed. by Esposito, A. Esposito, A.; Bourbakis, N.; Bourbakis, N.; Avouris, N.; Avouris, N. & Hatzilygeroudis, I. & Hatzilygeroudis, I. Vol. 5042. Lecture Notes in Computer Science. Springer, pp. 235–242.

Kotti, M.; F. Paterno & C. Kotropoulos (2010). 'Speaker-independent negative emotion recognition'. In: *2nd International Workshop on Cognitive Information Processing*. Elba, France: IEEE, pp. 417–422.

Krell, G.; M. Glodek; A. Panning; I. Siegert; B. Michaelis; A. Wendemuth & F. Schwenker (2013). 'Fusion of Fragmentary Classifier Decisions for Affective State Recognition'. In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Ed. by Schwenker, F. Schwenker, F.; Scherer, S.; Scherer, S. & Morency, L.-P. & Morency, L.-P. Vol. 7742. Lecture Notes in Computer Science. Springer, pp. 116–130.

Krippendorff, K. (2012). Content Analysis: An Introduction to Its Methodology. 3rd ed. Thousand Oaks, USA: SAGE Publications.

Krugman, H. (1965). 'The Impact of Television Advertising: Learning without Involvement'. In: *Public Opinion Quarterly* 29.3, pp. 349–356.

Kühn, T. & K.-V. Koschel (2010). Group discussions : a practical manual. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Kumar, R.; C. P. Rosé; Y.-C. Wang; M. Joshi & A. Robinson (2007). 'Tutorial Dialogue as Adaptive Collaborative Learning Support'. In: *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*. Amsterdam, The Netherlands: IOS Press, pp. 383–390.

Kuncheva, L. I. (2004). Combining Pattern Classifiers: Methods and Algorithms. Hoboken, USA: John Wiley.

Kwon, O.-W.; K. Chan; J. Hao & T.-W. Lee (2003). 'Emotion Recognition by Speech Signals'. In: *EUROSPEECH-2003*. Geneva, Switzerland: ISCA, pp. 125–128.

Lamnek, S. (2005). Gruppendiskussion. UTB für Wissenschaft : Uni-Taschenbücher. Weinheim, Germany: Beltz.

Lange, J. & J. Frommer (2011). 'Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion'. In: *Proceedings der 41. GI-Jahrestagung*. Vol. 192. Lecture Notes in Computer Science. Berlin, Germany: Bonner Köllen Verlag, pp. 240–254.

Lew, M.; E. M. Bakker; N. Sebe & T. S. Huang (2007). 'Human-Computer Intelligent Interaction: A Survey'. In: *Human–Computer Interaction*. Ed. by Lew, M. Lew, M.; Sebe, N.; Sebe, N.; Huang, T.; Huang, T. & Bakker, E. & Bakker, E. Vol. LNCS 4796. Lecture Notes in Computer Science. Springer, pp. 1–5.

Limbrecht-Ecklundt, K.; S. Rukavina; S. Walter; A. Scheck; D. Hrabal; J.-W. Tan & H. Traue (2013). 'The importance of subtle facial expressions for emotion classification in human-computer interaction.' In: *Emotional Expression: The Brain and The Face* 5.1. in press.

Looze, C. de; C. Oertel; S. Rauzy & N. Campbell (2011). 'Measuring dynamics of mimicry by means of prosodic cues in conversational speech'. In: *17th International Congress of Phonetic Sciences*. Hong Kong, China: ICPhS, pp. 1294–1297.

Lugger, M. & B. Yang (2008). 'Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters.' In: *Proceedings of the ICASSP-2008*. Las Vegas, USA: IEEE, pp. 4945–4948.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd. Mahwah, NJ: Lawrence Erlbaum Associates.

Makhoul, J.; F. Kubala; R. Schwartz & R. Weischedel (1999). 'Performance Measures For Information Extraction'. In: *Proceedings of DARPA Broadcast News Workshop*. Herndon, USA: Morgan Kaufmann Publishers Inc., pp. 249–252.

