OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INSTITUT FÜR INFORMATIONS- UND
KOMMUNIKATIONSTECHNIK (IIKT)

Emotional and User-Specific Cues for Improved Analysis of Naturalistic Interactions

DISSERTATION

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

von

Dipl.-Ing. Ingo Siegert

geb. am 13.05.1983 in Wernigerode

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

Gutachter: Prof. Dr. rer. nat. Andreas Wendemuth
Prof. Dr.-Ing. Christian Diedrich
Prof. Dr.-Ing. Michael Weber

Promotionskolloquium am 18.03.2015

Wenn wir das Muster nicht erkennen
dann heißt das noch lange nicht, dass es kein Muster gibt.

WITTGENSTEIN – LUMEN

Danksagung

Nach vielen Jahren intensiver Arbeit liegt sie nun vor Ihnen: meine Dissertation. Damit ist es an der Zeit, mich bei denjenigen zu bedanken, die mich in dieser spannenden Phase meiner akademischen Laufbahn begleitet haben.

Als Erstes möchte ich mich bei meinem Doktorvater Prof. Dr. Andreas Wendemuth bedanken. Nicht nur für die Möglichkeit diese Arbeit am Lehrstuhl Kognitive Systeme durchführen zu können sowie die Unterstützung während der Bearbeitung, sondern auch für das Vertrauen und die Wertschätzung, die mir während der gesamten Promotionszeit entgegen gebracht wurde.

Bedanken möchte ich mich auch bei Herrn Prof. Dr. Michael Weber (Universität Ulm) und Herrn Prof. Dr. Christian Diedrich (Otto-von-Guericke-Universität Magdeburg) für die Bereitschaft die vorgelegte Dissertation zu begutachten.

Ein nicht unwesentlicher Teil dieser Arbeit ist am Lehrstuhl für Kognitive Systeme entstanden. Dass dies gelingen konnte, habe ich auch meinen Kollegen und Kolleginnen zu verdanken. Dr. Ronald Böck danke ich besonders für die Nachsicht, wenn ich mal wieder mit meinen Ideen in sein Büro gestürmt bin, da daraus am Ende gute Veröffentlichungen entstanden sind. David Philippou-Hübner und Tobias Grosser danke ich dafür, mich in den Anfangszeiten in der Arbeitsgruppe willkommen geheißen zu haben. Kim Hartmann danke ich besonders für die vielen Diskussionen und Nachfragen, die mich dazu gebracht haben, meine Ideen noch besser auszuformulieren. Ich danke auch allen hier nicht namentlich genannten Kollegen und Kolleginnen für die bereichernden Tipps und Diskussionsbeiträge, die mich wiederholt in neue thematische Bahnen gelenkt haben.

Dem Land Sachsen-Anhalt sei für die Bereitstellung meiner Stelle gedankt, dadurch war es mir möglich viel Erfahrung in der Betreuung von Studierenden und in Lehrveranstaltungen zu sammeln.

Außerdem möchte ich dem Sonderforschungsbereich SFB/TRR 62 „Eine Companion Technologie für Kognitive Technische Systeme“, gefördert durch die Deutsche Forschungsgemeinschaft, danken. Dieses Projekt bot mir eine Plattform, um mich mit anderen Doktoranden auszutauschen und meine Forschung im Kontext der sprachbasierten Emotionserkennung auch interdisziplinär zu vertiefen.

Ganz besonders danken möchte ich auch meiner Familie und meinen Freunden, die mich in all der Zeit unterstützt haben. Ohne euch wäre diese Arbeit nicht möglich geworden und ich nicht der, der ich bin.

Einen besonderen Dank möchte ich an meine Eltern richten, die mir mein Studium überhaupt ermöglicht haben. Bastian Ansorge für eine besondere Freundschaft seit fast 30 Jahren. Meiner Tochter Lea danke ich dafür, mir meine Zeit mit Ihr immer zu ganz besonderen Momenten gemacht zu haben. Meiner Verlobten Stephanie danke ich von ganzem Herzen für ihre unermüdliche Unterstützung, ihre Liebe und Motivation.

Zusammenfassung

DIE Mensch-Maschine-Interaktion erfährt in letzter Zeit immer größere Aufmerksamkeit. Hierbei geht es nicht nur darum, eine möglichst einfache Bedienung von technischen Systemen zu ermöglichen, sondern auch darum, eine möglichst natürliche Interaktion abzubilden. Gerade der sprachbasierten Interaktion kommt hierbei eine erhöhte Aufmerksamkeit zu. Zum Beispiel bieten moderne Smartphones und Fernseher eine robuste Sprachsteuerung an, was auf vielfältige technische Verbesserungen der letzten Jahre zurückzuführen ist.

Dabei wirkt die Sprachsteuerung immer noch artifiziell. Es können nur in sich geschlossene Dialoge mit kurzen Aussagen geführt werden. Zudem wird nur der Sprachinhalt ausgewertet. Die Art und Weise, wie etwas gesagt wird, bleibt unberücksichtigt, obwohl von der menschlichen Kommunikation bekannt ist, dass insbesondere die geäußerte Emotion für eine erfolgreiche Kommunikation wichtig ist. Ein relativ neuer Forschungszweig, das „Affective Computing“, hat unter anderem zum Ziel, technische Geräte zu entwickeln, die Emotionen erkennen, interpretieren sowie adäquat darauf reagieren können. Hierbei kommt der automatischen Emotionserkennung eine gewichtige Rolle zu.

Für die Emotionserkennung ist es wichtig zu wissen, wie sich Emotionen darstellen und wie sie sich äußern. Hierfür ist es hilfreich, sich auf empirische Erkenntnisse der Emotionspsychologie zu stützen. Leider gibt es keine einheitliche Darstellung von Emotionen. Auch die Beschreibung geeigneter emotionsunterscheidender akustischer Merkmale ist in der Psychologie eher deskriptiv gehalten. Deshalb wird für die automatische Erkennung auf erprobte Methoden der automatischen Spracherkennung zurückgegriffen, die sich auch für die Emotionserkennung als geeignet gezeigt haben.

Die automatische Emotionserkennung ist, wie auch die Spracherkennung, ein Zweig der Mustererkennung und im Gegensatz zur Emotionspsychologie datengetrieben, d.h., die Erkenntnisse werden aus Beispieldaten gewonnen. In der Emotionserkennung lassen sich die Phasen „Annotation“, „Modellierung“ und „Erkennung“ unterscheiden. Die Annotation kategorisiert Sprachdaten nach vordefinierten Emotionsbegriffen. Die Modellierung erzeugt Erkenner, um Daten automatisch zu kategorisieren. Die Erkennungsphase führt eine vorher unbekannte Zuordnung von Daten zu Emotionsklassen durch.

In den Anfangszeiten hat sich die automatische Emotionserkennung aufgrund des Mangels an geeigneten Datensätzen meist auf gespielte und sehr expressive Emotionsausdrücke gestützt. Hier konnten, mit aus der Spracherkennung bekannten Merkmalen und Erkennungsmethoden, sehr gute Erkennungsergebnisse von über 80 % bei

der Unterscheidung von bis zu sieben Emotionen erzielt werden. Für die Mensch-Maschine-Interaktion waren diese Erkenner jedoch ungeeignet, da in dieser die Emotionen weniger stark ausgeprägt sind. Daher wurden in Zusammenarbeit mit Psychologen naturalistische Interaktionsszenarien beschrieben und entsprechende Datensätze mit Probanden aus unterschiedlichen Personengruppen erhoben, denen keine „zu spielenden“ Vorgaben gemacht wurden, da sie natürlich reagieren sollten. Das hat dazu geführt, dass sich die Erkennungsraten auf diesen Daten verschlechtert haben und nur noch um die 60 % betragen. Aus dieser Entwicklung ergeben sich offene Fragen, die in dieser Arbeit untersucht werden sollen. Es wird vor allem untersucht, ob zusätzlich technische beobachtbare Marker die Emotionserkennung und Interaktionssteuerung in natürlicher Mensch-Maschine-Interaktion verbessern.

Die *erste offene Frage* beschäftigt sich mit der Generierung einer reliablen Klassenzuordnung für Emotionsdaten. Da bei natürlichen Interaktionen die Emotionsreaktionen nicht mehr vorgegeben sind, muss eine Klassenzuordnung durch geeignete Annotation im Nachhinein erstellt werden. Dabei ist vor allem die erreichbare Reliabilität wichtig. In der vorliegenden Arbeit konnte gezeigt werden, dass für eine naturalistische Mensch-Maschine-Interaktion die Reliabilität gesteigert werden kann, wenn Audio- und Video-Daten in Verbindung mit dem Kontext zur Annotation genutzt werden. Eine weitere Steigerung der Reliabilität und die Vermeidung des zweiten Kappa-Paradoxes kann erreicht werden, wenn die emotionalen Bereiche der Daten vorselektiert werden. Damit ist es möglich, eine Annotation hoher Güte zu erhalten.

Die *zweite offene Frage* untersucht inwieweit bestimmte Sprechercharakteristiken zur Verbesserung der Emotionserkennung herangezogen werden können. Der Vokaltrakt unterscheidet sich zwischen Männern und Frauen und ist auch durch Alterserscheinungen einer Veränderung unterworfen. Dies beeinflusst akustische Merkmale, die für die Emotionserkennung charakteristisch sind. Diese Arbeit untersucht, ob sowohl das Geschlecht als auch die Altersgruppe der Sprecher für die Emotionserkennung berücksichtigt werden müssen. Anhand von Experimenten mit verschiedenen Datensätzen konnte gezeigt werden, dass die Erkennungsleistung durch Berücksichtigung des Geschlechts oder der Altersgruppe nicht nur verbessert wurden, sondern dies in vielen Fällen auch signifikant war. In einigen Fällen konnte die Kombination beider Sprechercharakteristiken sogar eine weitere Verbesserung erzielen. Ein Vergleich mit einer Technik, die die anatomischen Unterschiede des Vokaltrakts normalisiert, zeigt, dass diese zwar auch eine Verbesserung gegenüber einer Nichtnormalisierung bringt, aber hinter den geschlechts- und altersgruppenspezifischen Modellen zurückbleibt.

Anschließend wurde die geschlechts- und altersgruppenspezifische Modellierung für die Fusion kontinuierlicher, fragmentierter, multimodaler Daten genutzt. Es konnte gezeigt werden, dass auch in diesem Fall, obwohl die Sprachdaten nicht über den kompletten Datenstrom verfügbar waren, eine Verbesserung der Fusionsleistung möglich ist.

Die *dritte offene Frage* erweitert den Untersuchungsgegenstand auf Interaktionen und untersucht, ob bestimmte akustische Feedbacksignale für eine emotionale Auswertung genutzt werden können. Hierbei konzentriert sich diese Arbeit auf Diskurspartikel, wie z.B. „hm“ oder „äh“. Dies sind kurze sprachliche Äußerungen, die den Sprechfluss unterbrechen. Da sie semantisch bedeutungslos sind, ist ausschließlich Ihre Intonation relevant. Zuerst wird untersucht, ob diese Partikel als Indikator für eine Bedienungsunsicherheit beim Benutzer dienen können. Es konnte gezeigt werden, dass bei anspruchsvollen Dialogen signifikant mehr Diskurspartikel genutzt werden als bei unkomplizierten Dialogen. Das Besondere an den Diskurspartikeln ist weiterhin, dass sie je nach Intonationsverlauf bestimmte Bedeutungen im Dialog übernehmen. Sie können Nachdenken, Initiativübernahme oder Nachfragen ankündigen. In dieser Arbeit konnte gezeigt werden, dass alleine über den Intonationsverlauf die am häufigsten auftretende Bedeutung „nachdenkend“ robust von allen anderen Dialogfunktionen unterschieden werden kann.

Die Bearbeitung der *vierten und letzten offenen Frage* beschäftigt sich mit der zeitlichen Modellierung von Emotionen. Wenn im technischen System die Emotionen des Nutzers sprechergruppenspezifisch erfasst und auch die jeweils geäußerten Interaktionssignale richtig gedeutet werden können, muss das System adäquat reagieren. Diese Reaktion sollte jedoch nicht auf einer einzelnen Äußerung des Nutzers beruhen, sondern seine langfristige emotionale Entwicklung berücksichtigen. Für diesen Zweck wurde in der Arbeit ein Stimmungsmodell vorgestellt, welches durch beobachtete Emotionsverläufe die Stimmung berechnet. Weiterhin konnte auch der Individualität des Nutzers Rechnung getragen werden, indem das Persönlichkeitsmerkmal der „Extraversion“ in das Modell integriert werden konnte.

Natürlich ist es nicht möglich, die in dieser Arbeit identifizierten offenen Fragen restlos zu klären. Die Erweiterung der puren akustischen Emotionserkennung durch Berücksichtigung von anderen Modalitäten, Sprechercharakteristiken, Feedbacksignalen und Persönlichkeitsmerkmalen erlaubt es jedoch, länger andauernde natürliche Interaktionen zu untersuchen und dialogkritische Situationen zu erkennen. Technische Systeme, die diese erweiterte Emotionserkennung nutzen, passen sich an ihren Nutzer an und werden so zu seinem Begleiter und letztendlich zu seinem *Companion*.

Abstract

THE Human-Computer Interaction recently received an increased attention. This is not just a matter of making the operation of technical systems as simple as pissuble, but also to enable a possibly natural interaction. In this context, especially the speech-based operation gained an increased attention. For example, modern smart phones and televisions offer a robust voice control, which is attributed to various technical improvements in recent years.

Nevertheless, voice control still seems artificial. Only self-contained dialogues with short statements can be managed. Furthermore, just the content of speech is evaluated. The way in which something is said remains unconsidered, although it is known from human communication, that the transmitted emotion is important in order to communicate successfully. A relatively new branch of research, the “Affective Computing”, has, amongst other objectives, the aim to develop technical systems that recognise and interpret emotions and respond to them appropriately. In this case, speech-based automatic emotion recognition has a major role.

For emotion recognition, it is important to know how emotions can be presented and how they are expressed. For this purpose, it is helpful to rely on empirical evidences of the psychology of emotions. Unfortunately, there is no uniform representation of emotions. Also the definition of appropriate emotion-distinctive acoustic features is rather descriptive in psychology. Therefore, the automatic detection of emotions is based on proven methods of automatic speech recognition, which have also been shown as appropriate for emotion recognition.

Automatic emotion recognition is, as speech recognition, a branch of pattern recognition. Contrary to emotion psychology, it is data-driven, that means insights are gathered from sampled data. For emotion recognition the phases “annotation”, “modelling” and “recognition” are distinguishable. The annotation categorises speech data according to predefined emotion-terms. Modelling generates recognisers to categorise data automatically. Recognition performs a previously unknown allocation of data to emotional classes.

In the beginning, automatic emotion recognition was usually based – due to the lack of suitable data sets – on acted and very expressive emotional expressions. In this case, based on features and detection methods known from speech recognition, very good recognition results of over 80 % in distinguishing of up to seven emotions could be achieved. However, for human-machine interaction these recognisers were unsuitable because in this case emotions are not that expressive. Therefore, in collaboration with

psychologists, naturalistic interaction scenarios were developed to collect relevant data sets with subjects from different groups of persons, who were not given specifications for “acting”.

This led to decreased recognition rates of only 60 % on these data. From this development, open issues arise, which will be investigated in this thesis. In particular, this thesis examines if further technically observable cues improve the emotion recognition and interaction control in naturalistic human-machine interaction.

The *first open issue* deals with the generation of a reliable class assignment of emotional data. Since in natural interactions the emotional reactions are not specified, a class allocation has to be created after the recording by a suitable annotation. In doing so, the achievable reliability is particularly important. In the present thesis, it could be shown that for a naturalistic human-machine interaction the reliability can be increased if audio and video data in combination with the context are used for the annotation. A further increase of reliability and an avoidance of the second Kappa paradox can be achieved if the emotional phases of the data are preselected. This makes it possible to obtain a high quality annotation.

The *second open issue* examines to what extent certain speaker characteristics can be utilised to improve the emotion recognition. The vocal tract differs between male and female speakers and is also changed due to aging, which affects the acoustic features that are characteristic for emotion recognition. This work investigated whether both the gender and the age-group of speakers have to be considered for emotion recognition. Through experiments with different datasets it could be shown that the recognition performance was significantly improved considering the gender or the age group. In some cases a combination of both speaker characteristics could achieve an even further improvement. A comparison show that a method normalising the vocal tract’s anatomical differences improves the recognition in comparison to the non-normalised case, however it falls behind results using the gender and age-group specific models.

Subsequently, the gender and age-group specific modelling was extended to the fusion of continuous, fragmentary, multimodal data. It could be shown that also in this case, although the speech data were not available for the entire data stream, an improvement in the fusion recognition is possible.

The *third open issue* expands the object of investigation to interactions and examines whether certain acoustic feedback signals can be used for an emotional evaluation. This work focuses on discourse particles, such as “hm” or “uh”. These are short vocalizations, interrupting the flow of speech. As they are semantically meaningless, only their intonation is relevant. First, it is examined whether they can serve as an indicator of a user operating under uncertainty. It could been shown that in challenging dialogues significantly more discourse particles are used than in simple dialogues. A further special feature of discourse particles is that they have specific functions in a dialogue depending on their intonation. So they can denote thinking, turn-taking, or requests. In this work, it could be shown that the most common meaning “thinking” is robustly distinguishable from all other dialogue functions by using the intonation only.

The *fourth and final open issue* deals with a temporal modelling of emotions. If a technical system is able to capture emotions in a speaker-group-specific manner and to correctly interpret the uttered interaction patterns, the system finally has to respond properly. However, this reaction should not only be based on a single utterance of the user, but should also consider his long-term emotional development. For this purpose, a mood-model was presented, where the mood is calculated from the course of observed emotions. Furthermore, the individuality of the user is taken into account by integrating the personality trait of extraversion into the model.

Of course it is not possible to resolve the open issues identified in this thesis completely. The extension of the pure acoustic emotion recognition by considering further modalities, speaker characteristics, feedback signals and personality traits allows however to examine longer-lasting natural interactions and dialogues and to identify critical situations. Technical systems that use this extended emotion recognition adapt to their users and thus become his attendant and ultimately his *companion*.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

IN the present time, technology plays an increasingly important role in people's lives. Especially their operation requires an interaction between human and machine. But this interaction is predominantly unidirectional. The technical system offers certain input options, the human operator only utters a selected option in a command-transmitting fashion and the system (hopefully) performs the desired action or gives an appropriate respond. However, in the everyday use of modern technology, the Human-Computer Interaction (HCI) is getting more complex, whereas it should still remain user-friendly. This requires flexible interfaces allowing a bidirectional dialogue, where humans and machines are equal.

In this context, the importance of speech-based interfaces is increasing over the classical HCI interfaces, such as keyboard and display. Through continually improved speech recognition and speech understanding ability in recent years, the efficiency of dialogue systems has increased rapidly. Automatic Speech Recognition (ASR) systems get more robust and popular. Today they can be found in several everyday technologies, such as smart-phones or navigation devices (cf. [Carroll 2013]).

Although these technical systems imitate the human interaction to allow also naive users to easily operate them, they do not take into account that Human-Human Interaction (HHI) is always socially situated and that interactions are not just isolated but part of larger social interplays. Thus, today's HCI research argues that computer systems should be capable of sensing agreement or inattention, for instance. Furthermore, they should be and capable of adapting and responding to these social signals in a polite, unintrusive, or persuasive manner (cf. [Vinciarelli et al. 2009]). Therefore, these systems have to be adaptable to the users' individual skills, preferences, and current emotional states (cf. [Wendemuth & Biundo 2012]). This aim however, is only successful, if engineers, psychologists, and computer scientists cooperate.

This chapter introduces the development of HCI briefly with the purpose of explaining the need for an enriched HCI and provides the reader with its basic ideas. Afterwards, the basic principle of a technical affect recognition is discussed, which is the motivation for my specific research topics. At the end of this chapter, the structure of this thesis is presented.

1.1 Enriched Human Computer Interaction

HCI research has gone through a huge development in the last decades. The research was and is focused on developing easy-to-use interfaces that could be used by experts as well as by novices. Today, several kinds of interfaces can be distinguished. A historical overview on HCI is given in [Carroll 2013]. Important developments, discussed by Carroll, showing the need for an enriched HCI, are presented in the following.

In the beginning of computer systems, the operation was mostly reserved for the developing institutions and selected scientists. Only experts were able to interact with these systems. Furthermore, the development was focused on improving the system's performance. Thus, the development of easy-to-use user interfaces was of scant importance. Up to the 1970s, the way of interaction was fixed to Command Line Interfaces, where the user had to type textualised commands to control the system (cf. [Carroll 2013]). This changed in the 1980s when home computers and later personal computers became more important and the Graphical User Interface (GUI) emerged. At this time, the upcoming computers systems could be easily used by trained users. A GUI allowed the user to use a pointing device to control the interaction. The GUI mimics a real desktop with objects that can be placed[1]. Thus, the interaction is simplified as it is supposed that the user is used to working at a desk and hence, manages the interaction with a computer as well. Since the 1990s, the desktop metaphor went through several adjustments, for instance, additional menu bars or docks. Even today, the interaction with technical systems still follows this Window, Icon, Menu, Pointing device (WIMP) paradigm. Smart-phone devices still use WIMP elements. But they open up a new era of post-WIMP interfaces: Touch-screen-based interaction now allows new manipulation actions such as “pinching” and “rotating”, as well as presenting additional information more naturally (cf. [Elsholz et al. 2009]). This kind of interaction is considered as more natural, as the user now can directly manipulate the icons instead of making a detour by using a computer mouse (cf. [Elsholz et al. 2009]). But, icons and folders are still parts of these post-WIMP GUIs (cf. [Rogers et al. 2011]). Moreover a new era of ubiquitous computing

[1] This is known as the desktop metaphor (cf. [Carroll 2013]).

devices is showing up, demanding for a more natural way of interaction, as the standard computer devices are moving to the background and the interaction is more integrated (cf. [Elsholz et al. 2009; Carroll 2013]).

Unfortunately, this kind of interaction is quite unnaturalistic as the users have to manipulate iconic presentations. Thus, also research on speech-based interaction gained a lot of interest. The technological development of speech recognition systems is given in [Juang & Rabiner 2006]. The Speech User Interface (SUI) research was motivated by the fact that speech is the primary mode of communication. It has gained a great deal of attention since the 1950s. The first speech recognition systems could only understand a few digits or words, spoken in an isolated form and thus, implicitly assuming that the unknown utterance contained one and only one complete term (cf. [Davis et al. 1952]). Ten years later, the work of Sakay and Doshita involved the first use of a speech segmenter to overcome this limitation. Their work can be considered as a precursor to a continuous speech recognition (cf. [Sakai & Doshita 1962]). Another early speech recognition system made use of statistical information about the allowable phoneme sequence of words (cf. [Denes 1959]). This technique was later considered again for statistical language modelling. In the 1970s, speech recognition technology made major strides, thanks to the interest of and funding from the U.S. Department of Defense. The fundamental concepts of Linear Predictive Coding (LPC) were formulated by Atal & Hanauer. This technique greatly simplified the estimation of the vocal tract response from a speech waveform (cf. [Atal & Hanauer 1971]). Main efforts are made in n-gram language modelling, which are used to transcribe a sequence of words, and in statistical modelling techniques controlling the acoustic variability of various speech representations across different speakers (cf. [Jelinek et al. 1975]). Furthermore, the concept of dynamic time warping has become an indispensable technique for ASR systems (cf. [Sakoe & Chiba 1978]). In the 1980s, a more rigorous statistical modelling framework for ASR systems became popular. Although the basic idea of Hidden Markov Model (HMM) was known earlier[2], this technique did not have its breakthrough till then (cf. [Levinson et al. 1983]). In the 1990s, ASR systems became more sophisticated and supported large vocabulary and continuous speech. Furthermore, the first customer speech recognition products emerged. Also, well-structured systems arose for researching and developing new concepts. The Hidden Markov Toolkit (HTK) developed by the Cambridge University team was and is one of the most widely used software for ASR research (cf. [Young et al. 2006]).

Today the interaction can rely on plentiful resources. GUIs and SUIs co-exist in many technical devices. GUIs are well known and in transition due to touch-screen-based interfaces, allowing a more direct manipulation (cf. [Elsholz et al. 2009; Kameas

[2] The idea of HMMs was described first in the late 1960s (cf. [Baum & Petrie 1966]).

et al. 2009; Rogers et al. 2011]). Today's SUIs can be used to control technical systems speaker-independently and also under noisy conditions (cf. [Juang & Rabiner 2006]). But speaker dependent continuous large vocabulary transcription also achieves high accuracy rates (cf. [Zhan & Waibel 1997]). Although these types of interaction are quite reliable and robust, they are still very artificial, since only the speech content of the user input is processed. Especially in comparison to an HHI, these interfaces are still lacking of the opportunities of a HHI. In HHI, speech is the natural way of interaction, it is not only used to transmit the pure content of the message but also to transmit further aspects, as appeal, relationship, or self-revelation.

Two researchers are heavily related with the human communication theory, Thun and Watzlawick. Thun discussed the many aspects of human communication and introduced his "four-sides model" (cf. [Thun 1981]). This model illustrates that every communication has four aspects. Regarding its understanding, a message can be interpreted by both sender and receiver. The factual information is just one perspective and not always the most important one. The appeal, the relationship, and the self-revelation also play important roles. The self-revelation is of special importance for the appraisal of the speaker's message. Although this could complicate the human communication, it is very important to make assumptions about the user's affective state, his wishes and intentions (cf. [Thun 1981]). Watzlawick investigated human communication and formulates five axioms (cf. [Watzlawick et al. 1967]), where the axiom: "One cannot not communicate" [Watzlawick et al. 1967] is the most important. By this, they emphasised the importance of the non-verbal behaviour. Thus, for them HHI is usually understood as a mixture of speech, facial expressions, gestures, and body postures. This is what in HCI research is called multimodal interaction.

These considerations are not only valid for HHI but also for HCI. Although factual information is in the focus, users also create a relationship with the system (cf. [Lange & Frommer 2011]). Thus, it it important to know how something has be said in HCI as well. Further motivated by the book "Affective Computing" by Picard & Cook, the vision emerged that future technical systems should provide a more human-like way of interaction while taking into account human affective signals (cf. [Picard & Cook 1984]). This area of research has received increased attention since the mid-2000's, as more and more researchers combined psychological findings with computer science (cf. [Zeng et al. 2009]). The terms "affect" and "affective state" are used quite all-encompassing to describe the topics of emotion, feelings, and moods, even though "affect" is commonly used interchangeably with the term "emotion". In the following thesis I will use the term "affective state" when talking about affects in general and the term "emotion", when a specific emotional concept is meant.

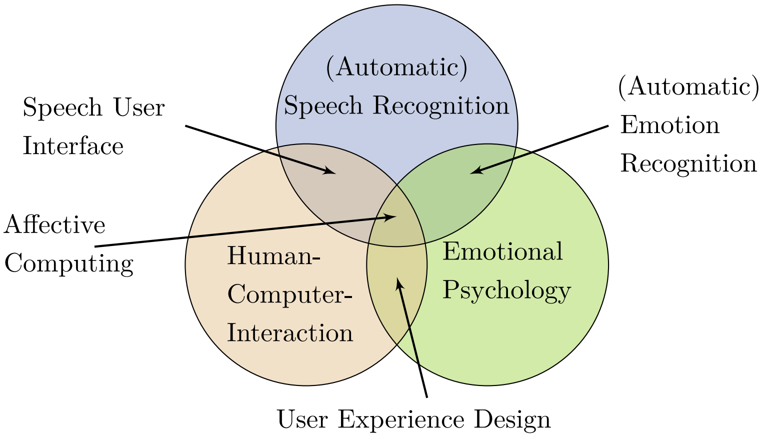


Figure 1.1: Influence of different disciplines on speech-based Affective Computing.

In Figure 1.1, the influence of different disciplines on affective computing are visualised. Affective computing incorporates the research disciplines speech recognition, emotional psychology and HCI[3]. Additionally there are also overlaps between pairs of these disciplines that are related to affective computing. For instance, (speech based) emotion recognition research is influenced by speech recognition and emotional psychology, speech user interfaces are a combination of HCI research and ASR research. The discipline combining HCI and emotional psychology deals with user experience.

Wilks was envisioned machines equipped with affective computing to become conversational systems for which he introduced the term “companion”.

> whose function will be to get to know their owners [..] and focusing not only on assistance [..] but also on providing company and Companionship [..] by offering aspects of personalization [Wilks 2005].

This kind of HCI-system needs more methods of understanding and intelligence than actually present (cf. [Levy et al. 1997; Wilks 2005]), to be able to adjust onto a user.

A DFG-founded research programme contributing to this aim was started in 2009, the SFB/TRR 62 “A Companion-Technology for Cognitive Technical Systems”, under which this work originated. The vision of this programme is to explore methods allowing technical systems to provide completely individual functionality, adapted to each user. Technical systems should adapt themselves to the user’s abilities, preferences, requirements, and current needs. These technical systems are called “Companion Systems” (cf. [Wendemuth & Biundo 2012]). Furthermore, a Companion System reflects the user’s current situation and emotional state. It is always available, cooperative,

[3] Although, in Affective Computing several input modalities are considered, for instance facial recognition, this thesis will only regard the speech channel.

trustworthy, and interacts with its users as a competent and cooperative partner. As main research task, future systems have to recognise automatically the user's emotional state. This task should be considered in parts in this thesis.

1.2 Emotion Recognition from Speech

In order to enable technical systems to recognise emotional states automatically, these systems have to measure input signals, extract emotional characteristics, and assign them to appropriate categories. This approach, known from pattern recognition, has been widely used since the 1980s in computer science. Pattern recognition is successfully applied for instance image processing, speech processing, or computer-aided medical diagnostics [Jähne 1995; Anusuya & Katti 2009; Wolff 2006].

Within pattern recognition, the community distinguishes between two types of learning, supervised and unsupervised learning. Supervised learning estimates an unknown mapping from given samples. Classification and regression tasks are common examples of this learning technique. In unsupervised learning the training data is not labelled and the algorithms are required to discover the hidden structure within the data. This learning technique is mostly used to cluster data or perform a dimension reduction. In my thesis, I concentrate on supervised learning approaches.

A necessary step for the supervised pattern recognition is to model the assignment of objects to categories. For this, two approaches are distinguished: the syntactic and the statistical approach. The syntactic approach (cf. [Fu 1982]) is the more traditional one. It models the assignment by sequences of symbols grouped together with objects of the same category by defining an interrelationship. Furthermore, it earns a hierarchical perspective where a complex pattern is composed from simpler primitives. Using specific knowledge of, for instance, the structure of the face to locate the mouth and eye regions, and afterwards applying different emotional classifiers which are finally combined, the recognition problem can be simplified [Felzenszwalb & Huttenlocher 2005]. The syntactic approach is most promising for problems, having a definite structure that can be captured by a set of rules [Fu 1982].

The statistical approach is currently most widely used [Jain et al. 2000]. Each object is represented by n measurements (data samples) and constituted as a cloud of points in a d-dimensional space covering the values of all measurements. These values are called features as they represent meaningful characteristics of the actual pattern recognition problem. The aim is to group these features into different categories, also known as clusters, by forming compact and disjoint regions. To separate the different categories, decision boundaries have to be established. In the statistical approach, the distinction of categories is modelled by probability distributions, whose parameters have to be learned [Bishop 2011]. An advantage of this approach is that no deeper knowledge of the underlying process generating the data samples is needed. The general process of a supervised pattern recognition is depicted in Figure 1.2.

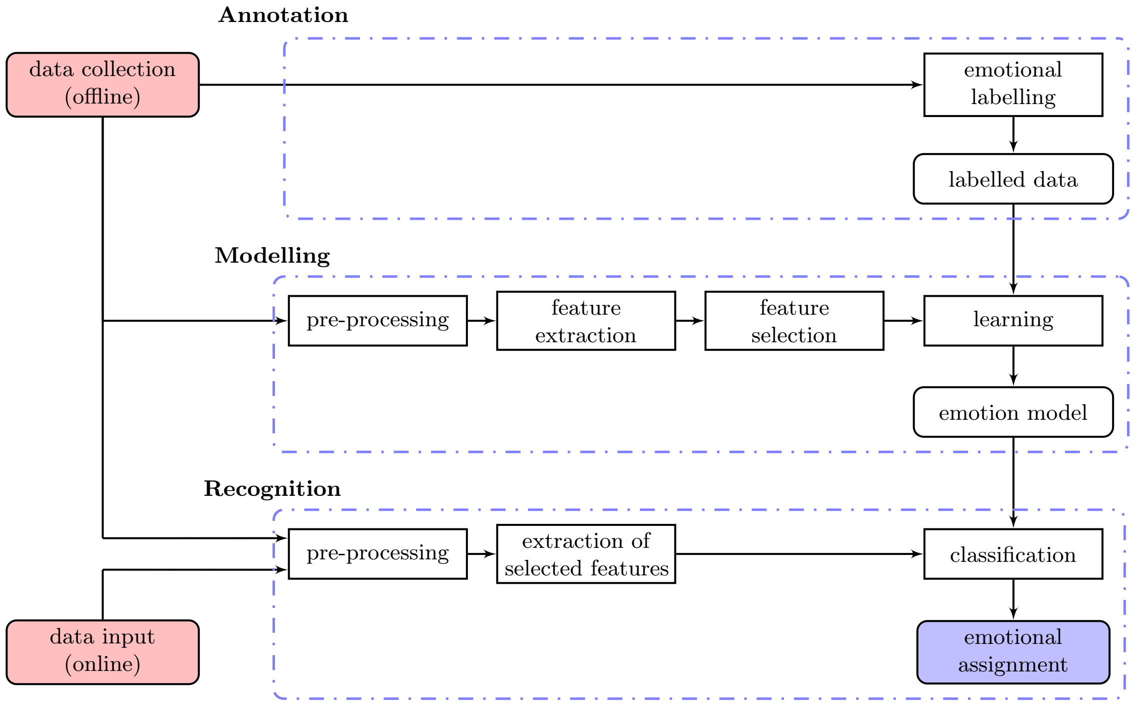


Figure 1.2: Overall scheme of a supervised pattern recognition system.

Three major parts are necessary to successfully develop a system that is capable of recognising an emotion: Annotation, modelling, and recognition. Within the annotation an emotional assignment, called label, is performed between a sample of the training material and an emotional category. For most applications of pattern recognition this task is quite easy. Data collected for specific objective phenomena can be used and categorised accordingly, for instance, recorded speech and its literal transcription. For affect or emotion recognition, this task could be quite challenging. The appearing classes are not as obvious and depending on their context (cf. Section 4.1). At first, the determining characteristics, called features, are extracted and pre-processed. Within the modelling part, a classifier is trained to automatically assign the labels to the collected data. Finally, in the recognition part, unknown or unseen data is processed by the classifier to obtain an emotional assignment.

Furthermore, additional steps are performed during modelling and recognition to enhance the classification performance. Pre-processing is used to remove or reduce unwanted and irrelevant signal components. This could include, for instance, a channel compensation. Afterwards, important characteristics are extracted automatically applying various signal processing methods. It is dependent on the particular application, whose features are essential. Spectral and prosodical features are mostly used for acoustic affect recognition. Furthermore, temporal information or higher order statistical context is also added to infer information about the temporal evolution of the affect (cf. Section 4.2). This can result in a very huge number of features. So far, a proper set of features for emotion recognition from speech covering all aspects is still missing and thus a whole bunch of features are used. By using an optional feature selection process, the huge set of features is reduced by eliminating less promising ones. Either an analysis of variance or a Principal Component Analysis (PCA) is utilised for this. The first approach tests, if one or more features have a good separation capability. The second one uses a space transformation to achieve a good representation of features and decide about a possible reduction of dimensionality.

The recognition can be pursued on material collected offline that has not been used for training to perform a classifier evaluation. This collection is called a “dataset” or “corpus”, including the assignment of labels. The trained classifier can also be applied to live data, this method of operation is called “online classification”.

1.3 Thesis structure

After the subject of investigation has been motivated and the general topics have been presented, the remaining parts of this thesis are structured as follows.

Chapter 2 presents the psychological aspects of emotion recognition and discusses the question how emotions can be described and how they can be measured. Additionally, further psychological concepts as moods and personality traits are discussed insofar as they are necessary for the subsequent work.

Chapter 3 reviews the state-of-the-art in emotion recognition from speech. Starting with the description of the development of emotional speech corpora, naturalistic affect databases as the recent object of investigation are introduced. Afterwards important features, classification methods, and evaluation aspects common for emotion recognition are reviewed. The review is followed by giving an overview of achievable classification performances using different datasets, features and classifiers. Finally, four open issues are identified, which will be pursued during this thesis.

Chapter 4 presents the various methods utilised in this thesis. First the emotional annotation methods are introduced. In this context the kappa-statistic as an important reliability measure for annotation is presented. Subsequently necessary acoustic features and their extraction are described, while distinguishing short-term segmental and longer-term supra-segmental features. Furthermore, their connection to emotional characteristics and further influences, such as ageing, are depicted. Then speech-based emotion recognition techniques, parameters and their optimisation are introduced. An outlook on concepts of classifier combination techniques is also given. This chapter is closed with a description of classifier validation and performance measures as well as statistical significance measures.

Chapter 5 presents the datasets used in this thesis in more detail. Here, one dataset of simulated affects and three datasets of naturalistic affects are distinguished. This illustrates the direction of this thesis by leaving simulated affects and turning towards naturalistic interactions with all their facets and problems.

Chapter 6 describes the author's own work and addresses the first two open issues. A toolkit for emotional labelling is described, followed by methodological improvements to find a reliable ground truth of emotional labels. Afterwards, the second open issue is addressed, by using a speaker group dependent modelling to utilise information about the speaker's age and gender for the improvement of speech-based emotion recognition. This method is applied to various emotional speech databases. Additionally, this method is applied within recent multimodal emotion recognition systems, to investigate the expectable performance gain.

Chapter 7 also addresses the author's own work and describes a new type of interaction pattern, whose usefulness for emotion recognition within naturalistic interactions is investigated. Its ability to indicate situations of higher cognitive load is shown especially. First experiments to automatically classify different types of this interaction pattern are presented. Furthermore, the influence of different user characteristics such as age, gender and personality traits is analysed.

Chapter 8 describes a further aspect needed to analyse naturalistic interactions investigated by the author. The presented mood modelling aims to allow the system to make a prediction on the longer-term affective development of its human conversational partner. The underlying techniques with an additional included personality factor as well as experimental model evaluations are presented and discussed.

In order to allow a strict separation of the authors own contribution, Chapters 1 to 5 introduce the requirements for this thesis with corresponding foreign authors given. The authors own work are discussed separately in the Chapters 6 to 8.

Finally, in Chapter 9 the presented work is concluded and the direction for future research is indicated.

Chapter 2

Measurability of Affects from a Psychological Perspective

Contents

IN the previous chapter, I introduced the aim of this thesis. The main concepts and definitions were given and discussed. The problem of the affect recognition could be traced back to a pattern recognition problem (cf. Section 1.2).

As a requirement for successfully training a pattern recognition system, the observed phenomena have to be entitled and measurable characteristics to distinguish the different phenomena which have to be found. For a technical implementation on recognising emotions, it is important to first review important impacts given by psychological research on emotions and its answers on what emotions are, how emotions can be described, and how they become manifest in measurable characteristics. In order to meet the variety of different theories of emotion, different approaches that deal with the emotion detection and identification were followed. The validity and reliability of these approaches depend very much on the employed theory and method. In the following, I will therefore depict only some of these theories that are of importance for the engineering perspective on emotion recognition.

First, I review common theories on the representation of emotions (cf. Section 2.1). This will be important later, when discussing different emotional labelling methods (cf. Section 4.1) and presenting my own research on methodological improvements of the annotation process (cf. Section 6.1).

Afterwards, I describe the problem of measuring emotional experiences (cf. Section 2.2). There, I depict the appraisal theory, which gives an explanation on the subjectiveness of the verbalisation. The correct verbalisation of emotional experiences is a quite error-prone subjective task. Furthermore, the appraisal theory makes predictions on bodily response patterns, showing that emotional reactions can be characterised by measurable features.

In the last section of this chapter, I briefly describe psychological insights on moods and personality as two longer lasting traits (cf. Section 2.3). These concepts are later important, for analysing the individuality in the HCI (cf. Chapter 7).