Manning, C. D. & H. Schütze (1999). Foundations of statistical natural language processing. Cambridge, USA: MIT Press.

Marquardt, D. W. (1963). 'An algorithm for least-squares estimation of nonlinear parameters'. In: *SIAM Journal on Applied Mathematics* 11.2, pp. 431–441.

Martin, O.; I. Kotsia; B. Macq & I. Pitas (2006). 'The eNTERFACE'05 Audio-Visual Emotion Database'. In: *Proceedings of the 22nd International Conference on Data Engineering Workshop*, s.p.

McCowan, I.; D. Gatica-Perez; S. Bengio; G. Lathoud; M. Barnard & D. Zhang (2005). 'Automatic Analysis of Multimodal Group Actions in Meetings'. In: *IEEE Transactions Pattern Analysis and Machine Intelligence* 27.3, pp. 305–317.

McCowan, I.; S. Bengio; D. Gatica-perez; G. Lathoud; F. Monay; D. Moore; P. Wellner & H. Bourlard (2003). 'Modeling Human Interaction in Meetings'. In: *Proceedings of the ICASSP-2003*. Hong Kong, China: IEEE, pp. 748–751.

McKeown, G.; M. Valstar; R. Cowie; M. Pantic & M. Schroder (2012). 'The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent'. In: *IEEE Transactions on Affective Computing* 3.1, pp. 5–17.

Mehrabian, A. (1996). 'Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament'. In: *Current Psychology* 14.4, pp. 261–292.

Melin, H. (1999). 'Databases For Speaker Recognition: Activities In Cost250 Working Group 2'. In: *Proceedings COST250 Workshop on Speaker Recognition in Telephony.* Rome, Italy: Springer, s.p.

Merriam-Webster (1998). Merriam-Webster's Collegiate Dictionary - Deluxe Edition. Merriam-Webster, Inc.

Merten, M.; A. Bley; C. Schroeter & H.-M. Gross (2012). 'A mobile robot platform for socially assistive home-care applications'. In: *Proceedings of 7th German Conference on Robotics.* Munich, Germany: IEEE, pp. 233–238.

Meudt, S.; L. Bigalke & F. Schwenker (2012). 'Atlas - Annotation tool using partially supervised learning and multi-view co-learning in human-computer-interaction scenarios'. In: *Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications.* San Francisco, USA: IEEE, pp. 1309–1312.

Molau, S.; M. Pitz; R. Schlüter & H. Ney (2001). 'Computing mel-frequency cepstral coefficients on the power spectrum'. In: *Proceedings of the ICASSP-2001.* Salt Lake City, USA: IEEE, pp. 73–76.

Muda, L.; M. Begam & I. Elamvazuthi (2010). 'Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques'. In: *Journal of Computing* 2.3, pp. 138–143.

Müller, V. C. (2011). 'Interaction and Resistance: The Recognition of Intentions in New Human-Computer Interaction'. In: *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues.* Ed. by Esposito, A. Esposito, A.; Esposito, A. M.; Esposito, A. M.; Martone, R.; Martone, R.; Müller, V. C.; Müller, V. C. & Scarpetta, G. & Scarpetta, G. Vol. 6456. Lecture Notes in Computer Science. Springer, pp. 1–7.

Nasoz, F.; C. Lisetti; K. Alvarez & N. Finkelstein (2003). 'Emotion recognition from physiological signals for user modeling of affect'. In: *International Journal of Cognition, Technology and Work – Special Issue on Presence* 6.1, pp. 4–14.

Nguyen, P. (2009). 'Techware: Speech recognition software and resources on the web'. In: *IEEE Signal Processing Magazine* 26.3, pp. 102–105.