2.1 Representation of Emotions

Psychologists' research on emotions has sought to determine the nature of emotions for a long time, starting from the description of emotions either in a categorial (cf. [McDougall 1908]) or dimensional way (cf. [Wundt 1919]) to the finding of universal emotions [Plutchik 1980; Ekman 2005], up to formulating emotional components [Schlosberg 1954; Scherer et al. 2006]. Emotions can generally be illustrated either in a categorial or a dimensional way.

2.1.1 Categorial Representation

At the beginning of the 20th century, McDougall introduced the concept of primary emotions as psychologically primitive building blocks (cf. [McDougall 1908]), allowing to assemble these primitives into "non-basic" mixed or blended emotions. The functional behaviour patterns are named with descriptive labels, such as `anger` or `fear`. Ekman extended this concept by investigating emotions that are expressed and recognised with similar facial expressions in all cultures (cf. [Ekman 2005]). These basic emotions are also called primary or fundamental emotions, whereas non-basic emotions are referred to as secondary ones.

Ortony & Turner established two concepts of grouping emotions into basic and non-basic emotions: a biological primitiveness concept and a psychological one. The first concepts emphasises the evolutionary origin of basic or primitive emotions, the second one describes them as irreducible components (cf. [Ortony & Turner 1990]). The concepts of McDougall and Ekman are examples of the first concept.

Another description made by Plutchik arranges eight basic emotional categories into a three-dimensional space to allow a structured representation of emotions (cf. [Plutchik 1980]). Plutchik makes use of the second concept by Ortony & Turner, which is also called the "palette theory of emotions" (cf. [Scherer 1984]). This representation of emotions is comparable to a set of basic colors with a specific relation used to generate secondary emotions. Regarding Plutchik's emotional model, this makes it possible to infer the effects of bipolarity, similarity, and intensity (cf. Figure 2.1). It is still an open question, which emotions are single categories or components of "emotion families" and also which categories should be taken into account for HCI [Ververidis & Kotropoulos 2006; Schuller & Batliner 2013].

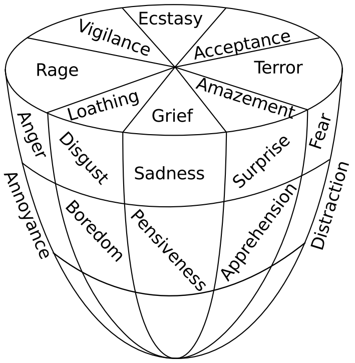


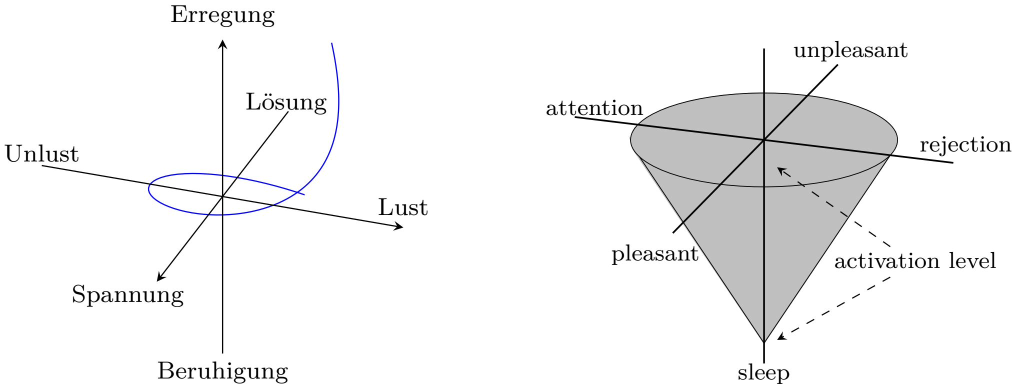
Figure 2.1: Plutchik's structural model of emotions (after [Plutchik 1991], p.157).

One disadvantage of the categorial theories presented here is the estimation of relationships between emotions. The similarity of emotions depends on the utilised type of measure: facial expression, subjective feeling, or perceived emotion. Furthermore, the concept of mixed emotions introduced by Plutchik leads to the problem that a uniform and especially distinctive naming of these new emotions is very difficult:

> That it is not always easy to name all the combinations of emotions may be due to one or more reasons: perhaps our language does not contain emotion words for certain combinations [..] or certain combinations may not occur at all in human experience, [..] or perhaps the intensity differences involved in the combinations mislead us. [Plutchik 1980]

2.1.2 Dimensional Representation

Another approach was made by Wundt, who found McDougalls concept of primary emotions misleading. He introduced a so-called "total-feeling" representing a mixture of potentially conflicting, elementary feelings consisting of a certain quality and intensity. The elementary feelings are constituted by a single point in a three dimensional emotion space with the axes "Lust" (`pleasure`) ↔ "Unlust" (`unpleasure`), "Erregung" (`excitement`) ↔ "Beruhigung" (`inhibition`), and "Spannung" (`tension`) ↔ "Lösung" (`relaxation`) (cf. Figure 2.2(a)). In Wundt's understanding an external event results in a specific continuous movement in this space, described by a trajectory. This theory provided for the first time a clear explanation for the transition of emotions. Furthermore, Wundt was able to verify a relation between pleasure and unpleasure and respiration or pulse changes (cf. [Wundt 1919]). An additional advantage of Wundt's approach is that emotions may be described independently of categories and that emotional transitions are inherent for this model. Unfortunately, this theory does not locate single emotions into the emotional space and does not explain how intensity could be integrated or determined given a distinct perception.



(a) Emotion space with emotion trajectory (-) (after [Wundt 1919], p.246).

(b) Schlosberg's conic emotion diagram (after [Schlosberg 1954], p. 87).

Figure 2.2: Representations of dimensional emotion theories.

Wundt's concept can be seen as a starting point for later research on dimensional emotion concepts, whereas later research groups deal with the exact configuration and number of the dimension axes. Schlosberg examined the activation axes (comparable to excitement ↔ inhibition), on the basis of emotional picture ratings (cf. [Schlosberg 1954]). He uses the dimensions pleasantness ↔ unpleasantness, attention ↔ rejection, and activation level. Schlosbergs activation can be identified as an intensity similar to Plutchik (cf. Figure 2.2(b)).

Particularly, the question of a need for a third dimension and their description is subject of ongoing discussions. In this way, Russel argued against the necessity of intensity as a third dimension. By a further investigation, Mehrabian & Russell could emphasise the fact that another dimension is needed to distinguish certain emotional states. They examined differences between anger and anxiety, by presenting emotional terms arguing for the need of a third dimension, they called dominance (cf. [Mehrabian & Russell 1977]). In this study they also presented the localization of 151 English emotional terms into their so-called Pleasure-Arousal-Dominance (PAD)-space. In a comprehensive study by [Gehm & Scherer 1988], using German words describing emotions, the findings by Russel and Mehrabian could not be replicated. Moreover, they found that `pleasure` and `dominance` are the dimensions having the most discriminating power to distinguish emotional terms. Gehm & Scherer criticise that Mehrabian & Russell did not take into account the underlying process of the subjects' ability to rate the emotionally relevant adjectives or pictures [Scherer et al. 2006]. This could be one indicator for the difference in the selected axes.

Becker-Asano summarised the discussions surrounding different dimensions and presented an overview of utilised components that can be condensed (cf. [Becker-Asano 2008]): The most important component is called either `pleasure`, `valence`, or `evaluation`. The valence of an emotion is always either positive or negative. The second component is mostly regarded as the `activation`, `arousal`, or `excitement` dimension. It determines the level of psychological arousal or neurological activation (cf. [Becker-Asano 2008]). For some researchers (cf. [Mehrabian & Russell 1977]) no further dimension is needed. But the works of [Schlosberg 1954] and [Scherer et al. 2006] highlight the need for a third dimension. Especially for cases of both, `high pleasure` and `high activation`, the incorporation of a third dimension indicating `dominance`, `control`, or `social power` is useful to distinguish certain emotions.

The reviewed research on emotional representation illustrates the dilemma the affective computing community has to fight with. Unfortunately, there are many concurrent emotional representations. When it comes to name individual reactions, reference is made to categories. But they are depending on the chosen setting and investigated question. There is agreement only on a small number of "basic emotions": `anger`, `disgust`, `fear`, `happiness/joy`, `sadness`, and `surprise`. They are used in most categorial systems (cf. [Ekman 2005; Plutchik 1980]). In addition, there are usually more categories, but there is no consensus on them (cf. [Mauss & Robinson 2009]). The emotion recognition community choose depending on the investigation several additional categories, mostly arbitrarily seeming ones. Therefore, a comparison of results is difficult and a rather artificial merging of emotional labels is needed, if results are compared across different corpora (cf. [Schuller et al. 2009a]).

If the variability of emotions is in the foreground, the dimensional approach is rather preferable. The emotion is presented as a point in a (multi-)dimensional space. This perspective argues that emotional states are organised by underlying factors such as `valence` and `arousal`. However, type and exact number of dimensions is still a subject of research. It is agreed that `valence` is the most important dimension, but whether `arousal` and/or `dominance` are further needed has not been definitively resolved (cf. [Mehrabian & Russell 1977; Scherer et al. 2006]). Of special appeal for the affective computing community is the PAD-space, as it allows to distinguish many different emotional states. Additionally, dimensional and discrete perspectives can be reconciled to some extent by conceptualising discrete emotions in terms of combinations of multiple dimensions (e.g., `anger` = `negative valence`, `high arousal`) that appear discrete because they are salient (cf. [Mauss & Robinson 2009]).

2.2 Measuring Emotions

In the previous section, I presented concepts to distinct emotions, but moved over to the question of their origin. In psychological research it is common sense that emotions reflect short-term states, usually bound to a specific event, action, or object (cf. [Becker 2001]). Hence, an observed emotion reflects a distinct user assessment related to a specific experience. The appraisal theory (cf. [Scherer 2001]) now states that emotions are the result of the evaluation of events causing specific reactions.

In appraisal theory, it is supposed that the subjective significance of an event is evaluated against a number of variables. The important aspect of the appraisal theory is that it takes into account individual variances of emotional reactions to the same event. Thus, according to this theory, an emotional reaction is occurring after the interpretation and explanation of such an event. This results in the following sequence: event, evaluation, emotional body reaction (cf. Figure 2.3). The body reactions are than resulting in specific emotions. The appraisal theory is quite interesting for the process of automatic emotion recognition as it also defines specific bodily response patterns for appraisal evaluations [Scherer 2001].

One appraisal theory model that is considered in this thesis is the Multi-level Sequential Check Model by Scherer (cf. [Scherer 1984]). It helps to explain the underlying process between appraisals and the elicited emotions and captures the dynamics of emotions by integrating a dynamic component.

The basic principles of the Multi-level Sequential Check Model were proposed by Scherer in 1984, focussing on the underlying appraisal processes in humans (cf. [Scherer 1984]). The proposed model explains the differentiation of organic subsystem responses. Therefore, it includes a specific sequence of evaluation checks, which allow to observe the stimuli at different points in the process sequence: 1) a relevance check (novelty and relevance to goals), 2) an implication check (cause, goal conduciveness, and urgency), 3) coping potential check (control and power), and 4) a check for normative significance (compatibility with one's standards). Each check uses other appraisal variables, for instance relevance tests for `novelty` and `intrinsic pleasantness` whereas implication checks for `causality` and `urgency`. This results in a sequence of specific event evaluation checks (appraisals), where the organic subsystems NES, SNS, and ANS are synchronised. These subsystems manifest themselves in response patterns, which can be described using emotional labels. Furthermore, during event evaluation cognitive structures are involved and considered respectively (cf. Figure 2.3). This model encouraged several theoretical extensions over the past decades (cf. [Marsella & Gratch 2009; Smith 1989; Scherer 2001]).

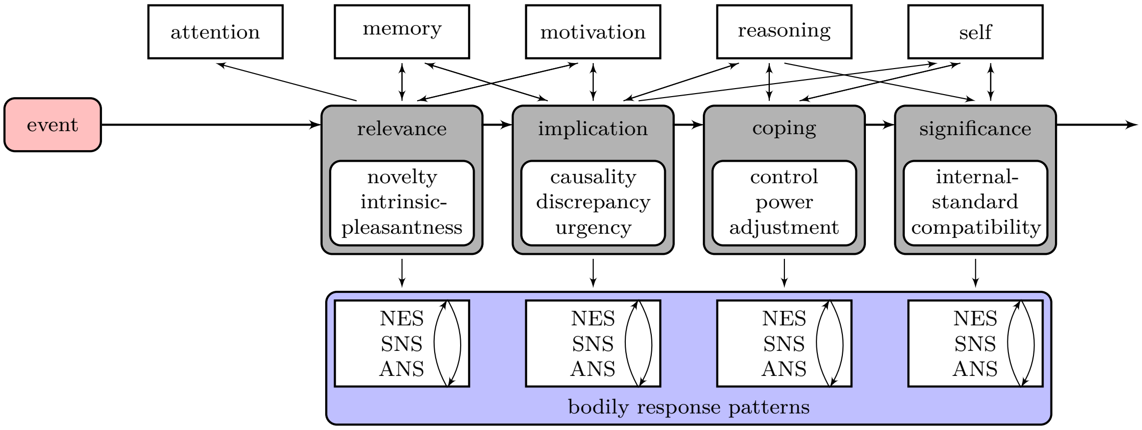


Figure 2.3: Scherer's Multi-level Sequential Check Model, with associated cognitive structures, example appraisal variables and peripheral systems (after [Scherer 2001], p. 100).

According to Scherer, verbal labels are language-based categories for frequently and universally occurring events and situations, which undergo similar appraisal profiles [Scherer 2005a]. This consideration seems to be connected to Ekman's investigations of basic emotions. They are expressed and recognised universally through similar facial expressions, regardless of cultural, ethical, gender, or age differences. However, the theories have a contrary view on the emotional response: basic emotion theorists assume integrated response patterns for each (basic) emotion, while appraisal theorists believe that the response pattern is a result of the appraisal process, which than is observed as a specific emotional reaction. Both theories predict that there is a similarity between response patterns and emotions, only the temporal order of that similarity is object of an ongoing debate [Scherer 2005a; Colombetti 2009].

2.2.1 Emotional Verbalisation

Another impact that Scherer's appraisal model implies is the problem of the verbalisation and the communicative ability of emotional experiences. In his understanding, the changes in the emotion components can be divided into three modes: unconsciousness, consciousness, and verbalisation. To illustrate the notion, Scherer uses a Venn diagram to show the possible relation between the modes (cf. Figure 2.4).

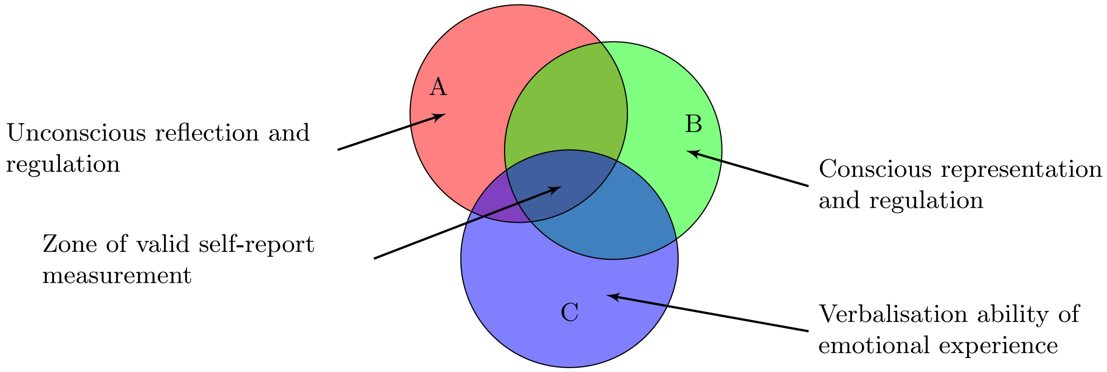

A
B
C
Unconscious reflection and regulation
Conscious representation and regulation
Zone of valid self-report measurement
Verbalisation ability of emotional experience

Figure 2.4: Scherer's three modes of representation of changes in emotion components (after [Scherer 2005a], p. 322).

In this figure, circle (A) represents the raw reflection in all synchronised components. These processes are unconscious but of central importance for response preparation. Scherer called the content of this circle "integrated process representation". The second circle (B) becomes relevant, when the "integrated process representation" becomes conscious. Furthermore, it represents the quality and intensity generated by the triggering event [Scherer 2005a]. The third circle (C) reflects processes enabling a subject to verbally report its subjective emotional experience. Scherer notes by the incomplete overlap it should be pointed out that we can verbalise only a small part of our conscious experience. He gives two reasons for this assumption. First, the lack of appropriate verbal categories and second, the intention of a subject to control or hide specific feelings [Scherer 2005a]. This problem of valid and comprehensible emotional labelling is reviewed in Section 4.1 and further investigated in Section 6.1.

2.2.2 Emotional Response Patterns

The appraisal theory also tries to make predictions of *bodily response patterns* (cf. [Scherer 2001]). They describe measurable changes in the nervous systems and derived changes in face, voice, and body, which then can be observed by the sensors of a technical system. Appraisal theorists argue that only if conscious schemata for the type of event are established, the various nervous systems' processing modules (memory, motivation, hierarchy, reasoning) are involved [Scherer 2001]. Furthermore, different action tendencies are invoked, which will activate parts of the Neuro-Endocrine System (NES), Autonomic Nervous System (ANS), and Somatic Nervous System (SNS). The general assumption is that different organic subsystems are highly interdependent. Changes in one subsystem will affect others. These changes will even affect observable responses in voice, face, and body.

Examining facial expressions to be able to indicate the underlying unobservable emotional processes has a long tradition in emotion psychology. In quantified facial behaviour using componential coding, trained coders detect facial muscle movements or Action Units (AUs) using reliable scoring protocols (cf. [Mauss & Robinson 2009]). The most widely used componential coding system is the Facial Action Coding System (FACS). FACS is an anatomically based, comprehensive measurement system that assesses 44 different muscle movements (e.g., raising of the brows, tightening of the lips) (cf. [Ekman & Friesen 1978]). As such, it measures all possible combinations of movements that are observable in the face rather than just movements that have been theoretically postulated. Researchers were able to define prototypical patterns (AUs) for some basic emotions in still images [Ekman 2005], but they are not able to interpret the various occurring facial expressions in spontaneous interactions.

In contrast, several appraisal theories suggest to link specific appraisal dimensions with certain facial actions (cf. [Frijda 1969; Smith 1989]). Thus, facial expressions are not seen as a “direct readout” of emotions but as indicators of the underlying mental states and evaluations. A study taking into account the temporal structure is presented in [Kaiser & Wehrle 2001]. By analysing facial expressions, some “pure” appraisal and reappraisal processes could be specified, but the full variety of observed facial expressions is still not described.

Another very prominent response pattern are vocal characteristics. Scientific studies have examined them most commonly by decomposing the acoustic waveform of speech and afterwards, assessing whether such acoustic properties are associated with the emotional state of the speaker. In [Johnstone et al. 2001], the authors notice that only limited research has been done in terms of acoustic emotional patterns. Studies typically investigated acted basic emotions or real, but bipolar inductions (e.g. `low` vs. `high stress`). Also other studies could define only in a few cases a regularity between acoustic characteristics and emotions. They mostly refer to investigations found in [Johnstone et al. 2001; Scherer et al. 1991].

Most acoustic research reported correlations between `arousal` and pitch: higher levels of `arousal` have been linked to higher-pitched vocal samples (cf. [Bachorowski 1999; Mauss & Robinson 2009]). Only a few studies could differentiate emotional responses on the `valence` or `dominance` dimension (cf. [Frijda 1986; Iwarsson & Sundberg 1998]). A broad study conducted by Banse & Scherer, measured differences in acoustic changes of 14 emotions, including various intensities of the same emotion family (e.g., `cold anger`, `hot anger`) and 29 acoustic variables (cf. [Banse & Scherer 1996]). Therefore, twelve actors were provided with emotion-eliciting scenarios. To avoid influences of different phonetic structures, two meaningless German utterances are used. For analysis the acoustic parameters fundamental frequency, intensity, and speech rate were analysed. The authors found that a combination of ten acoustic properties distinguish discrete emotions to a greater extent than could be attributed to `valence` and `arousal` alone. However, these links were complex and multivariate in nature, involving post hoc comparisons. Furthermore, the acoustic characteristics are described mostly in a qualitative manner, such as medium low frequency energy or increase of pitch over time (cf. Table 2.1). Unfortunately, the predictions made by the utilised appraisal theory and the actual measured acoustics differed remarkable. It should be noted that work of this complex type is just beginning and much remains to be learned (cf. [Juslin & Scherer 2005; Mauss & Robinson 2009]).

Table 2.1: Vocal emotion characteristics (after [Scherer et al. 1991], p. 136). The symbol + denotes increase, − denotes decrease compared to neutral utterances; double symbols of the same type indicate the strength of the change, double symbols with opposite direction signs indicate that both increase and decrease are possible.

	Fear	Joy	Sadness	Anger
Fundamental frequency	++	+	+−	+−
Fundamental frequency variance	+	+	−	+
Intensity mean	+	+	−−	+
Intensity variance	+	+	−	+
High-frequency energy	++	+−	+−	++
Speech rate	++	+	−	+

2.3 Mood and Personality Traits

As already stated in the previous section, emotions reflect short-term states, usually bound to a specific event, action, or object (cf. [Becker 2001]). Hence, an observed emotion reflects a distinct user assessment that is related to a specific experience.

In contrast to emotions, *moods* reflect medium-term states, generally not related to a concrete event (cf. [Morris 1989]), but a subject of certain situational fluctuations that can be caused by emotional experiences [Morris 1989; Nolen-Hoeksema et al. 2009]. In general, moods are distinguished by their positive or negative value. As moods are not directly caused by intentional objects, they do not have a specific start or end point. Thus, moods can last for hours, days, or weeks and are more stable states than emotions. In [Mehrabian 1996] the PAD-space octands are used to distinct specific mood categories.

Moods influence the user's cognitive functions directly as they can manipulate how subjects interpret and perceive their actual situation, which influences their behaviour and judgements. A study conducted by Niedenthal et al. reveals that subjects tend to assess things as `negative`, when a negative mood is induced (cf. [Niedenthal et al. 1997]). The mood's motivational influence is especially important in HCI. Several studies found that a positive mood enhances the individual (creative) problem solving ability and expands the attention so that relevant information becomes more accessible (cf. [Rowe et al. 2007; Nadler et al. 2010; Carpenter et al. 2013]). The measuring of the users' mood has to rely on self-reports, for instance the Positive and Negative Affect Schedule (PANAS) (cf. [Watson et al. 1988]).

Personality reflects long-term states and individual differences in mental characteristics. According to [Nolen-Hoeksema et al. 2009], personality comprises

> [..] distinctive and characteristic patterns of thought, emotion, and behaviour that make up an individual's personal style of interacting with the physical and social environment.

To this end, in the beginning of personality research mostly describing adjectives were used, to describe different personality traits (cf. [Allport & Odbert 1936; Nolen-Hoeksema et al. 2009]). Today, it is agreed that personality is a rather complex entity containing different aspects. Personality traits are important for the user's behaviour in both HHI and HCI [Daily 2002]. Certain personality traits such as optimism and neuroticism could also influence certain types of moods [Morris 1989].

To identify the specific personality traits of a user, mostly questionnaires are used. There are several questionnaires used depending on which trait should be analysed. For HCI the subject's affinity for technology can be additionally measured by AttrakDiff (cf. [Hassenzahl et al. 2003]) or TA-EG (cf. [Bruder et al. 2009]). In the following, I will only present one most important questionnaires used for the present work.

A common model to characterise personality is the "Five Factor Model". The "Five Factor Model" describe five broad dimensions of personality. Initial work on the theory of the Five Factor Model has been published by Allport & Odbert, who used a lexical approach to find essential personality traits through language terms (cf. [Allport & Odbert 1936]). Through factor analysis Goldberg could identify five very strong, independent factors, the "Big Five" [Goldberg 1981]. Based on these findings, Costa & McCrae developed the NEO[4] five-factor personality inventory (cf. [Costa & McCrae 1992]), widely used today. The questionnaire uses 60 items on a five point Likert scale, capturing the five personality dimensions. The NEO-FFI focus on the "general

[4] In a former version of their personality inventory Costa & McCrae only considered the three factors `neuroticism`, `extraversion`, and `openness` (NEO-I). This inventory was later revised including the presently known five traits and renamed to NEO Personality Inventory (NEO PI), where "NEO" is now considered as part of the name and no longer as acronym (cf. [Costa & McCrae 1995]).

population" in a non-clinical environment (cf. [Nolen-Hoeksema et al. 2009]). The dimensions are described as follows (cf. [Doost et al. 2013]):

Openness for experience Appreciation for art, adventure, and unusual ideas. Openness reflects the degree of intellectual curiosity, creativity, and a preference for novelty and variety. Some disagreement remains about how to interpret the openness factor, which is sometimes called "intellect" rather than openness to experience.

Conscientiousness A tendency to show self-discipline, act dutifully, and aim for achievement; planned rather than spontaneous behaviour; organised and dependable.

Extraversion Energy, positive emotions, assertiveness, sociability and the tendency to seek stimulation in the company of others, and talkativeness.

Agreeableness A tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others.

Neuroticism The tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, or vulnerability. Neuroticism also refers to the degree of emotional stability and impulse control, and is sometimes referred by its low pole – "emotional stability".

Nowadays, the "Big Five" are widely confirmed and represent the most influential personality model (cf. [John et al. 1991; Ozer & Benet-Martinez 2006]).

In contrast to the "Big Five", other theories of personality focuses more on interpersonal relationships, like the inter-psychic model of personality as proposed by [Sullivan 1953] focussing on interpersonal relationships, covered by the Inventory of Interpersonal Problems (IIP) [Horowitz et al. 2000]. It models conceptualising, organising, and assessing interpersonal behaviour, traits, and motives. Eight scales mark the interpersonal circumplex by selecting items (`domineering`, `vindictive`, `cold`, `socially avoidant`, `nonassertive`, `exploitable`, `overly nurturant`, and `intrusive`). The questionnaire uses 64 items on a five point Likert scale.

A questionnaire, dealing with the stress-coping ability is the Stressverarbeitungsfragebogen (stress-coping questionnaire) (SVF) [Jahnke et al. 2002]. It includes 20 scales, for instance `deviation`, `self-affirmation`, or `control of reaction`, for different types of responses to an unspecific selection of situations that could impair, adversely affect, irritate, or disturb the emotional stability or balance of the subject.

In HCI, personality plays an important role as well (cf. [Cuperman & Ickes 2009; Funder & Sneed 1993]). For instance, Weinberg identified personality traits as well as interpersonal relationship as relevant aspects in the field of HCI (cf. [Weinberg 1971]). Research studies that investigated `extraversion` as a personality trait discovered that people with `high extraversion` values are more satisfied and emotional stable

[Pavot et al. 1990]. According to Larsen & Ketelaar, extroverted persons responded more strongly to positive emotions than to negative emotions during an emotion induction experiment (cf. [Larsen & Ketelaar 1991]). Furthermore, Tamir claims that extroverted persons regulate their emotions more efficiently, showing a slower decrease of positive emotions (cf. [Tamir 2009]). Summarising, there is some evidence suggesting that `extraversion` is related to computer aptitude and achievement [Veer et al. 1985]. Furthermore, many user characteristics are discussed having an influence on the interaction towards technical systems, for instance attributional style, anxiety, problem solving and stress coping abilities.

2.4 Summary

The presented studies show that psychological research has already covered substantial ground in the field of emotion research. Although most studies investigate specific areas as medical diseases, social interaction, or the behaviour after strong experiences, they deliver a quite adequate understanding of how to categorise and describe emotional observations. Important for automatic emotion recognition is the consensus that emotions are reflected in (external) observable phenomena. These can be measured in both facial expressions and acoustic characteristics.

Ekman's investigation provides a very accurate description, how certain basic emotions are reflected in facial expressions. Unfortunately, there are no similar studies for emotional acoustic characteristics, yet. The investigations presented by [Scherer 2001] lead in such direction, but as they stick on the appraisal level, they are not directly usable for automatic emotion recognition from speech.

But the appraisal theory provides another important finding, as it postulates that a reliable self-assessment of emotions is not possible in all cases. This makes it difficult for automatic emotion recognition to generate a valid ground truth, and further methods have to be applied to secure it (cf. Section 4.1).

Two concepts that receive little attention within the HCI research are mood and personality. Although, psychological research revealed a huge influence on the users behaviour, the conceptualisation of both traits does not go beyond a description. A reliable measuring of mood changes or specific personality traits can only be assured by using questionnaires, which is not feasible for technical systems in an actual ongoing interaction. An approach, how a mood-like representation based on emotional observations can be modelled is presented in Chapter 8.

Chapter 3

State-of-the-Art

Contents

IN the previous chapters, I introduced the aim of this thesis. As pointed out in Section 2.2 there is only limited psychological research on acoustic emotional patterns and thus, for automatic emotion recognition new approaches have to be developed, which are mainly related to pattern recognition methods. As from a technical perspective, emotion recognition from speech is very similar to Automatic Speech Recognition (ASR), hence the community uses similar characteristics and classifiers (cf. [Schuller et al. 2011c]). Recently emotion specific techniques, such as gender specific modelling or perceptually more adequate features for emotion recognition are incorporated (cf. [Dobrišek et al. 2013; Cullen & Harte 2012]).

In the following, the state-of-the-art in emotion recognition from speech will be presented. Research activities in this field are heavily related to the field of affective computing, arising in the mid 1990s by the book “Affective Computing” by Rosalind Picard. In this book Picard states “many computers will need the ability to recognise human emotion.” As speech is the most natural way to interact, it seems adequate to analyse emotions expressed in it. Automatic emotion recognition from speech is a quite emerging field of research starting with spurious papers in the late 1990s and getting a growing interest since 2004 (cf. [Schuller et al. 2011c]).

Initially, I review the development of sets of emotional data, highlighting the recent change from acted emotions towards naturalistic interactions (cf. Section 3.1). Afterwards, I focus on important trends for emotion recognition from speech by reviewing specific developments in this area necessary to position my own research. This includes utilised features and pre-processing steps, applied classifiers as well as methods in evaluating the results (cf. Section 3.2). Additionally, I review the development of achieved recognition results on different types of emotional material, which are later used for my own research (cf. Section 3.3). This chapter is completed by emphasising certain open issues, which are under-represented in the actual research-discussions (cf. Section 3.4). These topics serve as framework for my own research.

It is obvious that this chapter can only be a spotlight of the research activities due to the enormous amount of work done in this field. Especially the speech recognition community has already been established for a long time. The same holds true in general for pattern recognition. Thus, I only discuss issues where affective speech recognition and pattern recognition fields overlap and that have a strong relation to my own research and I am aware that it is not possible to cover all aspects of the community within this thesis.

3.1 Reviewing the Evolution of Datasets for Emotion Recognition

Classifiers trained on emotional material are needed to efficiate automatic emotion recognition. This material is usually denoted as “dataset” or “corpus”. A dataset commonly focuses a certain set of emotions or emotional situations. Furthermore, it covers a specific language within a certain domain. To train optimal classifiers, it would be desirable if the dataset is quite large and covers a broad set of emotions within a widely valid domain. Furthermore, it should contain additional information about the users (age, gender, dialect, or personality) and should be available to and accepted by the research community. As one can easily imagine, there is no single corpus that meets all of these requirements.

The demand on high quality emotional material was raised at times where no corpus meeting these requirements was available. Thus, the research community started with small mostly self-recorded datasets containing acted non-interactional emotions. Later the community switched to larger datasets of induced emotional episodes. Afterwards, naturalistic emotions got in the focus. Just recently the community switched to corpora containing longer lasting naturalistic interactions. Several surveys give an overview on the growing number of emotional corpora. [Ververidis & Kotropoulos 2006] as well as [Schuller et al. 2010a] are good sources to compare well-known emotional speech databases. An almost comprehensive list of emotional speech databases for various purposes and languages can be found in the Appendix of [Pittermann et al. 2010]. I distinguish between databases with simulated emotions, containing acted or induced emotions, and databases with naturalistic emotions, containing spontaneous or naturalistic emotional interactions. In this section, I give a general non-exhaustive overview about the various available databases used for emotion recognition. An in-depth description about the the corpora used in this thesis is given in Chapter 5.

3.1.1 Databases with Simulated Emotions

The research on emotion recognition started in the late 1990s, where no common emotional corpora were available. Thus, based on the established methods for speech recognition database generation, emotional material was generated. The first available corpora on emotional speech were quite small and consisted only of a few subjects, not more than ten. These corpora have a quite high recording quality, since they are recorded in studios. This also shows their direct relationship to speech corpora.

As the spoken content was mostly pre-defined, consisting of single utterances, these corpora contain acted emotions without any interaction. Due to its pre-defined characteristic, the emotional content was evaluated via perception tests to secure the observability of emotions and their naturalness [Burkhardt et al. 2005]. These corpora were rarely made publicly available. Additionally, some of these databases were not originally recorded to serve as material for emotion recognition classifiers. Instead, their purpose was to serve as quality assessment of speech synthesis. Well-known representatives are the Berlin Database of Emotional Speech (emoDB) [Burkhardt et al. 2005] and the Danish Emotional Speech (DES) database [Engberg & Hansen 1996] using actors to pose specific emotions in German or Danish. These databases mostly have the flaw of very over-expressed emotional statements that are easy to recognise but hardly present in naturalistic scenarios. Furthermore, they only cover a specific range of emotions, mostly comparable with Ekman's set of basic emotions (`anger`, `sadness`, `happiness`, `fear`, `disgust`, `sadness`, and `boredom`) [Ekman 1992]. But recently corpora with a broader set of emotional states were generated, for instance the GEneva Multimodal Emotion Portrayals (GEMEP) corpus, featuring audio-visual recordings of 18 emotional states, with different verbal contents and different modes of expression [Bänziger et al. 2012].

Another approach is to extract emotional parts from movies. This procedure should assure a more naturally expressed emotion. Databases using this method are the Situation Analysis in a Fictional and Emotional Corpus (SAFE) [Clavel et al. 2006] for English or the New Italian Audio and Video Emotional Database (*NIAVE*) [Esposito & Riviello 2010] for Italian. These types of corpora need a more elaborative annotation to segment the video material and annotate these segments afterwards. But using a proper selection of movie material, specific emotions, such as different types of `fear` (SAFE) or `irony` (*NIAVE*) can be collected.

The next step in data collection was the emotional inducement. Emotional stimuli are presented to a subject, whose reactions were recorded. From psychological research it is known that movies, music, or pictures are useful to elicit an emotional reaction. Although these methods are quite common in psychological research (cf. [Pedersen et al. 2008; Forgas 2002]), there are nearly no speech databases available that are generated utilising these methods. Instead, participants are instructed to form images of emotional memories or hypothetical events, of which they afterwards have to talk about or react onto properly. These answers are recorded to build the database. Corpora of this type are the emotional speech corpus on Hebrew [Amir et al. 2000] and the eNTERFACE'05 Audio-Visual Emotion Database (eNTERFACE) [Martin et al. 2006]. The emotional episodes the participants should memorise and react to mostly cover basic emotions comparable to Ekman's set [Ekman 1992]. Another method that is used to generate emotional speech databases is to confront subjects with a task that has to be solved under stress. This method is used in the Speech Under Simulated and Actual Stress Database (SUSAS) [Tolkmitt & Scherer 1986], the Airplane Behavior Corpus (ABC) [Schuller et al. 2007b], or the emotional Speech DataBase (*emoSDB*) (cf. [Fernandez & Picard 2003]).

All of these inducement methods have in common that the emotional content has to be assessed afterwards. For this, a perception test is sufficient to select successful inducements as the intended emotional reaction is pre-defined. An overview of all mentioned databases with simulated affects can be found in Table 3.1 on page 31.

3.1.2 Databases with Naturalistic Affects

The databases presented so far mostly contain single phrases not originated from a longer interaction. Therefore, researchers also used excerpts of human to human interactions, which are expected to contain emotional episodes. Typically, excerpts from TV-shows are used, especially chat shows, for instance the Belfast Naturalistic Database (*BNDB*) [Douglas-Cowie et al. 2000] or the Vera am Mittag Audio-Visual Emotional Corpus (VAM) [Grimm et al. 2008]. Furthermore, reality TV-shows are utilised, for instance the Castaway database [Devillers & Vidrascu 2006]. Another preferred sources are interviews as in the EmoTV corpus [Abrilian et al. 2005]. These databases are similar to databases based on movie excerpts. It is assumed that the occurring emotions are more natural and spontaneous than in databases of acted emotions (cf. [Devillers & Vidrascu 2006; Grimm et al. 2008]).

In order to carry out an emotional assessment, these databases have to be annotated using several annotators, ranging from 2 to 17. Furthermore, different emotional representations are utilised, for instance, emotion categories, dimensional labels, or dimensional emotion traces (cf. Table 3.1 on page 31).

Another method is used for the ISL Meeting Corpus, where three or more individuals are recorded during a meeting with a pre-defined topic. By this procedure an increased expressiveness of the interaction is assured. But the resulting emotional annotation bears no relationship to the effort of generation, as the so far discerned emotions are `positive`, `negative`, and `neutral` (cf. [Burger et al. 2002]).

Another method for collecting emotional speech data is to use recordings from telephone based dialogue systems. These dialogues are mostly from a very specific domain, and the collection is very easy, as for call-center agencies the data-recording is already established. The only difficulty is to generate valid and reliable labels. Thus, the Messages corpus only contains the assessments `positive emotion`, `negative emotion`, and `no emotion` (`neutral`) [Chateau et al. 2004]. Other telephone based corpora put a lot more effort into the annotation, but mostly covering `negative` and `high arousal` emotions. Representatives are the CEMO corpus containing recordings obtained from medical emergency call centers [Devillers & Vidrascu 2006], the Affective Callcenter Corpus (*ACC*) for English [Lee & Narayanan 2005] and the UAH emotional speech corpus (UAH)[5] for Spanish [Callejas & López-Cózar 2008]. In contrast, the "Emotional Enriched LDC CallFriend corpus" (*CallfriendEmo*) [Yu et al. 2004] contains several emotions and uses general telephone conversations. But all of these corpora have the disadvantage that the material is of varying quality and the interaction is totally uncontrolled. Thus, the question whether and which emotions arise cannot be controlled and the material has to be labelled by an extensive number of labellers. Moreover, these databases utilize HHI, but it is not clear and still a matter of research whether the same emotions are occurring within HCI.

To be able to conduct interaction studies under controlled surroundings and to investigate the emotions within HCI, several databases are recorded in a so-called Wizard-of-Oz (WOZ) scenario. In this case, the application is controlled by an invisible human operator, while the subjects believe to talk to a machine. The system can be directly used to frustrate the user and provoke emotional reactions within a game-like setting as used in the NIMITEK Corpus [Gnjatović & Rösner 2008]. The experiment can be focussed on specific emotional situations, for instance the level of interest, as used in the Audivisual Interest Corpus (TUM AVIC) [Schuller et al. 2009b]. It could also be focussing on a pure interaction task as with an imperfect system as in SmartKom (cf. [Wahlster 2006]) or ITSPOKE (cf. [Ai et al. 2006]). These corpora also depend on an exhaustive manual annotation. But due the WOZ-style the expected

[5] UAH is the abbreviation for the spoken dialogue system "University on the Phone" (Universidad Al Habla) developed at the University of Granada (cf. [Callejas & López-Cózar 2005]).

user reactions are in control of the experimentators. A speciality can be seen in the corpus EmoRec [Walter et al. 2011]. Although this corpus uses a WOZ-controlled game scenario to evoke emotional reactions as well, this corpus aims to put the subjects into distinct emotional states within the emotional PAD-space. This is controlled via a calibration phase and bio-sensor monitoring. Thus, an annotation is not needed afterwards. The FAU AIBO corpus (cf. [Batliner et al. 2004]) also uses WOZ-simulated interactions, but with children instead of adults. It records emotional interactions of children playing with a WOZ-controlled robot in English and German. This corpus was labelled during the Combining Efforts for Improving automatic Classification of Emotional user States (CEICES)-initiative (cf. [Batliner et al. 2006]) and contains some basic emotions but also a broad spectrum of secondary emotions, for instance `motherese` or `reprimanding`. The WOZ data corpus (*WOZdc*) collected by Zhang et al. worked with children and recorded emotional episodes within an intelligent tutoring system (cf. [Zhang et al. 2004]). This system was used to analyse the users' reaction on system malfunctions and thus, the emotional annotation also covers these kind of states.