Novick, D.; G. Adoneth; D. Manuel & I. Grís (2012). 'When the Conversation Starts: An Empirical Analysis'. In: *Joint Proceedings of the IVA 2012 Workshops.* Santa Cruz, USA: Otto von Guericke University Magdeburg, pp. 67–74.

Nwe, T. L.; S. W. Foo & L. C. De Silva (2003). 'Speech emotion recognition using hidden Markov models'. In: *Speech Communication* 41.4, pp. 603–623.

Oertel, C.; F. Cummins; N. Campbell; J. Edlund & P. Wagner (2010). 'D64: A corpus of richly recorded conversational interaction'. In: *Proceedings of LREC 2010 Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality.* Valetta, Malta: ELRA, pp. 27–30.

Oertel, C. (2010). 'Identification of Cues for the Automatic Detection of Hotspots'. MA thesis. Bielefeld University.

Oertel, C.; S. Scherer & N. Campbell (2011a). 'On the Use of Multimodal Cues for the Prediction of Degrees of Involvement in Spontaneous Conversation'. In: *INTERSPEECH-2011.* Florence, Italy: ISCA, pp. 1541–1544.

Oertel, C.; C. de Looze; S. Scherer; A. Windmann; P. Wagner & N. Campbell (2011b). 'Towards the Automatic Detection of Involvement in Conversation'. In: *Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issues.* Ed. by Esposito, A. Esposito, A.; Vinciarelli, A.; Vinciarelli, A.; Vicsi, K.; Vicsi, K.; Pelachaud, C.; Pelachaud, C. & Nijholt, A. & Nijholt, A. Vol. 6800. Springer, pp. 163–170.

Olson, D. L. & D. Delen (2008). Advanced Data Mining Techniques. 1st. Berlin/Heidelberg, Germany: Springer. ISBN: 3540769161.

Panning, A.; I. Siegert; A. Al-Hamadi; A. Wendemuth; D. Rösner; J. Frommer; G. Krell & B. Michaelis (2012). 'Multimodal affect recognition in spontaneous HCI environment'. In: *2012 IEEE International Conference on Signal Processing, Communication and Computing.* Hong Kong, China: IEEE, pp. 430–435.

Picard, R. (2000). Affective computing. Cambridge, USA: MIT Press.

Planet, S. & I. Iriondo (2012). 'Children's Emotion Recognition from Spontaneous Speech Using a Reduced Set of Acoustic and Linguistic Features'. In: *Cognitive Computation.* without issue assignment, pp. 1–7.

Plutchik, R. (2001). 'The Nature of Emotions'. In: vol. 89. 4. Sigma Xi, pp. 344–350.

Rabiner, L. R.; M. J. Cheng; A. E. Rosenberg & C. A. McGonegal (1976). 'A comparative performance study of several pitch detection algorithms'. In: *IEEE Transactions on Audio, Signal, and Speech Processing* 24, pp. 399–417.

Rabiner, L. & B. Juang (1993). Fundamentals of Speech Recognition. Upper Saddle River, USA: Prentice Hall.

Rasheed, U.; Y Tahir; J. Dauwels; S. Dauwels; D. Thalmann & N. Magnenat-Thalmann (2013). 'MULTI-MODAL BEHAVIOR DETECTION USING SPEECH CUES FOR REAL-TIME SOCIO-FEEDBACK'. In: *Proceedings of the ICASSP-2013.* Vancouver, Canada: IEEE.

Ratzka, A. (2010). Patternbasiertes User Interface Design für multimodale Interaktion. Glückstadt, Germany: Verlag Werner Hülsbusch.

Rogers, Y.; H. Sharp & J. Preece (2011). Interaction Design - Beyond Human-Computer Interaction. Hoboken, USA: John Wiley.

Rösner, D.; J. Frommer; R. Andrich; R. Friesen; M. Haase; M. Kunze; J. Lange & M. Otto (2012). 'LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing'. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation.* Istanbul, Turkey: ELRA, pp. 82–89.