A data corpus emphasising rather the dialogical character, but using a WOZ scenario to induce emotional reactions is the Belfast Sensitive Artificial Listener (SAL) database. This corpus uses interactive characters with different personalities, to induce emotions during an interaction (cf. [McKeown et al. 2010]). Special characteristic of this corpus is its method of emotional annotation. Different emotional dimensions are labelled continuously, to be able to follow the emotional evolvement (cf. Section 4.1.2). A similar corpus emphasising the naturalistic interaction even more is the LAST MINUTE corpus (LMC). It also used a WOZ-scenario, but, instead of inducing emotional reactions, relied more on specific critical dialogue parts, called "barriers", to force the subjects to re-plan their interaction strategy (cf. [Rösner et al. 2012]). This provokes more natural reactions, as the subjects are not forced to show emotional reactions. They could also use verbal markers as swearwords or feedback signals indicating their trouble within the communication (cf. [Prylipko et al. 2014a]).

SAL and LMC represent a "new generation" of corpora, as they place a greater value on the naturalness of the interaction and particularly for LMC, the emotion inducement is relegated to the side. An overview of all mentioned corpora is given in Table 3.1. Some of the presented corpora contain several modalities, audio, video, or even bio-physiological data. As the focus of this thesis is on acoustical information, these additional modalities are just denoted in Table 3.1 and not used in the presented experiments.

Table 3.1: Overview of selected emotional speech corpora.

Name	Language	Emotions	Length/Recordings	Subjects	Type	Modality
emoDB	German	ang bor dis fea hap neu sad	00:22	5m 5f	sim	a
DES	Danish	ang hap neu sad sur	00:28	2m 2f	sim	a
GEMEP	French	han des anx amu int ple pri joy rel pan irr sad adm ten dis cot sur	1 260 utt	5m 5f	sim	a v
SAFE	English	fea neg pos neu	07:00	14m 12f 2c	sim	a v
ABC	German	agg che inx ner	11:30	4m 4f	sim	a v
NIAVE	Italian	hap iro fea ang sur sad	00:07	13m 13f	sim	a v
Amir	Hebrew	ang dis fea joy sad neu	15:30	16m 15f	ind	a b
eNTERFACE	English	ang dis fea hap sad sur	01:00	34m 8f	ind	a v
SUSAS	English	hst mst neu scr	01:01	13m 13f	ind	a
emoSDB	English	four conditions of stress	00:20	4	ind	a
BNDB	English	continuous traces	01:26	31m 94f	nat	a v
VAM	German	discrete values of A and V	00:48	15m 32f	nat	a v
Castaway	English	36 ’everyday’ emotions	05:00	10	nat	a v
EmoTV	French	ang des dis dou exa fea irr joy pai sad ser sur wor neu	00:12	48	nat	a v
ISL Meeting	English	pos neg neu	103:00	660	nat	a
ACC	English	neg nne	1 367 utt	691m 776f	nat	a
Messages	French	pos neg neu	478 rec	103	nat	a
CEMO	French	ang fea hur pos rel sad sur neu	20:00	271m 513f	nat	a
UAH	Spanish	ang bor dou neu	02:30	60	nat	a
CallfriendEmo	English	bor hap han int pan sad neu	1 888 utt	4m 4f	nat	a
NIMITEK	German	*ang ner sad joy com bor fea dis neu*	*15:00*	3m 7f	evo	a v
TUM AVIC	English	LoI−2 LoI−1 LoI0 LoI+1 LoI+2	10:30	11m 10f	nat	a v
ITSPOKE	English	pos neg neu	100 rec	20	nat	a
EmoRec	German	four quadrants of VA emotion space	33:00	35m 65f	evo	a v b
AIBO	German	ang bor emp hel joy mot rep sur irr neu oth	912 rec	51c	nat	a
	English		*01:30*	30c	nat	a
WOZdc	English	cof puz hes	*00:50*	17c	nat	a v
SAL	English	continuous traces	10:00	20	ind	a v
LMC	German	four dialogue barriers (bsl lst cha wai)	*56:00*	64m 66f	nat	a v (b)

Times given in *italic* denote the complete corpus, the emotional content may be less. Abbreviations: a: audio, v: video, b: bio-physiological, sim: simulated, ind: induced, evo: evoked, nat: naturalistic, emotional terms are given in the appendix.

3.2 Reviewing the speech-based Emotion Recognition Research

As already mentioned, automatic emotion recognition is based on pattern recognition methods. Thus, this chapter first introduced datasets which are suitable to train classifiers to recognise simulated and naturalistic emotions. Afterwards, emotional acoustic characteristics have to be extracted and classifiers to recognise different emotions have to be modelled. Therefore, in the following, commonly used methods for feature extraction, feature selection and classification are described and several attempts to evaluate classifiers for emotional speech are discussed. 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.

For general pattern recognition problems, as well as for emotion recognition from speech, the first and fore-most important step is to extract a meaningful and informative set of features. From speech two basic characteristics can be distinguished: 1) “What has been said?” and 2) “How has it been said?”. The first aspect is described by linguistic features, the second aspect by acoustic features, which will be of greater interest for this thesis. Mostly, the acoustic features are further divided into short-term acoustic features also called Low Level Descriptors (LLDs) and longer-term supra-segmental features. Furthermore, researchers distinguish between spectral, prosodic, and paralinguistic (non-linguistic) features [Schuller et al. 2010a].

As a starting point a comparably small set of features consisting of static characteristics are employed. In this case, mostly pitch, duration, intensity (energy) and spectral features – such as Mel-Frequency Cepstral Coefficients (MFCCs), Perceptual Linear Predictions (PLPs) (cf. [Bitouk et al. 2010; Böck et al. 2010]) – or formants are applied (cf. [Bozkurt et al. 2011; Gharavian et al. 2013]). Less frequently, voice quality features as Harmonics-to-Noise Ratio (HNR), jitter, or shimmer are used (cf. [Li et al. 2007]). However, they recently gain greater attention (cf. [Kane et al. 2013; Scherer 2011]). Lately, also the supra-segmental nature of emotions is taken into account. When investigating supra-segmental speech units, as for instance words or turns, the extracted short-term features have to be normalised over time by using descriptive statistical functionals, as such speech units vary over time. In this case, the application of statistical functionals assures that each entity has the same number of feature vectors independent of the unit’s spoken length (cf. [Schuller et al. 2011c]). Popular statistic functionals covering, for instance, the first four moments (mean, standard deviation, skewness, and kurtosis), order statistics, quartiles, and regression statistics. A very comprehensive list can be found in [Batliner et al. 2011]. This approach results in very large feature vectors containing thousands of features. On the other hand, they are showing promising results in emotion recognition [Cullen & Harte 2012; Schuller et al. 2009a]. The application of supra-segmental information has become very popular. But it is still unclear what is the best unit for emotion recognition from speech (cf. [Batliner et al. 2010; Vlasenko & Wendemuth 2013]).

Other feature extraction approaches, rarely used, utilise expert-based features known to characterise hard to find but perceptually more adequate information. Applied for emotion recognition are the Teager Energy Operator (TEO) (cf. [Cullen & Harte 2012]), perceptual quality metrics features (cf. [Sezgin et al. 2012]), or formant shift information (cf. [Vlasenko 2011]).

Unfortunately, to date there is neither a large-scale comprehensive comparison of the usefulness of various feature sets for emotion recognition, nor a psychologically derived description of emotional speech patterns comparable to the FACS for facial expressions. Some preliminary efforts are made by [Vogt & André 2005; Cullen & Harte 2012], comparing some feature sets to a greater extend. It has to be further mentioned that only few research even investigated automatically extractable emotion specific acoustic feature sets (cf. [Albornoz et al. 2011; Cullen & Harte 2012]), although such a set is predicted by psychological research (cf. [Johnstone et al. 2001]).

In addition to a discriminating feature set, a powerful classification technique is also needed. As emotion recognition from speech is related to Automatic Speech Recognition (ASR), the same dynamic classifiers as Hidden Markov Models (HMMs) are often used (cf. [Nwe et al. 2003; Zeng et al. 2009]). They implicitly warp the observed features over time and thus allow to skip additional computation steps to obtain the same number of feature vectors for different lengths of investigated units. Also the one-state HMM, called Gaussian Mixture Model (GMM), is used (cf. [Böck et al. 2010; Zeng et al. 2009]). This classifier is known to be very robust for speaker and language identification tasks, where the acoustic characteristics are just slowly changing over time (cf. [Kockmann et al. 2011; Vlasenko et al. 2014]). These classifiers have been proven to achieve quite high and robust recognition results for different types of emotions. Furthermore, several methods for model adaptation exists, for instance Maximum Likelihood Linear Regression (MLLR) or Maximum A Posteriori (MAP) estimation, allowing to overcome the problem of few data (cf. [Gajšek et al. 2009]) or to perform a speaker adaptation (cf. [Hassan et al. 2013; Kim et al. 2012a]). Another quite popular classifier is the Support Vector Machine (SVM) (cf. [Schuller et al. 2011c]), it is able to handle very large feature spaces and can avoid the "curse of dimensionality" (cf. [Bellman 1961]). Also discriminative classifiers such as Artifical Neural Networks (ANNs) and decision trees are used (cf. [Glüge et al. 2011; Wöllmer et al. 2009]). But, as these classifiers are less robust to overfitting and thus require greater amounts of data, they are used just rarely (cf. [Schuller et al. 2011c]).

Just recently a further approach was used by combining several classifiers and thus improving the training stability (cf. [Albornoz et al. 2011; Ruvolo et al. 2010]). This approach is called ensemble-classifiers. Therein, the most crucial point is the combination of the different classifier outputs. Simple but also promising approaches are based on majority voting (cf. [Anagnostopoulos & Skourlas 2014]), more complex approaches integrate a “meta-classifier” that learns how to combine the outputs of the “base-classifiers” (cf. [Wagner et al. 2011]). This method of combining several classifiers is called fusion, when various modalities are combined (cf. Section 4.3.4).

Another important topic is the classifier validation. Here the majority of researchers still relies on speaker-dependent strategies such as cross-validation (cf. [Schuller et al. 2011c]). Within the last five years some researchers ensured true speaker independence by using a Leave-One-Speaker-Out (LOSO) or Leave-One-Speaker-Group-Out (LOSGO) validation strategy (cf. [Schuller et al. 2009a])), sometimes also called interindividual validation (cf. [Böck et al. 2012b]). When reporting about performance measures, significance tests are mostly ignored in the speech emotion recognition community, with the exception of [Seppi et al. 2010], for instance. It should be noted that significant improvements can only rarely be achieved by a single new method. However, it is important to report about this method and by several methods in combination significant improvements are possible (cf. [Seppi et al. 2010]).

Especially in [Schuller et al. 2011c], it is emphasised that the comparability between research results is quite low, as differences in the applied feature sets, classifiers, and mainly in the validation strategy exist. Moreover, the utilised data differ in their acoustic conditions, speaker-groups and emotional labels. This issue on the diverse emotional corpora was already discussed in more detail in Section 3.1.

The research community attempted to overcome these problems by announcing several competitions and benchmark datasets, allowing the competition and comparison of ones own methods. The first attempts where made by the CEICES-initiative, where seven research groups combined their efforts in generating an emotionally labelled database and unified their feature selection [Batliner et al. 2006]. The correspondingly generated database, the FAU AIBO Corpus, served as a basis for the INTERSPEECH 2009 Emotion Challenge (cf. [Schuller et al. 2009c]), the first open public evaluation of speech-based emotion recognition systems and starting point for an annual challenge series. The corresponding challenges defined test, development, and training partitions, and provided the acoustic feature set with baseline results, allowing a real comparison of all participants. The INTERSPEECH 2009 EMOTION Challenge had three sub-challenges: Open Performance, Classifier and Feature, where five-class or two-class emotion problem had to be solved. In Open Performance participants could use their own features. The best results were 41.7 % Unweighted Average Recall (UAR) for the five-class task (cf. [Dumouchel et al. 2009]) and 70.3 % UAR for the two task class (cf. [Kockmann et al. 2009]). In the Classifier sub-challenge participants had to use a set of provided features. Only for the five-class task a UAR of 41.6 % could be achieved (cf. [Lee et al. 2009]). In the two-class task, the baseline was not exceeded by any of two participants. In the Feature sub-challenge participants had to design the 100 best features to be tested under equal conditions. The feature sets provided could not exceed the baseline feature set provided by the organisers.

The INTERSPEECH 2010 Paralinguistic Challenge (cf. [Schuller et al. 2010b]), aimed to provide an agreed-upon evaluation set for the use of paralinguistic analysis. In three different sub-challenges, researchers were encouraged to compete in the determination of the speakers' age, the speakers' gender, and the speakers' affect. The organisers used the "level-of-interest" as affect. The TUM AVIC corpus,having five different classes, was utilised for the affect sub-challenge (cf. [Schuller et al. 2009b]). An extended version of the INTERSPEECH 2009 Emotion Challenge was provided as feature set extended with paralinguistic features, such as F_0-envelope, jitter and shimmer (cf. [Schuller et al. 2010b]). The best Correlation Coefficient (CC) for the affect sub-challenge is 0.627 (cf. [Jeon et al. 2010]).

This type of challenges was continued with the INTERSPEECH 2011 Speaker State Challenge (cf. [Schuller et al. 2011a]). It aimed to evaluate speech-based speaker state recognition systems on the mid-term states of intoxication and sleepiness. Two sub-challenges addressed both two-class problem with a provided corpus. For the intoxication challenge, a UAR of 70.5 % (cf. [Bone et al. 2011]) and for the sleepiness challenge a UAR of 71.7 % (cf. [Huang et al. 2011]) could be achieved at best.

The INTERSPEECH 2012 Speaker Trait Challenge provided a basis to assess speech-based trait evaluation. It consisted of three sub-challenges, a five-class personality sub-challenge, a two-class likability sub-challenge and a two-class pathology sub-challenge (cf. [Schuller et al. 2012a]). The best UARs are 69.3 % for the personality (cf. [Ivanov & Chen 2012]), 64.1 % for the likability (cf. [Montacié & Caraty 2012]), and 68.9 % for the pathology sub-challenge (cf. [Kim et al. 2012b]). The INTERSPEECH 2013 Computational Paralinguistics Challenge was centred around the evaluation of social signal, conflict, emotion, and autism detection (cf. [Schuller et al. 2013]). For the social signal sub-challenge, non-linguistic events such as laughter or sigh from a speaker had to be detected. The conflict sub-challenge analysed group discussions to detect conflict situations. In the autism sub-challenge the type of pathology of a speaker had to be determined. Again, this challenge has an emotional sub-challenge, consisting of a 12-class problem using the GEMEP corpus of acted emotions (cf. [Bänziger et al. 2012]). The best participant achieved 42.3 % UAR in this sub-challenge (cf. [Gosztolya et al. 2013]). The last challenge of this series to date is the INTERSPEECH 2014

Computational Paralinguistics Challenge having the two sub-challenges Cognitive Load and Physical Load, the results will be presented at the INTERSPEECH 2014 held from 14th to 18th September in Singapore.

In 2011 another series of challenges started with the Audio/Visual Emotion Challenge and Workshop (AVEC). This challenge aimed at combining the acoustic and the visual emotion recognition community efforts on naturalistic material. It always consisted of three sub-challenges, focussing on audio analysis, video analysis and the combined audio-visual analysis. The first and second challenge of this series aimed on the detection of emotions in the SEMAINE SAL corpus in terms of `positive/negative valence`, and `high/low arousal`, `expectancy` and `power` on pre-defined chunks for the 2011 Challenge (cf. [Schuller et al. 2011b]) and continuous time and dimension values for the 2012 Challenge (cf. [Schuller et al. 2012b]). As best recognition result in the audio sub-challenge 2011 on pre-defined chunks a UAR of 57.7 % could be achieved (cf. [Meng & Bianchi-Berthouze 2011]). In 2012 testing against continuous time and dimension values a CC of 0.168 % could be achieved (cf. [Savran et al. 2012]).

The 2013 and 2014 AVEC challenges were extended to investigate a more complex mental state, the depression, of the user. Thus, these challenges consisted of two sub-challenges: First, the fully-continuous emotion detection from audio, video and audio-video information on the three dimensions `arousal` and `valence`. The second sub-challenge dealt with the detection of depression, also from audio, video and audio-video information (cf. [Valstar et al. 2013]). The 2013 best participant achieved a CC of 0.168 % on the affective audio-visual sub-challenge (cf. [Meng et al. 2013]).

These challenges, where test conditions are strictly pre-defined, have the disadvantage that due to the pre-defined set of features merely different classification systems and learning methods are evaluated. These challenges did not help for the evaluation of new feature sets or in identifying new characteristic patterns. For this purpose certain databases are established in the community. These databases are publicly available, well described and widely used (cf. [Böck et al. 2010; Ruvolo et al. 2010; Schuller et al. 2007a; Schuller et al. 2009a; Vogt & André 2006]). Therefore, they are generally referred as benchmark corpora. Unfortunately, these corpora are not used in the above mentioned challenges. The benchmark corpora include emoDB, eNTERFACE for simulated databases and VAM for spontaneous interactions. These databases also serve as basis for my investigations. So far, there is no generally accepted representative database for naturalistic interactions. In my investigations, I rely on the LMC, as it contains naturalistic interactions within a WOZ-controlled HCI. The effort in creating such a database is quite high and thus, these databases are often not fully public available. Furthermore, the annotation process is quite expensive and emotional labels are not as reliable as for acted emotional data.

3.3 Classification Performances in Simulated and Naturalistic Interactions

As stated earlier, emotion recognition from speech started with a few small databases of acted emotions, for instance emoDB. The results achieved on these databases were quite promising. In the following section, I sketch the achieved results and conducted efforts on selected databases with simulated and naturalistic affects. For this, I restrict the report on results that are comparable since the same corpora and validation methods are used. Some corpora will later serve as database for my own investigations as well (cf. Chapter 6) and are presented in more detail in Chapter 5. Finally, a comparison of the reported achieved recognition results on acted and naturalistic databases is given in Table 3.2 on page 39.

The authors of [Böck et al. 2010] compared different feature sets and several architectures of HMMs on emoDB's seven emotional classes `anger`, `boredom`, `disgust`, `fear`, `joy`, `neutral`, and `sadness`. Using a ten-fold cross validation and 39 spectral features, an overall accuracy of 79.6 % was reported. The authors in [Schuller et al. 2007a] employed the same emotions on emoDB with roughly 4 000 acoustic and prosodic features and trained a Random Forest (RF) classifier. Additionally, they applied a two-fold cross-validation with a division into two stratified sub-folds to assure speaker-independence. They achieved an recognition rate (accuracy) of 72.3 %. In [Schuller et al. 2009a] an UAR of 73.2 % for GMMs and 84.6 % for SVMs with linear kernel could be achieved by using the same features. For validation a two-fold cross-validation with a division into two stratified sub-folds was used as well.

By using a novel feature extraction, the Spectro-Temporal-Box-Filters containing short time scale, medium time scale, and long time scale features, the authors in [Ruvolo et al. 2010] achieved an overall accuracy of 78.8 % on emoDB's seven emotions, by performing a hierarchical classification with late fusion of multi-class SVMs. For validation a LOSO strategy is applied. The authors of [Vogt & André 2006] performed a gender differentiation to improve the automatic emotion recognition from speech. They generated a gender-independent and gender-specific set of features. By using a Naïve Bayes classifier with a LOSO validation strategy, the emotion identification performance for emoDB's seven emotion classes improved to an accuracy of 86.0 %. The authors used the gender information known a-priori. When employing an automatic gender identification system, the emotion recognition achieves just 82.8 %.

In [Schuller et al. 2009a], a two class UAR of 91.5 % for GMMs and 96.8 % for SVMs with linear kernel could be achieved by clustering emoDB's emotional sentences into representatives of `low arousal` (`A`−) and `high arousal` (`A`+) emotions. For this, the authors used 6 552 features from 56 acoustic characteristics and 39 functionals and applied a ten-fold cross-validation strategy with a division into two stratified sub-folds having an equal portion of male and female speakers to assure speaker-independence.

In recent years, the research community shifted from data with simulated emotional expressions to more naturalistic emotional speech data, also due to results stating that in realistic recordings the expressions are of higher variability and not obvious at all (cf. [Truong et al. 2012]). Faced with such kind of data, the remarkable results of emotion recognition achieved on simulated data dropped, when using corpora with naturalistic emotions. Two explanations can be given: First, naturalistic interactions contain more blended emotions and second the expressiveness is lower than for simulated emotions [Batliner et al. 2000; Mower et al. 2009].