Rösner, D.; R. Friesen; M. Otto; J. Lange; M. Haase & J. Frommer (2011). 'Intentionality in Interacting with Companion Systems – An Empirical Approach'. In: *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments.* Ed. by Jacko, J. A. Jacko, J. A. Vol. 6763. Lecture Notes in Computer Science. Springer, pp. 593–602.

Sapra, A.; N. Panwar & S. Panwar (2013). ‘Emotion Recognition from Speech’. In: *International Journal of Emerging Technology and Advanced Engineering* 3.2, pp. 341–345.

Scherer, K. (2005). ‘What are emotions? And how can they be measured?’ In: *Social Science Information* 44.4, pp. 695–729.

Scherer, K. R. (2001). ‘Appraisal considered as a process of multilevel sequential checking’. In: *Appraisal processes in emotion: Theory, methods, research*. Ed. by Scherer, K. R. Scherer, K. R.; Schorr, A.; Schorr, A. & Johnstone, T. & Johnstone, T. Oxford University Press, pp. 92–120.

Scherer, S.; M. Glodek; G. Layher; M. Schels; M. Schmidt; T. Brosch; S. Tschechne; F. Schwenker; H. Neumann & G. Palm (2012). ‘A generic framework for the inference of user states in human computer interaction: How patterns of low level behavioral cues support complex user states in HCI’. In: *Journal on Multimodal User Interfaces* 6.3-4, pp. 117–141.

Scherer, S. (2011). ‘Analyzing the user’s state in HCI: from crisp emotions to conversational dispositions’. PhD thesis. Ulm University.

Schmidt, T. & W. Schütte (2010). ‘FOLKER: An Annotation Tool For Efficient Transcription Of Natural, Multi-Party Interaction’. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. Valletta, Malta: ELRA, pp. 2091–2096.

Schuller, B.; M. Wimmer; D. Arsic; G. Rigoll & B. Radig (2007a). ‘Audiovisual Behavior Modeling by Combined Feature Spaces’. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Honolulu, Hawaii, USA: IEEE, pp. 733–736.

Schuller, B.; R. Müller; B. Hörnler; A. Höthker; H. Konosu & G. Rigoll (2007b). ‘Audiovisual recognition of spontaneous interest within conversations’. In: *Proceedings of the 9th International Conference on Multimodal Interfaces*. Nagoya, Japan: ACM, pp. 30–37.

Schuller, B.; B. Vlasenko; D. Arsic; G. Rigoll & A. Wendemuth (2008a). ‘Combining Speech Recognition and Acoustic Word Emotion Models for Robust Text-Independent Emotion Recognition’. In: *Proceedings of the 2008 IEEE*

International Conference on Multimedia and Expo, Hannover, Germany: IEEE, pp. 1333–1336.

Schuller, B.; X. Zhang & G. Rigoll (2008b). ‘Prosodic and spectral features within segment-based acoustic modeling’. In: *INTERSPEECH-2008.* Brisbane, Australia: ISCA, pp. 2370–2373.

Schuller, B.; B. Vlasenko; F. Eyben; G. Rigoll & A. Wendemuth (2009a). ‘Acoustic Emotion Recognition: A Benchmark Comparison of Performances’. In: *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop.* Merano, Italy, pp. 552–557.

Schuller, B.; M. Wöllmer; T. Moosmayr & G. Rigoll (2009b). ‘Recognition of noisy speech: A comparative survey of robust model architecture and feature enhancement’. In: *Eurasip Journal on Audio, Speech, and Music Processing* 2009.1, pp. 1–18.

Schuller, B.; B Vlasenko; F. Eyben; M. Wollmer; A. Stuhlsatz; A. Wendemuth & G. Rigoll (2010a). ‘Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies’. In: *IEEE Transactions on Affective Computing* 1.2, pp. 119–131.

Schuller, B.; M. Valstar; F. Eyben; G. McKeown; R. Cowie & M. Pantic (2011a). ‘AVEC 2011-the first international audio/visual emotion challenge’. In: *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II.* Memphis, USA: Springer, pp. 415–424.

Schuller, B.; Z. Zhang; F. Weninger & G. Rigoll (2011b). ‘Using multiple databases for training in emotion recognition: To unite or to vote?’ In: *INTERSPEECH-2011*, pp. 1553–1556.

Schuller, B.; S. Steidl; A. Batliner; A. Vinciarelli; K. Scherer; F. Ringeval; M. Chetouani; F. Weninger; F. Eyben; E. Marchi; M. Mortillaro; H. Salamin; A. Polychroniou; F. Valente & S. Kim (2013). ‘The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism’. In: *INTERSPEECH-2013.* Lyon, France: ISCA, pp. 148–152.

Schuller, B.; S. Steidl & A. Batliner (2009c). 'The INTERSPEECH 2009 Emotion Challenge'. In: *INTERSPEECH-2009*. Brighton, UK: ISCA, pp. 312–315.

Schuller, B.; S. Steidl; A. Batliner; F. Burkhardt; L. Devillers; C. A. Müller & S. S. Narayanan (2010b). 'The INTERSPEECH 2010 paralinguistic challenge'. In: *INTERSPEECH-2010*. Makuhari, Japan: ISCA, pp. 2794–2797.

Schuller, B.; A. Batliner; S. Steidl & D. Seppi (2011c). 'Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge'. In: *Speech Communication* 53.9-10, pp. 1062–1087.

Schuller, B.; S. Steidl; A. Batliner; F. Schiel & J. Krajewski (2011d). 'The INTERSPEECH 2011 Speaker State Challenge'. In: *INTERSPEECH-2011*. Florence, Italy: ISCA, pp. 3201–3204.

Schuller, B.; S. Steidl; A. Batliner; E. Nöth; A. Vinciarelli; F. Burkhardt; R. van Son; F. Weninger; F. Eyben; T. Bocklet; G. Mohammadi & B. Weiss (2012). 'The INTERSPEECH 2012 Speaker Trait Challenge'. In: *INTERSPEECH-2012*. Portland, USA: ISCA, s.p.

Schulz von Thun, F. (1981). Miteinander reden 1 - Störungen und Klärungen. Reinbek, Germany: Rowohlt.

Selting, M. et al. (2011). 'A system for transcribing talk-in-interaction: GAT 2'. In: *Gesprächsforschung Online-Zeitschrift zur verbalen Interaktion* 12.1, s.p.

Shneiderman, B. & C. Plaisant (2010). Designing the User Interface: Strategies for Effective Human-computer Interaction. Boston, USA: Addison-Wesley.

Shorter Oxford English Dictionary (2002). 5. Oxford University Press.

Siegert, I.; K. Hartmann; S. Glüge & A. Wendemuth (2012b). 'Modelling of Emotional Development within Human-Computer-Interaction'. In: *Proceedings of the 2. Interdisziplinärer Workshop*. Kognitive Systeme 1. Duisburg, Germany, s.p.

Simpson, D. (2002). Situatedness, or, Why We Keep Saying Where We're Coming From. Duke University Press.

Stefan, S.; I. Siegert; L. Bigalke & S. Meudt (2010). 'Developing an Expressive Speech Labeling Tool Incorporating the Temporal Characteristics of Emotion'. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation.* Valletta, Malta: ELRA, pp. 1172–1175.

Tan, J.-W.; S. Walter; A. Scheck; D. Hrabal; H. Hoffmann; H. Kessler & H. Traue (2012). 'Repeatability of facial electromyography (EMG) activity over corrugator supercilii and zygomaticus major on differentiating various emotions'. In: *Journal of Ambient Intelligence and Humanized Computing* 3.1, pp. 3–10.