One of the first freely available databases containing naturalistic data was VAM. As this database is labelled within the `arousal-valence` space, a clustering into discrete emotional clusters has to be performed. In [Tarasov & Delany 2011] the dimensions are discretised in `low`, `middle` and `high` values of the dimension. Using an SVM with Radial Basis Function (RBF) Kernel, they achieved a weighted accuracy of 62.0 % on `arousal` in a five-fold cross validation with 384 acoustical and spectral features[6].

The authors of [Schuller et al. 2009a] also conducted experiments on VAM, for comparison with their results achieved on emoDB. Using the same 6 552 features and LOSGO validation, they achieved a UAR of 76.5 % using GMMs and 72.4 % for SVMs with linear kernel on VAM to distinguish between `high` and `low arousal`. Another approach investigated by Zhang et al. tries to compensate the data sparseness by agglomerating different emotional speech data for training. The normalised acoustic material of the databases ABC, TUM AVIC, DES, SAL, and eNTERFACE are use for this to train the SVM with a linear Kernel using 6 552 features. They uses the same emotional recombination as presented in [Schuller et al. 2009a] to perform the cross-corpora training. An UAR of 69.2 % could be achieved (cf. [Zhang et al. 2011]).

Another approach is presented in [Sezgin et al. 2012]. The authors introduce a new set of acoustic features for emotion recognition based on perceptual quality metrics instead of the speech production modelling used as basis for common spectral features (cf. Section 4.2.1). They extracted seven perceptual features[7]. The motivation for using these perceptual features is the fact that “the harmonic structure of emotional speech is much more similar to a periodical signal with stable harmonics with respect

[6] This feature set is identical to the Interspeech 2009 Challenge feature set [Schuller et al. 2009c].

[7] Their perceptual features are: a) average harmonic structure magnitude of the emotional difference, b) average number of emotion blocks, c) perceptual bandwidth, d) normalised spectral envelope, e) normalised spectral envelope difference, f) normalised emotional difference, and g) emotional loudness.

to unemotional speech" [Sezgin et al. 2012]. To evaluate their perceptual features, they applied them on the same two-class `arousal` recognition problem defined by [Schuller et al. 2009a] for VAM. They achieved an UAR of 69.4 % using an SVM applying an improved Soft-Majority Vote (S-MV) (cf. [Sezgin et al. 2012]).

As already discussed in Section 3.1.2 another type of corpora recently emerged, namely naturalistic interactions without the purpose to induce specific emotions but rather to evoke general emotional reactions during an interaction. One representative is the audio-visual SAL database, which is part of the final HUMAINE database [McKeown et al. 2010]. The data contains audio-visual recordings from a naturalistic HCI, where users are driven through a range of emotional states. The data has been labelled continuously by four annotators with respect to the `activation` dimension[8] using FEELtrace (cf. [Cowie et al. 2000], Section 4.1.2). The authors in [Schuller et al. 2009a] extracted 1 692 turns by an automatic voice activity detection system and averaged the continuous `arousal` labels over one complete turn to decide between `A`− (mean below zero) and `A`+ (mean above zero). Afterwards, they extracted their set of 6 552 acoustic features and achieved an UAR of 55.0 % on an SVM with a linear kernel and 61.2 % utilising a GMM. Furthermore, the cross-corpora approach by Zhang et al. is applied on SAL as well. For this, the normalised acoustic material ABC, TUM AVIC, DES, VAM, and eNTERFACE is used to train an SVM with a linear kernel on the same two class problem as for [Schuller et al. 2009a]. By using 6 552 features, an UAR of 61.6 % could be achieved for the classification of `A`− and `A`+ [Zhang et al. 2011].

Table 3.2: Classification results in percent on different databases with simulated and naturalistic emotions. Furthermore, comparable results between several corpora are highlighted.

Corpus	Result	Classes	Comment
emoDB			
	72.3 *Acc*	7	4000 acoustic and prosodic features, two-fold cross validation, RF [Schuller et al. 2007a]
	73.2 UAR	2	6 552 acoustic features, LOSO validation, GMM [Schuller et al. 2009a]
	78.8 *Acc*	7	STBF features, LOSO, hierarchical classification of multi-class SVMs [Ruvolo et al. 2010]
	79.6 *Acc*	7	39 spectral features, ten-fold cross validation, HMM [Böck et al. 2010]

Continued on next page

[8] As stated in Section 2.1, `activation` is a synonymously used term for `arousal`.

Table 3.2 – *Continued from previous page*

Corpus	Result	Classes	Comment
emoDB			
	84.6 UAR	2	6 552 acoustic features, LOSO validation, SVM with linear kernel [Schuller et al. 2009a]
	86.0 *Acc*	7	Naive Bayes, gender-differentiation, a-priori gender information [Vogt & André 2006]
	91.5 UAR	2	6 552 acoustic features, LOSO validation, GMM [Schuller et al. 2009a]
	96.8 UAR	2	6 552 acoustic features, LOSO validation, SVM with linear kernel [Schuller et al. 2009a]
VAM			
	62.0 *Acc*	3	384 acoustic features, five-fold cross validation, SVM-RBF [Tarasov & Delany 2011]
	69.2 UAR	2	6 552 acoustic features, cross-corpora, LOSGO validation, SVM with linear kernel [Zhang et al. 2011]
	69.4 UAR	2	9 perceptual features, LOSGO validation, SVM-RBF and S-MV [Sezgin et al. 2012]
	72.4 UAR	2	6 552 acoustic features, LOSGO validation, SVM with linear kernel [Schuller et al. 2009a]
	76.5 UAR	2	6 552 acoustic features, LOSGO validation, GMM [Schuller et al. 2009a]
SAL			
	55.0 UAR	2	6 552 acoustic features, LOSGO validation, SVM with linear kernel [Schuller et al. 2009a]
	61.2 UAR	2	6 552 acoustic features, LOSGO validation, GMM [Schuller et al. 2009a]
	61.6 UAR	2	6 552 acoustic features, cross-corpora, LOSGO validation, SVM with linear kernel [Zhang et al. 2011]

This comparison demonstrated that the promising results using databases with simulated affects cannot be reproduced with naturalistic affect databases, even if the number of classes is reduced. The results drop down from 96.8 % for a two-class problem and simulated emotions to 61.6 % for naturalistic emotions. This decrease cannot even be compensated, when sophisticated feature extraction methods or thousands of features are being used. The investigations presented in this section will be later considered again to discuss my own contributions on the improvement of emotion recognition from speech. This comparison further reveals that the achieved results by SVMs and GMMs are quite similar. Sometimes an SVM approach achieves better results (two-class problem on emoDB), sometimes the GMM has a better performance (two-class problem on VAM or SAL). Thus, both methods promise a good classification performance. In [Böck 2013] it is shown that GMMs can be used on a wide range of corpora. Hence, they appear better suited for naturalistic emotional corpora.

3.4 Open issues

Although the review of the previous section shows that the field of emotion recognition receives a strong attention, there are still (many) open questions. In the following, I discuss certain developments and show the gaps, which I would like to close with my work. The first two open issues directly follow from considerations made in the ASR community and thus are examined together in one chapter. These issues examine methodological improvements for the emotion recognition from speech. The next two issues go beyond this classical emotion recognition approach as they broaden the short-term emotion recognition towards a longer-term interactional emotion recognition. In this context, the third issue expands the type of acoustical patterns that have to be considered to understand the conversational relationship within an HCI, and the fourth open issue describes the problem that to date in affective computing only short-term emotions are considered although longer-term affective states are known, amongst others, to influence the user's behaviour and problem solving ability.

3.4.1 A Reliable Ground Truth for Emotional Pattern Recognition

A first prerequisite for emotion recognition is the availability of data for training and testing classifiers. In the psychological research it is still an open debate whether a categorial or dimensional approach should be used, and which number of categories or dimensions are needed for an adequate modelling of the users behaviour (cf. Section 2.1). Schuller et al. pointed out that for the emotion recognition community a rather straightforward engineering approach was used: “we take what we get, and so far, performance has been the decisive criterion” [Schuller et al. 2011c]. This very practical approach allows the evaluation of various feature extraction and classification methods. If the results should be used to regulate and control the reactions of the system adequately, according to the user's behaviour, the interpretation of the data labels can no longer be neglected. This problem has been addressed by many researchers in the community (cf. [Batliner et al. 2011; Ververidis & Kotropoulos 2006; Zeng et al. 2009]), but a proper solution has not appeared yet.

Especially for naturalistic interactions one has to rely on the emotional annotation of data, as the emotional labels are not a-priori given. But as already shown in [Scherer 2005a], it is very difficult to correctly identify and entitle one's own emotional experiences. Truong et al. also shows that it makes a difference whether a self- or foreign rating is carried out. However, both types of ratings are reliable (cf. [Truong et al. 2012]). Besides this and some other investigations (cf. [Grimm & Kroschel 2005; Lefter et al. 2012]), there are only a few emotional speech corpora giving an annotation reliability, as it is known and common, for instance, in computational linguistic analysis (cf. [Artstein & Poesio 2008; Hayes & Krippendorff 2007]).

Also the label generation has so far just been insufficiently treated. No known publication deals with the usefulness of different methods for emotional speech labelling. The expected reliability is also just rarely being investigated on single corpora (cf. [Grimm & Kroschel 2005; McKeown et al. 2012]). Furthermore, there are only few publications that deal with methods to improve and increase the reliability of emotional speech labelling. Mostly only single datasets are considered for methodological improvements (cf. [Callejas & López-Cózar 2008; Clavel et al. 2006; McKeown et al. 2012]). Furthermore, it can be seen in Table 3.1 on page 31 that there are barely two datasets, using the same emotions. Thus, for cross-corpora analyses researchers have to apply a trick and, for instance, cluster different emotional classes (cf. [Schuller et al. 2010a]). This evidences that a unifying labelling method is needed. The issue of a proper emotional labelling as well as its reliability is addressed in Section 6.1.

3.4.2 Incorporating Speaker Characteristics

Another issue that has only been rarely investigated for emotion recognition is the speaker clustering or speaker adaptation, to improve the emotional modelling. Although the incorporation of age and gender differences has been used to improve speech and speaker recognition [Kelly & Harte 2011; Kinnunen & Li 2010] it has only rarely been used for emotion recognition. As a first factor for speaker adaptation, a gender differentiation was used. The authors of [Dobrišek et al. 2013] utilised a gender-specific Universal Background Model (UBM)-MAP approach to improve the recognition performance, but did not compare their results on a broader context. Another publication used a two-stage approach to automatically detect the gender and afterwards perform the emotion classification (cf. [Shahin 2013]). Based on the "Emotional Prosody Speech and Transcripts database" (Six basic emotions including the neutral state), they could improve the classification performance by an average of

11% utilising supra-segmental HMMs. The authors of [Vogt & André 2006] applied a gender differentiation to improve the automatic emotion recognition from speech. They noticed an absolute difference of approx. 3 % between the usage of correct and recognised gender information.

But these publications only investigate the obvious gender-dependency. No other factors as, for instance, age is considered. Although it is known that age has an impact on both the vocal characteristics (cf. [Harrington et al. 2007; Linville 2001]) and the emotional response (cf. [Gross et al. 1997; Lipovčan et al. 2009]), it is also not investigated whether the previously mentioned improvements by [Shahin 2013; Vogt & André 2006] are dependent on the utilised material or other factors. Additionally, these studies are conducted on databases of simulated affects. Thus a proof that these methods are suitable for natural interactions as well is missing. My extension to these studies is presented in Section 6.2. A further integration of my approach into a fusion of fragmentary data is discussed in Section 6.3.

3.4.3 Interactions and their Footprints in Speech

Speech contains more than just emotions. It includes information about the speaker's feelings and his mental state. More importantly it also determines the nature and quality of the user's relationship to its interlocutor. Vinciarelli et al. called this "behavioural cues", which in most cases accompanies the verbal communication. The cues are sensed and interpreted outside conscious awareness, but greatly influence the perception and interpretation of messages and situations (cf. [Vinciarelli et al. 2009]).

From psychological as well as linguistic research these cues consist of linguistic vocalisations, including all non-words typically such as "uhm", and non-linguistic vocalisations, including non-verbal sounds like laughing or crying. Linguistic vocalisations serve as a replacement of words that actually cannot be uttered and thus, indicate a high cognitive load (cf. [Corley & Stewart 2008]). They are also used as so-called "back-channelling" to signalise the progress of the dialogue and regulate turn-taking (cf. [Allwood et al. 1992]). Non-linguistic vocalisations provide some information about the speakers attitude towards situations and are mostly uttered as "vocal outbursts" (cf. [Hawk et al. 2009; Schröder 2003]).

Although linguistic and non-linguistic vocalisations have been a subject of extensive research, they have rarely been interpreted in terms of behavioural cues within HCI. The detection has mostly aimed at the improvement of ASR systems, where these vocalisations are seen as a form of noise rather than a source of information (cf. [Liu et al. 2005]). The function within the dialogue is well known for linguistic vocalisations, but to the best of my knowledge only preliminary efforts have been made so far to investigate the use and purpose of linguistic vocalisations within HCI. If they are investigated, then the main research question is how to embed these behavioural cues into conversational agents to mimic a human being and improve the conversation (cf. [Kopp et al. 2008]). But, the detection and interpretation of human uttered linguistic vocalisations with the aim to distinguish several interactional functions has not been pursued. An exception is the detection of the non-linguistic vocalisations laughter and crying, which are of special interest because of their ubiquitous presence in interactions (cf. [Knox & Mirghafori 2007; Scherer et al. 2012]). My research presented in Chapter 7 deals with the issue whether specific linguistic vocalisations can be used to interpret an ongoing HCI and which user characteristics have to be taken into account.

3.4.4 Modelling the Temporal Sequence of Emotions in HCI

The observation of emotional states and interaction signals alone is not sufficient to understand or predict the human behaviour and intelligence needed for future Companion systems. In addition to emotional observations and interaction patterns, a description about the progress of an interaction is necessary as well. Only by a long-term emotional observation of a user, the individual behaviour can be estimated. The pure observation of short-term affective states, the emotions, is not able to provide such a description of the user's behaviour. Thus, besides short-term emotions the system should also observe the user's longer-term affective development. These longer-term affective states are called moods. Moods influence the user's cognitive functions, his behaviour and judgements, and – of importance for HCI – also the individual (creative) problem solving ability (cf. [Morris 1989; Nolen-Hoeksema et al. 2009]). As moods cannot be observed directly from human's bodily reactions, they have to be estimated indirectly.

This aspect has to date not been addressed from a human perspective in the HCI. Instead the community aims to equip the computational agent with a more human-like behaviour by using a mood modelling that changes the agents behaviour (cf. [Becker-Asano 2008; Gebhard 2005]). Techniques that try to model an emotional development within the ongoing interaction, to predict, for instance, changes of the mood from `positive valence` to `negative valence` are not presented so far. In Chapter 8, I present my mood modelling technique to enable a technical system to predict the mood development of the human dialogue partner based on the observation of the user's directly assessable emotional responses.

Chapter 4

Methods

Contents

As discussed in Chapter 3, several steps are necessary, in order to be able to recognise emotions within a naturalistic interaction. First, the training material have to be recorded and enriched with emotional labels. These labels should cover the variety of occurring emotions. Furthermore, the reliability of these labels need to be assured. Hence, in Section 4.1 I introduce methods for the emotional annotation of datasets and utilised reliability measures.

Afterwards, acoustic characteristics like changes in pitch or loudness describing the emotional state of the speaker have to be identified. Such obtained features for emotion recognition and their origin are introduced in Section 4.2. Furthermore, a common arrangement in short-term and longer-term features as well as spectral and prosodic ones is applied. Additionally, side-effects such as age or gender influences are also discussed.

Based on the extracted features and the labelled data suitable classifiers can be trained. In my research I focussed on HMMs and GMMs as classifiers. The main principles behind these classifiers are introduced in Section 4.3. Furthermore, optimal feature and parameter sets are investigated and will serve as a best practice for my later research.

Finally, common evaluation methods are presented (cf. Section 4.4). These cover the different arrangements of the utilised data material, to generate the validation set. Furthermore, different classifier performance measures are introduced and their differences are discussed. Additionally, utilised significance measures are depicted.

4.1 Annotation

As I have stated in Section 1.2, the emotional annotation of speech material is the very first step for affect recognition. In the case, where simulated emotional data is used, this task can be solved quite easy. In the case of emotional acted data, the *label* is clearly instructed to the actor by the experimenter. The expressive quality can afterwards easily be assessed via perception tests [Burkhardt et al. 2005]. For induced emotions, the experimental design mostly ensures a valid ground truth. In this case, this can be secured via perception tests as well [Martin et al. 2006]. But, for data gathered within a naturalistic interaction or, if the previous mentioned procedure of perception tests cannot be conducted, manual annotation is needed.

Unfortunately, this task is quite challenging, as it has to be done fully manual by well-trained labellers, who need to be familiar with assessing emotional phenomena. As emotions and their assessments are quite subjective (cf. Section 2.2.1), a large number of labellers is commonly utilised to label the emotional data. Afterwards a majority vote is used to come up with an assessment that can be regarded as valid. To furthermore evaluate the labelling quality and give a statement on the correctness of the found phenomena the Inter-Rater Reliability (IRR) has to be computed.

4.1.1 Transcription, Annotation and Labelling

Several terms are used, to distinguish the different labelling pre-proocessing steps, namely (literal) *transcription*, (phonetic) *annotation*, and (emotional) *labelling*.

Transcription denotes the process of translating the spoken content into a textual description. Hereby, it is purely written down what has been said, including mispronunciations, dialects or paralinguistic expressions e.g. laughter or discourse particles. It depends on the subsequent task, at which precision the timing information are provided. This precision can be done on (dialogue) turn, sentence, utterance, or word level. It should be noted that there is no assessment of the spoken content.

The second term, *annotation*, describes the optional step of adding the information how something has been said. This provides the option of easily interpreting the previous gained textualization and finding particular phenomena. These phenomena contains for instance pauses, breathing, accents/emphases, word lengthening. Particularly linguistic research developed several methods to add meaning to the pure textual description. Such approaches are, for instance, the Gesprächsanalytisches Transkriptionssystem (dialogue analytic transcription system) (GAT) [Selting et al. 2009], halb-interpretative Arbeits-Transkription (semi-interpretive working transcription) (HIAT) [Rehbein et al. 2004], or Codes for the Human Analysis of Transcripts (CHAT) [MacWhinney 2000].

The last term *labelling* describes further levels of meaning. These levels are detached from the textual *transcription* and describe e.g. affects, emotions, or interaction patterns. To assess these meanings, a broader textual context of the interaction is needed, which could even imply the utilisation of further information as facial expressions [Lefter et al. 2012]. These phenomena are detached from specific timing levels and heavily depend on subjective interpretations. Several labellers are needed for a valid assessment and a reliability calculation has to be performed.

Unfortunately, the introduced terms are not strictly separated. Since in linguistic analysis, the knowledge about how something is said is the base for research, the method described here is referred to as *transcription*, too. In pattern recognition the terms *annotation* and *labelling* are used synonymously to define the task of adding additional information to the pure data material. A literal transcription is normally not needed for affect recognition.

4.1.2 Emotional Labelling Methods

As I have mentioned before, the data gathered in naturalistic interactions lacks of emotional labels. Therefore, this information has to be added in an additional labelling step. A broad field of research deals with the question of how to label emotional affects within experiments. The most promising method of obtaining a valid assignment would be a self-assessment by the observed subject itself [Grimm & Kroschel 2005]. Unfortunately, this is not always feasible and reproducing valid labels has some flaws, as a subject is not always able to verbalise the emotional state actual felt (cf. [Scherer 2005a]). Especially in situations where the subject is highly involved in the experimental scenario, it is counterproductive to interrupt the experiment for regular self-assessments. Therefore, several labelling methods are devised where another subject, called labeller or rater, has to assess the experimental data and has to assign an emotional label. They are mostly based on questionnaires. To obtain a valid assessment mostly several labellers (>6) are required.

Sadly, also this kind of labelling does not necessarily reflect the emotion truly felt by a subject, as a number of “input- and output-specific issues” [Fragopanagos & Taylor 2005] influence the assessment, such as display rules and cognitive effects, or felt emotions are not always perceivable by observers [Truong et al. 2012]. To overcome these issues it is advisable to employ several raters to label the same content and use a majority voting [Fragopanagos & Taylor 2005]. Furthermore, [Truong et al. 2008] found that the averaged agreement between repeated self-rating was lower than the inter-rater agreement of external raters.

Several studies also investigated the influence of contextual information on the annotation. In [Cauldwell 2000] the authors present a study, where they investigate the influence of context on the perception of `anger`. They argue that traditional associations between tones and attitudes are misleading and that contextual factors can neutralise the anger perception. A further study (cf. [Lefter et al. 2012]) investigates the role of modality information onto aggression annotation utilising three different settings: audio only, video only, and multimodal (audio plus video). They stated that for 46 % of their material the annotation of all three settings differs.

These considerations show that the labellers have to be experienced in emotional assessment and should be supported by suitable labelling methods. Furthermore, the results gained have to be secured by reliability measures (cf. Section 4.1.3).

Word Lists

A common method for assigning emotional labels is an Emotion Word List (EWL). Descriptive labels, usually not more than ten, are selected to describe the emotional state of an observed subject (see Table 4.1). These labels can be formed from counterparts like `positive` vs. `negative`, or designed for a specific task (e.g. aggression detection) [Devillers & Vasilescu 2004; Lee & Narayanan 2005; Lefter et al. 2012]. EWLs are utilised in several databases and comprise different emotional terms.

Table 4.1: Common word lists and related corpora.

Emotional Labels	Used in
negative, non-negative	*ACC* [Lee & Narayanan 2005]
angry, bored, doubtful, neutral	UAH [Callejas & López-Cózar 2008]
fear, negative, positive, neutral	SAFE [Clavel et al. 2006]
positive, neutral, negative	ISL Meeting Corpus [Burger et al. 2002]
Ekman's six basic emotions and neutral	DES [Engberg & Hansen 1996], emoDB [Burkhardt et al. 2005]
joyful, emphatic, suprised, ironic, helpless, touchy, angry, bored, motherese, reprimanding, rest	AIBO database [Batliner et al. 2004]

One disadvantage of EWLs is the missing relationship between labels. This makes it difficult for the labeller to give an evaluative assessment [Sacharin et al. 2012]. The labels and their meaning also have to be introduced to the labeller as the subjective interpretation can differ from labeller to labeller [Morris 1995]. Furthermore, the application of EWLs to other languages requires a complex task of translation and validation [Bradley & Lang 1994]. Another flaw is that the selection of emotional terms is mostly limited to a specific domain or selection of emotions. This can cause some emotional phenomena within the data to get lost or to be merged into one emotional term. This may later complicate the emotion recognition since the emotional characteristics of these merged phenomena also differ.

Geneva Emotion Wheel

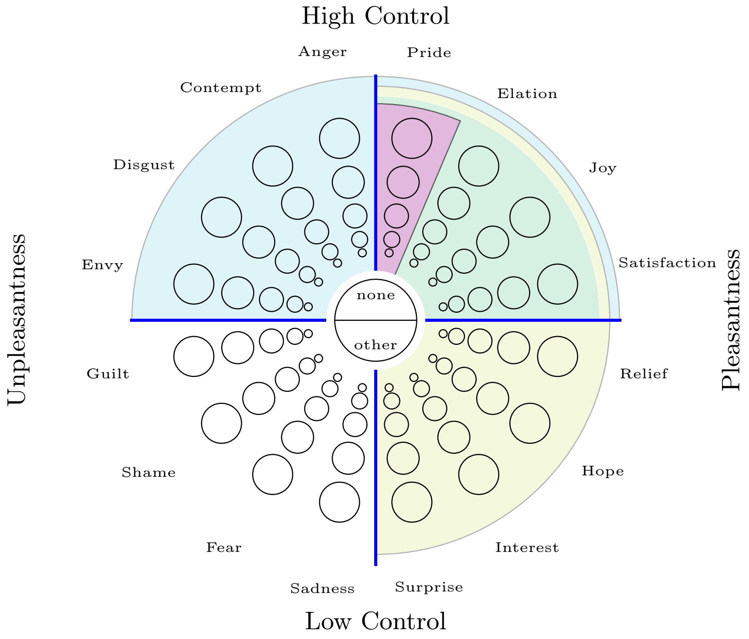


Figure 4.1: Geneva Emotion Wheel as introduced by Scherer (cf. [Siegert et al. 2014b]). The arrangement supports the labelling process: e.g. decision for *high control*, the ● semicircle, then decision for *pleasantness*, the ● semicircle. This defines the resulting quadrant (●). Than the labeller has to choose from four emotions and chooses *pride*, marked with ●.

A prominent solution to overcome the mentioned problems of EWL is the Geneva Wheel of Emotions (GEW) (cf. Figure 4.1) by Scherer [Scherer 2005b]. This assessment tool is highly related to Scherers' appraisal theory [Scherer 2001]. It is a theoretically derived and empirically tested instrument to measure emotional reactions to objects, events, and situations and consists of 16 emotion categories, called "emotion families", each with five degrees of intensity, arranged in a wheel shape in the control and pleasantness space. Additionally, the options `no emotion` and `neutral` are added to provide the rater with the opportunity to assign neutral or unspecific situations. This arrangement supports the labeller in assessing a single emotion family with a specific intensity by guiding him with the axes and quadrants (cf. [Sacharin et al. 2012]).

Unfortunately, the labelling effort using GEW is quite high since the labeller has to mark the emotion via a multi-step approach: 1) decide on the control axis to get the semicircle, 2) choose the value for pleasantness to get the quadrant, and 3) decide between the remaining four emotion families. This gets even more complex, when the intensity of an emotion should be assessed, too. A newer version of the GEW utilises 20 newly arranged emotion families, containing, for instance amusement, interest, regret, and disappointment (cf. [Scherer et al. 2013]).

A disadvantage of the GEW is its reference to `control` as second dimension instead of the strong physiologically related dimension of `arousal` [Grandjean et al. 2008]. Commonly the dimensions `pleasure` and `arousal` are used to describe human emotions [Grimm & Kroschel 2005; Russel 1980]. The dimension `control`, also called `dominance`, is an object of ongoing discussions (cf. Section 2.1), mostly because the studies rely on different methods and interpretations, varying from “not needful” [Russel & Mehrabian 1974; Yang et al. 2007] to “immanent” [Gehm & Scherer 1988] to distinguish certain emotions.

(Self)-Assessment Manikins

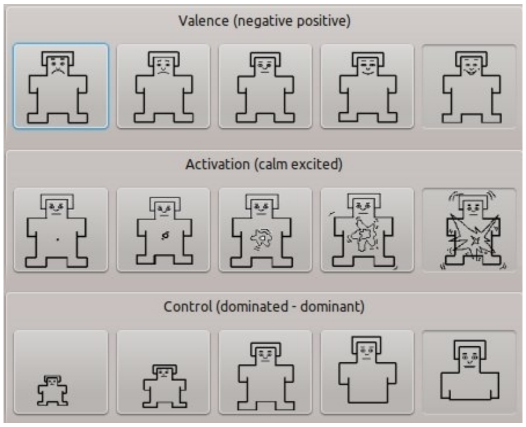

Valence (negative positive)
Activation (calm excited)
Control (dominated - dominant)

Figure 4.2: Five-scale Self-Assessment manikins, each row represents another dimension within the PAD-space [manikins after Lang 1980].

Having a verbal description of emotional affects can cause some challenges as well since the application for another language requires a translation and validation [Bradley & Lang 1994]. Furthermore, the relation between each literal label can differ from labeller to labeller. Thus, the relations differ not only from the subjective observation but also from the subjective interpretation of the verbal description [Morris 1995]. To address these issues, Lang invented a picture-oriented instrument to assess the `pleasure`, `arousal`, and `dominance` dimension directly [Lang 1980]. In their opinion, a dimensional representation supports the labellers to judge the relation between observed emotions much better than a literal transcription is able to. At the same time, it reduces the evaluation effort. For example, the Semantic Differential Scale uses 18 bipolar adjective pairs to generate judgements along three axes (cf. [Mehrabian 1970]), whereas the so-called Self Assessment Manikins (SAM) depict the same dimensions by 3 × 5 figures, see Figure 4.2. These figures depict the main characteristic for each dimension in changing intensity, for instance changing from a happy smiling manikin to a weeping, unhappy one to represent `pleasure`.

The granularity of the representation is adjustable and spans from five figures for each dimension [Morris & McMullen 1994] to a nine-point scale with intermediate steps between the figures [Ibáñez 2011]. This method has been used to assess different scenarios. It is also usable with labellers that are not “linguistically sophisticated”, like children (cf. [Morris 1995]). But resulting from the dimensional description, the ability to evaluate distinct or blended emotions is missing.

FEELTRACE

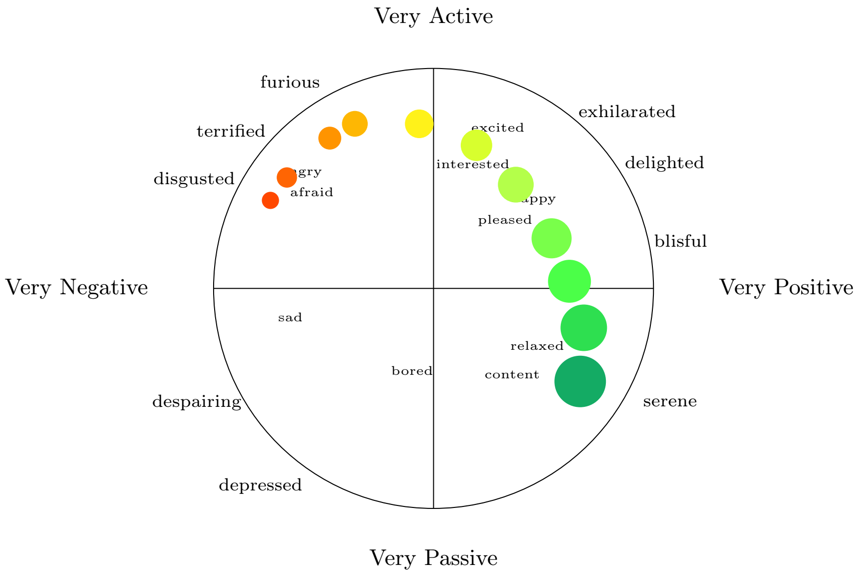


Figure 4.3: FEELTRACE as seen by a user. The color-scheme is derived from Plutchik’s wheel of emotions. Furthermore, “verbal landmarks” are visible [after Cowie et al. 2000].

An entirely different approach to assess emotional affects is introduced by FEELTRACE [Cowie et al. 2000]. This framework is designed to track the assessed affect over time, so that the emotional evolvement can be examined. These assessments are stored in a numeric format, which allows a statistical handling. FEELTRACE is based on the `arousal-valence` space and is circularly arranged, see Figure 4.3. This space can be seen as a variant of the `pleasure-arousal` space (cf. Section 2.1).

Using a mouse, the labeller can change the assessment by modifying the trajectory in this two-dimensional space. This trajectory represents the emotional change over time. To differentiate the actual cursor position from previous positions the older positions are gradually shrinking over time. To support the labeller, two further types of feedback are implemented (cf. [Cowie et al. 2000]). First "verbal landmarks" are added. Hereby, strong archetypal emotions associated with broader sectors are placed at the periphery and less extreme emotions, where a location within the circle is possible, are placed at these coordinates. Furthermore, the cursor is colour coded, utilising a colour scheme derived from Plutchik. It uses the colours red, green, yellow, and blue to indicate specific positions within the `activation-evaluation` space. A white pointer indicates the origin of the space (cf. [Cowie et al. 2000]).

This tool induces a high cognitive load on the labeller, as a real-time processing of the observations is needed. As it is indicated in [Koelstra et al. 2009], there is a gap between observation and verbalisation. Thus, the generated traces cannot be directly mapped onto the underlying material. Furthermore, it is necessary for the labeller to assess the whole material, as the contextual influence for a correct label is much higher than for other labelling tools (cf. [Cowie et al. 2000]).

In 2010, Cowie & McKeown presented GTrace, the successor of FEELTRACE (cf. [Cowie & McKeown 2010]). This tool allows to customise the set of used emotional scales. It contains a set of 50 different scales and was used to label the SEMAINE database (cf. [McKeown et al. 2010]). This tool has the advantage of decoupling the two-dimensional space into separate axes. Thus, the labeller can concentrate solely on one observation. But this advantage is associated with a further increased processing time, as for each scale the material has to be processed completely.

Although this tool allows to handle intermediate emotional states and to capture long term and short term temporal progress, the evaluation of the resulting labels is problematic, since each labeller produces a constant track with a step width of 0.02 s on the time axis. The minimal resolution of the emotional values is 0.0003 in the interval of $[-1, 1]$. As the relations of observed changes are very individual, only a trend can be extracted, rather than a distinct point within the emotion space (see Figure 4.4 for an example of FEELTRACE/GTrace assessments).

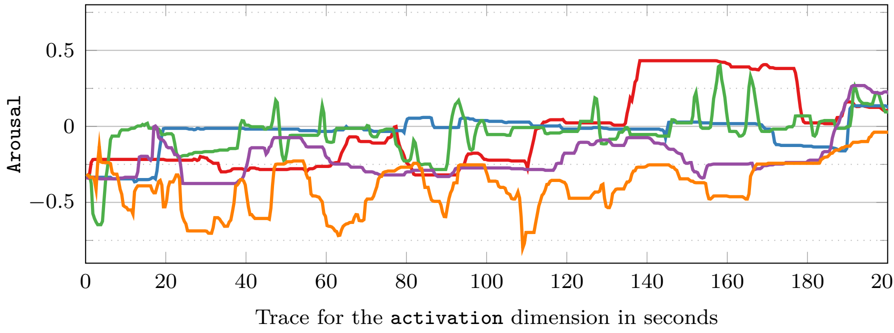


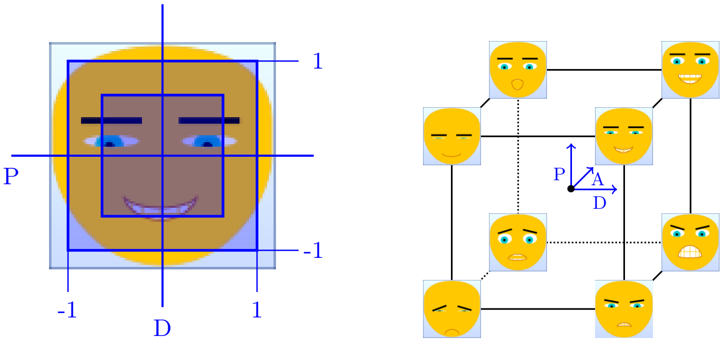
Figure 4.4: Example FEELTRACE/GTrace trace plot from five labellers for the female speaker 1 trace 29 from SAL (cf. [McKeown et al. 2012]).

Further Methods

The Product Emotion Measurement Tool (PrEmo) is designed to measure typical emotions related to a commercial product [Desmet et al. 2007]. It uses different animated cartoon characters to express 14 emotional categories, seven positive (`inspiration`, `desire`, `satisfaction`, `pleasant surprise`, `fascination`, `amusement`, and `admiration`) and seven negative (`disgust`, `indignancy`, `contempt`, `disappointment`, `dissatisfaction`, `boredom`, and `unpleasant surprise`). Each animation is about 1 s and consists of a gesticulating character with sound to increase the comprehensibility of each emotion. But PrEmo "focuses on the emotions that constitute a wow-experience" [Desmet et al. 2007]. So its usefulness for the emotional assessment of HCI is questionable.

Another tool, introduced by Broekens & Brinkman is related to SAM and measures emotions within the PAD-space, by using an interactive mouse-controlled button. The resulting label is extracted from the x-y coordinates of the cursor within the AffectButton window, which I will explain in the following. As a three-dimensional space has to be mapped onto a two dimensional plane, the authors of the AffectButton neglected the independence of `arousal` from the other dimensions in some cases. Therefore, the button consists of three parts, a center part, an inner border and an outer border (cf. Figure 4.5(a)). In the center only `P` and `D` are changed, with the center point being $[0, 0]$. In the inner border the user also controls A, which is interpolated from -1.0 (start of inner border) to 1.0 (start of outer border) based on its distance to the outer border. The following outer border is only added to allow the expression of extreme affects, without moving outside the AffectButton. In this case, `A` is always 1.0 while `P` and `D` are mapped to their nearest point on the inner border [Broekens & Brinkman 2009]. The three different parts of the button are not visualised, so that a longer training phase is needed, before a rater can use that method.

The resulting emotional expressions are depicted by different positions of eyebrows, eye and mouth. They consist of ten prototypical expressions for each extreme case within the PAD-space (e.g., -1,1,-1 for afraid), the neutral case, which represents the centre of the emotion-space, and transitions between them, see Figure 4.5(b).



(a) Mapping of `pleasure`, `arousal`, and `dominance` axis onto AffectButton [after Broekens & Brinkman 2009].

(b) The AffectButton extreme cases and their location within the `PAD`-space [after Broekens & Brinkman 2009].

Figure 4.5: The AffectButton graphical labelling method

By this arrangement the labeller can assess transitions of emotional observations. This method is validated by different experiments assessing emotional words or emotional annotation of music with a following questionnaire (cf. [Broekens & Brinkman 2013]). But, due to its design to manually adjust the intensity of the perceived emotion and the difficulty to reproduce former x-y coordinates, the effort using this method is quite high. Furthermore, the P and D axis are mapped onto the button quite unintuitively and the labeller cannot track back his assessment transitions as the AffectButton only depicts the actual assessment and not the trace of former assessments, as in FEELTRACE. Furthermore, the advantage of a non-verbal scale that is also usable with not "linguistically sophisticated" labellers gets lost since this tool requires a lot of explanation beforehand.

In clinical trials questionnaires covering multiple scales are mostly used to assess affective dimensions. The PANAS measures the current feeling using a verbal self-report. It measures the two higher order affects negative (NA) and positive affectivity (PA) (cf. [Watson et al. 1988]) Each affect comprises ten mood items, for instance attentive and enthusiastic for PA and distressed or nervous for NA utilising two five-point scales, each. This method is reliable and valid, but [Crawford & Henry 2004] rejected a complete independence of PA and NA. The 26-item scale Berlin Everyday Language Mood Inventory (BELMI) [Schimmack 1997] is used to assess the current mood. The 5-point Differential Emotions Scale (Version 4) (DES-IV) [Izard et al. 1993] can be used to distinguish ten emotional terms. The 18-items bipolar Semantic Differential Scale is used to assess emotions within the three dimensions `evaluation`, `potency`, and `activity` [Mehrabian & Russell 1974]. All these reported methods are pursued in the form of self-reports and using lexical items to measure specific emotional scales. They support the self-rating by forcing the subject to reflect on the current situation. But they cannot be utilised for rating observations of other subjects.

4.1.3 Calculating the Reliability

As I have stated in Section 2.2.1, emotions are very subjective. When applying annotation methods (cf. Section 4.1.2), human coders, also called raters or annotators, subjectively judge the emotional information of the data. Thus, an objective measure is needed. Applying such measures, other researchers can make a reliable judgement of the research. Additionally this measure should also allow a statement on the validity of the utilised labelling scheme, where reliability is one prerequisite [Artstein & Poesio 2008]. This means the gathered annotations should provide a measure that allows a comparison with other investigations.

For this purpose, Inter-Rater Reliability seems to be a good measure [Carletta 1996; Artstein & Poesio 2008]. It determines the extent to which two or more raters obtain the same result measuring a certain object [Kraemer 2008]. In contrast, the Intra-Rater Reliability compares the variation of the assessment, which is completed by the same rater on two or more occasions. Here the self-consistency of the rater's subsequent labellings is in the focus [Gwet 2008a].

Kappa-like statistics are the most common used measures for assessing agreement on categorial classes showing that independent coders agree to a determined extent on the categories assigned to the samples. [Carletta 1996] discussed the kappa statistic as a general measure for quality of a labelling that fulfils the requirement of a reliable judgement for linguistic dialogue annotation. There are several variants of kappa-like coefficients in the literature, whose advantages and disadvantages are object of multiple discussions, especially in medical research, e.g. [Berry 1992; Kraemer 1980; Soeken & Prescott 1986]. Galton was the first who mentioned a kappa-like statistic (cf. [Galton 1892]). All follow the general formula presented in Eq. 4.1, where A_o denotes the observed agreement and A_e denotes the expected agreement.

$$\text{reliability} = \frac{A_o - A_e}{1 - A_e} \tag{4.1}$$

The term $A_o - A_e$ is used to determine the actually achieved degree of agreement above chance. Whereas, the term $1 - A_e$ defines the degree of agreement that is generally attainable above chance. So the ratio between both terms determines the proportion of the actual agreement beyond chance.