Thiel, C.; S. Scherer & F. Schwenker (2007). 'Fuzzy-Input Fuzzy-Output One-Against-All Support Vector Machines'. In: *Knowledge-Based Intelligent Information and Engineering Systems.* Ed. by Apolloni, B. Apolloni, B.; Howlett, R. J.; Howlett, R. J. & Jain, L. & Jain, L. Vol. 4694. Lecture Notes in Computer Science. Springer, pp. 156–165.

Tolkmitt, F. J. & K. R. Scherer (1986). 'Effect of experimentally induced stress on vocal parameters.' In: *Journal of experimental psychology. Human perception and performance* 12.3, pp. 302–313.

Traum, D.; P. Aggarwal; R. Artstein; S. Foutz; J. Gerten; A. Katsamanis; A. Leuski; D. Noren & W. Swartout (2012). 'Ada and Grace: Direct Interaction with Museum Visitors'. In: *Intelligent Virtual Agents.* Ed. by Nakano, Y. Nakano, Y.; Neff, M.; Neff, M.; Paiva, A.; Paiva, A. & Walker, M. & Walker, M. Vol. 7502. Lecture Notes in Computer Science. Springer, pp. 245–251.

Veitch, J. A. & S. M. Kaye (1988). 'Illumination effects on conversational sound levels and job candidate evaluation'. In: *Journal of Environmental Psychology* 8.3, pp. 223–233.

Ververidis, D. & C. Kotropoulos (2006). 'Emotional speech recognition: Resources, features, and methods'. In: *Speech Communication* 48.9, pp. 1162–1181.

Viterbi, A. (1967). 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm'. In: *IEEE Transactions on Information Theory* 13.2, pp. 260–269.

Vlasenko, B.; B. Schuller; A. Wendemuth & G. Rigoll (2007). 'Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech'. In: *INTERSPEECH-2007*. Antwerp, Belgium: ISCA, pp. 2249–2252.

Vlasenko, B.; B. Schuller; K. Tadesse Mengistu; G. Rigoll & A. Wendemuth (2008). 'Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest'. In: *INTERSPEECH-2008*. Brisbane, Australia: ISCA, pp. 805–808.

Vlasenko, B.; D. Prylipko; D. Philippou-Hübner & A. Wendemuth (2011a). 'Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions'. In: *INTERSPEECH-2011*. Florence, Italy: ISCA, pp. 1577–1580.

Vlasenko, B. (2011). 'Emotion Recognition within Spoken Dialog Systems.' PhD thesis. Otto von Guericke University Magdeburg.

Vlasenko, B. & A. Wendemuth (2009). 'Processing affected speech within human machine interaction'. In: *INTERSPEECH-2009*. Brighton, UK: ISCA, pp. 2039–2042.

Vogt, T. & E. André (2005). 'Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition'. In: *IEEE International Conference on Multimedia and Expo 2005*. Amsterdam, The Netherlands: IEEE, pp. 474–477.

Wageman, R. (1995). 'Interdependence and Group Effectiveness'. In: *Administrative Science Quarterly* 40.1, pp. 145–180.

Wagner, J.; E. André; F. Lingenfelser & J. Kim (2011). 'Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data'. In: *IEEE Transactions on Affective Computing* 2.4, pp. 206–218.

Wahlster, W. (2006). SmartKom: Foundations of Multimodal Dialogue Systems. Cognitive Technologies. Berlin/Heidelberg, Germany: Springer.

Walter, S.; J. Kim; D. Hrabal; S. C. Crawcour; H. Kessler & H. C. Traue (2013). 'Transsituational Individual-Specific Biopsychological Classification of Emotions'. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 43.4, pp. 988–995.

Wendemuth, A. (2004). Grundlagen der stochastischen Sprachverarbeitung. München, Germany: Oldenbourg.

Wendemuth, A. & S. Biundo (2012). 'A Companion Technology for Cognitive Technical Systems'. In: *Cognitive Behavioural Systems. COST 2102.* Ed. by Esposito, A. Esposito, A.; Esposito, A. M.; Esposito, A. M.; Vinciarelli, A.; Vinciarelli, A.; Hoffmann, R.; Hoffmann, R. & Müller, V. C. & Müller, V. C. Vol. 7403 LNCS. Dresden, Germany: Springer, pp. 89–103.

Wendt, B. & H. Scheich (2002). 'The "Magdeburger Prosodie Korpus" - a spoken language corpus for fMRI-Studies'. In: *Proceedings of the Speech Prosody 2002.* Aix-en-Provence, France: SProSIG, s.p.

Wilks, Y. (2005). 'Artificial companions'. In: *Interdisciplinary Science Reviews* 30.2, pp. 145–152.

— (2006). Artificial Companions as a new kind of interface to the future Internet. Tech. rep. Oxford, UK: Oxford Internet Institute, University of Oxford.

Williams, R. J. & D. Zipser (1995). 'Gradient-based learning algorithms for recurrent networks and their computational complexity'. In: *Back-propagation: Theory, Architectures and Applications.* Ed. by Chauvin, Y. Chauvin, Y. & Rumelhart, D. E. & Rumelhart, D. E. Hillsdale, USA: L. Erlbaum Associates Inc. Chap. 13, pp. 433–486.

Witten, I. H. & E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques. 2nd. San Francisco, USA: Morgan Kaufmann.

Wöllmer, M.; F. Eyben; B. Schuller & G. Rigoll (2009). 'Robust vocabulary independent keyword spotting with graphical models'. In: *2009 IEEE Workshop on Automatic Speech Recognition & Understanding.* Merano, Italy: IEEE, pp. 349–353.

Wrede, B. & E. Shriberg (2003). 'Spotting "hot spots" in meetings: human judgments and prosodic cues'. In: *INTERSPEECH-2003.* Geneva, Switzerland: ISCA, pp. 2805–2808.

Xiao, Z.; E. Dellandrea; L. Chen & W. Dou (2009). ‘Recognition of emotions in speech by a hierarchical approach’. In: *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Amsterdam, The Netherlands: IEEE, pp. 312–319.

Young, S.; G. Evermann; M. Gales; T. Hain; D. Kershaw; X. Liu; G. Moore; J. Odell; D. Ollason; D. Povey; V. Valtchev & P. Woodland (2009). The HTK Book, version 3.4. Cambridge University Engineering Department.

Yu, C.; P. M. Aoki & A. Woodruff (2004). ‘Detecting User Engagement in Everyday Conversations’. In: *Proceedings of the 8th International Conference on Spoken Language Processing*. Jeju Island, Korea: ISCA, pp. 1329–1332.

Zeng, Z.; M. Pantic; G. I. Roisman & T. S. Huang (2009). ‘A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.1, pp. 39–58.

Ehrenerklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die Hilfe eines kommerziellen Promotionsberaters habe ich nicht in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Verwendete fremde und eigene Quellen sind als solche kenntlich gemacht.

Ich habe insbesondere nicht wissentlich:

- Ergebnisse erfunden oder widersprüchliche Ergebnisse verschwiegen,
- statistische Verfahren absichtlich missbraucht, um Daten in ungerechtfertigter Weise zu interpretieren,
- fremde Ergebnisse oder Veröffentlichungen plagiiert,
- fremde Forschungsergebnisse verzerrt wiedergegeben.

Mir ist bekannt, dass Verstöße gegen das Urheberrecht Unterlassungs- und Schadensersatzansprüche des Urhebers sowie eine strafrechtliche Ahndung durch die Strafverfolgungsbehörden begründen können.

Ich erkläre mich damit einverstanden, dass die Dissertation ggf. mit Mitteln der elektronischen Datenverarbeitung auf Plagiate überprüft werden kann.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 25.09.2013

Dipl.-Inf. Ronald Böck