The observed agreement A_o is similar to the "percent agreement" (cf. [Holsti 1969]). Therefore, the agreement value agr_i is defined for each item i, to denote agreement with 1 and disagreement with 0 (cf. Eq. 4.2). Afterwards, A_o is defined as the arithmetic mean of the agreement value agr_i for all items $i \in I$:

$$\text{agr}_i = \begin{cases} 1 & \text{if the two coders assign } i \text{ to the same category} \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

$$A_o = \frac{1}{I} \sum_{i=1}^{I} agr_i \tag{4.3}$$

The expected agreement A_e is defined by Artstein & Poesio as the probability of the raters r_1 and r_2 agreeing on any category c. Eq. 4.4 presents A_e for a general two raters' case. Hereby the calculation is based on the independence assumption that the raters assigning the labels acting independently. Thus, the chance of r_1 and r_2 agreeing on any given category c is expressed as the produkt of the chance of each of them assigning an item to that category:

$$A_e = \sum_{c \in C} P(c|r_1) \cdot P(c|r_2) \tag{4.4}$$

Common Kappa-Like Coefficients

The difference between the individual kappa-like coefficients is the particular definition of the expected agreement, which ultimately falls into two categories: 1) using a global probability distribution for the expected agreement of all raters or 2) using an individual probability distribution for each rater. Further differences are the number of supported raters and the possibility to apply distance metrics. [Artstein & Poesio 2008] visualised a coefficient cube to demonstrate the relationship between the different kappa-like coefficients (cf. Figure 4.6)[9]. The most common coefficients will be presented briefly in the following.

[8] The independence assumption has been subject of much critisism, as pointed out in [Powers 2012].

[9] The coefficient in the lower left rear corner is not defined. But as this generalises Scott's π only along the application of weighting, it is only of theoretical interest. This case can be easily adopted by Krippendorff's α_K.

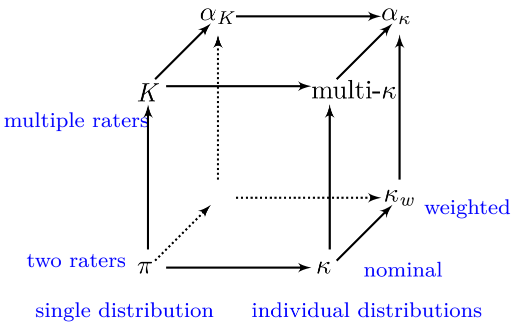


Figure 4.6: Generalising π along three dimensions, according to [Artstein & Poesio 2008]. The depicted abbreviations are discussed in the text.

Nominal Agreement Coefficients: π, κ, K, and multi-κ To calculate the expected agreement Scott assumed that the raters have the same distribution of responses $P(c|r_1) = P(c|r_2) = P(c)$, which is the ratio of the total number of assignments $\mathbf{n}_c$ to category c by both raters r_1 and r_2 and the overall number of assignments, which for the two-coders case is twice the number of items I (cf. Eq. 4.5, [Scott 1955]).

$$A_e^\pi = \sum_{c \in C} \hat{P}(c) \cdot \hat{P}(c) = \frac{1}{4I^2} \sum_{c \in C} \mathbf{n}_c^2 \tag{4.5}$$

Cohen moved away from Scott's assumption that raters have the same response-distribution [Artstein & Poesio 2008]. Instead, he measures the individual proportion $P(c|r_i)$ for each rater r_i assigning items to a category [Cohen 1960]. This individual probability is estimated by $\mathbf{n}_{rc}$ the number of assignments to a category by a coder divided by the number of items I. Cohens A_e can be formulated as the sum of the joint probabilities of each rater providing the assignment $\mathbf{n}_{rc}$ independently.

$$A_e^\kappa = \sum_{c \in C} \hat{P}(c|r_1) \cdot \hat{P}(c|r_2) = \frac{1}{I^2} \sum_{c \in C} \mathbf{n}_{r_1 c} \mathbf{n}_{r_2 c} \tag{4.6}$$

Scott's π (cf. Eq. 4.5) and Cohen's κ (cf. Eq. 4.6) are only for a two coders' case, but for emotional annotation this is usually not considered as a sufficient number of raters (cf. Section 4.1.2). Thus, a multi-coder coefficient is needed.

Fleiss' K (cf. [Fleiss 1971]) is able to calculate the degree of agreement for more than two coders[10]. In order to accomplish this, the observed agreement A_o cannot

[10] Fleiss itself called the coefficient κ, which led to much confusion. Artstein & Poesio called it multi-π, seeing it as an extension of Scott's π, as also a single probability distribution for all raters is used. In contrast, Siegel & Castellan called it K, since Fleiss himself does not span the link to π.

be defined as the percentage of items on which the observers agree. In a multi-coder scenario items may exist on which only some coders agree. Thus, Fleiss defined the amount of agreement on an item as the proportion of agreement on judgement pairs given the total number of judgement pairs for each item (cf. Eq. 4.7). A detailed description for the calculation can be found in [Fleiss 1971] and [Artstein & Poesio 2008]. Let $\mathbf{n}_{ic}$ be the number of coders who assigned item i to category c and R be the total number of raters, the pairwise agreement A_o^K can be estimated as follows:

$$A_o^K = \frac{1}{I} \sum_{i \in I} P_i \qquad \text{where} \quad P_i = \frac{1}{IR(R-1)} \left(\sum_{i \in I} \sum_{c \in C} \mathbf{n}_{ic}(\mathbf{n}_{ic} - 1) \right) \tag{4.7}$$

The same method, calculating the pairwise agreement, is also used to estimate the expected agreement [Artstein & Poesio 2008]. Fleiss further assumes that a single probability distribution can be used. For this, Artstein & Poesio span the link to Scott's π (cf. Figure 4.6). The probability that two random raters assign an item to a category can then be expressed by the ratio of the joint probability and the number of items I multiplied with the number of categories (cf. Eq. 4.8). In this case, $\mathbf{n}_c$ expresses the total number of items assigned by all raters to category c.

$$A_e^K = \sum_{c \in C} (\hat{P}(c))^2 \qquad \text{where} \quad \hat{P}(c) = \frac{1}{IR} \mathbf{n}_c \tag{4.8}$$

Another coefficient, also able to deal with multi-coder agreement, was suggested by [Davies & Fleiss 1982]. This coefficient represents a generalization of Cohen's κ and utilises separate distributions for every annotator. It is called multi-κ in the literature. The calculation of A_o follows Fleiss' definition, but A_e is expressed by the individual probability distributions of each rater. An implementation of such an expected agreement can be found in [Artstein & Poesio 2008].

Weighted Agreement Coefficients: α_K, κ_w and α_κ The former presented inter-rater reliability measures are only suitable for nominal values, where the differences between all categories have an equal effect on the reliability. But especially for affective or emotional observations, disagreements are not all alike. Even for simple categorial emotions the disagreement between an emotional valence of `positive` (`arousal`) and `negative` (`arousal`), for instance, is more serious than a disagreement between `positive` (`arousal`) and `neutral` (`arousal`). For such tasks, where reliability is determined by measuring agreement, an allowance for degrees of disagreement becomes essential. Under these circumstances, the nominal kappa statistics attain low values, which does not necessarily reflect the true reliability.

Hence, coefficients taking into account the degree of disagreements are needed, since a distance measure between the given labels is applied. The resulting agreement is given by the complementary event (cf. Eq. 4.9). In this work three reliability measures are discussed, namely Krippendorff's α_K, Cohen's κ_w, and Artstein & Poesio's α_κ. To obtain comparability, the terms D_o denoting the observed disagreement and D_e denoting the expected disagreement are used. The disagreement-reliability measures are generally defined as follows:

$$\alpha, \kappa_w = 1 - \frac{D_o}{D_e} \tag{4.9}$$

Given $D_o = 1 - A_0$ and $D_e = 1 - D_e$, the reliability coefficients α_K, κ_w and α_κ are equivalent to π and κ using the agreement formulation of Eq. 4.1 on page 55.

A distance metric is needed in order to be able to specify the different disagreements. In [Krippendorff 2012] several distance metrics for nominal, ordinal, interval, and ratio data are presented[11]. Generally, $d(c_a, c_b)$ is a function that maps category pairs to non-negative real numbers that specify the quantity of unlikeness between these categories. The appropriate distance metric is determined by the nature of the categories for an individual coding task. For the given introduction example the position can be assigned as follows: `positive` 1, `neutral` 0, and `negative` -1. Thus, when using a Euclidean distance, the disagreement between `positive` and `negative` is weighted with 2. Different metrics for emotional labelling are presented in Section 6.1. [Artstein & Poesio 2008] define two constraints that a general distance metric should fulfil:

(1) For every category $c \in C$, $d(c_a, c_a) = 0$.
(2) For every two categories $c_a, c_b \in C$, $d(c_a, c_b) = d(c_b, c_a)$.

To calculate the observed disagreement D_o an average disagreement value is defined, where in contrast to the observed agreement all disagreement values for a specific class are considered[12]. The calculation is done over pairs of judgement. One disagreement pair $\mathbf{n}_{ic_a}\mathbf{n}_{ic_b}$ for one item i can be considered as the number of raters coding the item either as class c_a or c_b multiplied with the distance $d(c_b, c_a)$ between these classes.

$$\text{disagr}_i = \sum_{j=1}^{C} \sum_{l=1}^{C} \mathbf{n}_{ic_j}\mathbf{n}_{ic_l} d(c_j, c_l) \tag{4.10}$$

[11] By using distance metrics Krippendorff's α_K comprises several known reliability coefficients, like Scott's π for two-rater nominal data and Pearson's intraclass-correlation coefficient for two-rater interval data [Hayes & Krippendorff 2007].

[12] Hereby, it is not necessary to exclude the agreement pairs, following Artstein & Poesio's definition of the distance for agreement pairs $d = 0$.

The overall observed disagreement is the arithmetic mean of these pairs of judgement (disagr_i) over all items I and the number of all ordered judgement pairs $R(R-1)$:

$$D_o = \frac{1}{IR(R-1)} \sum_i \text{disagr}_{i \in I} \tag{4.11}$$

As already stated for the nominal versions of the IRR, α_K and κ_w also differ mostly in the choice of the definition for the expected disagreement D_e. Krippendorff assumes that the expected disagreement is the result of a single probability distribution [Krippendorff 2012], whereas in Cohen's weighted kappa an individual probability distribution is assumed [Cohen 1968]. The overall probability $\hat{P}^{\alpha}(c)$ (cf. Eq. 4.12) for α_K is defined as $\mathbf{n}_c$, the total number of assignments of an item to category c by all raters, divided by the overall number of assignments IR. The overall probability $\hat{P}^{\kappa_w}(c|r)$ for κ_w is defined as the probability of the number of assignments $\mathbf{n}_{rc}$ of an item to category c by rater r divided by the number of items I (cf. Eq. 4.13).

$$\hat{P}^{\alpha}(c) = \frac{1}{IR}\mathbf{n}_c \tag{4.12}$$

$$\hat{P}^{\kappa_w}(c|r) = \frac{1}{I}\mathbf{n}_{rc} \tag{4.13}$$

Both coefficients interpret the expected disagreement as a distinct probability distribution for each rater. D_e is defined as the mean of the distance between categories weighted by these distinct probabilities for all category pairs. Artstein & Poesio state that Krippendorff used a slightly different definition for the expected disagreement [Artstein & Poesio 2008]. Krippendorff defines it as the mean of distances without any regard to items. Hence, he normalises with $IR(IR-1)$ instead of $(IR)^2$ (cf. Eq. 4.14). Cohen's κ_w is restricted to two coders, as shown in Eq. 4.15. Through division by the maximum weight $\mathbf{d}_{max}$ the observed disagreement is normalised to the interval $[0,1]$.

$$D_e^{\alpha} = \frac{1}{IR(IR-1)} \sum_{j=1}^{C} \sum_{l=1}^{C} \mathbf{n}_{c_j} \mathbf{n}_{c_l} \mathbf{d}_{c_j c_l} \tag{4.14}$$

$$D_e^{\kappa_w} = \frac{1}{\mathbf{d}_{max}} \frac{1}{I^2} \sum_{j=1}^{C} \sum_{l=1}^{C} \mathbf{n}_{r_1 c_j} \mathbf{n}_{r_2 c_l} \mathbf{d}_{c_j c_l} \tag{4.15}$$

Artstein & Poesio proposed an additional agreement coefficient, which can be applied for multiple coders and calculates the expected agreement utilising an individual probability distribution for each coder. The observed disagreement is equivalent to α_K and κ_w (cf. Eq. 4.11). The expected disagreements distinguish the individual distributions for each pair of coders. Hence, the expected disagreement infers the number of items assigned to a category by a specific rater $\mathbf{n}_{rc}$ instead of the number of items assigned to a category by all raters $\mathbf{n}_c$ (cf. Eq. 4.16).

$$D_e^{\alpha_\kappa} = \frac{1}{I^2\binom{R}{2}} \sum_{j=1}^{C}\sum_{l=1}^{C}\sum_{m=1}^{R-1}\sum_{n=m+1}^{R} \mathbf{n}_{r_m c_j}\mathbf{n}_{r_n c_l}\mathbf{d}_{c_j c_l} \tag{4.16}$$

Interpretation of Reliability Measures

Although kappa statistics are often used to state the reliability, they have some flaws that sometimes makes the calculation of the measurement inappropriate. Feinstein & Cicchetti addressed two paradoxa of kappa calculation (cf. [Feinstein & Cicchetti 1990; Cicchetti & Feinstein 1990]).

The first paradox occurs when a relatively high value of the observed agreement A_o is not accompanied with a high inter-rater reliability. Kraemer justifies this by the fact that the proportion of agreement is not equally distributed over all classes, which also simultaneously enlarged the expected agreement A_e. Kraemer first identified this problem as the "prevalence problem": The tendency for raters to identify one class more often due to highly skewed events in the data [Kraemer 1979]. Artstein & Poesio further stated that chance-corrected coefficients are sensitive to agreements on rare categories. Thus, they suggest in cases where the reliability is low despite a high observed agreement have been found to report A_o, too (cf. [Artstein & Poesio 2008]).

The second paradox can occur for the counterpart of the prevalence problem, called the "bias problem" [Feinstein & Cicchetti 1990]. The bias is the extent to which the raters disagree: the larger the bias, the higher the resulting kappa value, despite the value of the observed agreement. However the second paradox is appreciated to be less severe, because observed agreement and expected agreement are not independent, as "both are gathered from the same underlying ratings" [Artstein & Poesio 2008].

However, the choice of the coefficient is dependent on the desired information. To measure the reliability of the used coding, the single-distributed coefficients (e.g. π, K or α_K) should be used. Independent-distributed coefficients (e.g. κ or κ_w) are appropriate to measure data correctness [Artstein & Poesio 2008]. These considerations have not such a strong effect, if more than two annotators are used [Artstein & Poesio 2008]. In this case, the impact of the variance of the raters' distribution decreases as the number of labellers grows, and becomes more similar to random noise. The numerical difference is also very small for a high agreement [Artstein & Poesio 2008].

The kappa statistics presented examine the level of concordance that is archived in contrast to reachable agreement through "random estimation". Typical values for are between 1 (observed agreement = 1) and $-A_e/(1-A_e)$ (no observed agreement), with a value of 0 signifying chance agreement ($A_o = A_e$), see Eq. 4.17.

$$-A_e/(1-A_e) \leq \kappa \leq 1 \tag{4.17}$$

There are several interpretations of kappa values. In medical diagnosis, where kappa-like statistics are used as well, the interpretation suggested by [Landis & Koch 1977] is used. This interpretation is similar to that used for correlation coefficients and seen as an appropriate interpretation, as values above 0.4 are seen as *adequate* [Artstein & Poesio 2008]. The interpretation by [Altman 1991] also has the same origin, he just denoted every value lower than 0.2 as *poor*. In contrast, Fleiss et al. and Krippendorff proposed different interpretations of agreement, directly related to content analysis. In this area of research, a more stringent convention is utilised, as the assessment of content analysis-categories leaves less room for interpretations or subjective evaluation. Fleiss et al. states that values greater than 0.75 depict a very good agreement and values below 0.4 a poor agreement. Values in between are *fair to good*. Hereby the author of [Krippendorff 2012] express an even stronger interpretation than in [Fleiss et al. 2003]. Krippendorff considers a reliability higher than 0.75 or 0.8 as *good*, values between 0.67 and 0.8 are *good* and all values below 0.67 are *poor*.

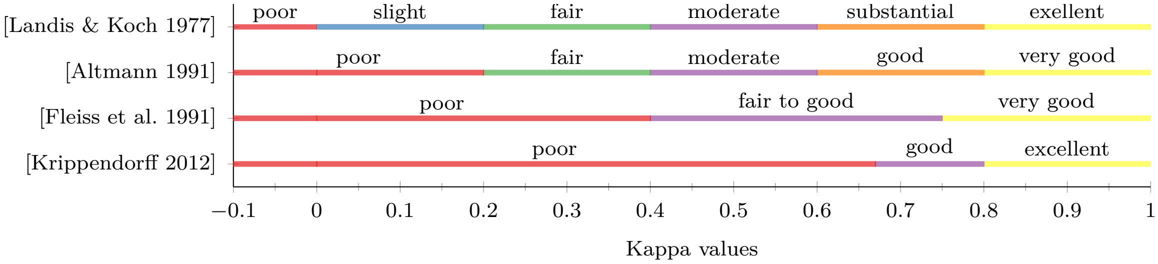


Figure 4.7: Comparison of different agreement interpretations of kappa-like coefficients utilised in medical diagnosis and content analysis.

These differences in the interpretation intervals make it hard to examine the values, see Figure 4.7 for a comparison. Therefore, besides the interpretation value, the pure kappa coefficient should also be given. For content analysis mostly the interpretation of [Krippendorff 2012] is used, but for the subjective emotional annotation no preferred interpretation exists so far. In my thesis, I will thus consider the interpretation suggested by [Landis & Koch 1977], as they offer the most divisions, which allows a more graduated statement.

4.2 Features

In this section, I will give an overview of all features utilised in this thesis. As I mentioned in Section 2.2, the psychological research has made some assumptions about acoustical features involved in the emotional response patterns, but characterisation is still on a descriptive level. Therefore, pattern recognition researchers, dealing with automatic emotion recognition from speech, started with features they could extract robustly and investigated their usability for emotion recognition. Researchers utilised well known acoustic features used for automatic speech recognition and speaker verification [Kinnunen & Li 2010; Schuller et al. 2011c; Pieraccini 2012]. Most of the utilised features are based on a model representation of human speech production, see Figure 4.8. For speech production, three systems must be taken into account: the respiratory system, the vocal system, and the resonance system.

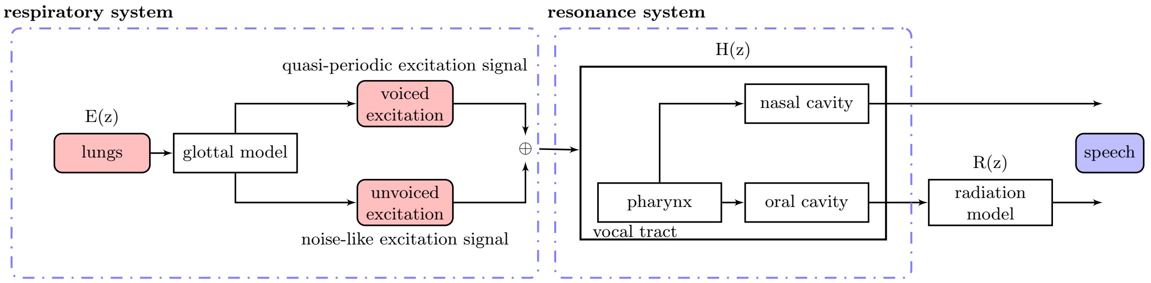


Figure 4.8: Acoustic speech production model [after Fant 1960]. The red boxes denotes the possible input and the blue box denotes the produced speech signal.

The lungs in the respiratory system generate an airflow, which is pressed through the glottis. If the vocal chords are tensed, a periodic signal with fixed frequency is produced. In this case a quasi-periodic excitation signal is generated. Otherwise, a white noise-like excitation signal is produced. Thus, in the respiratory system either a voiced or an unvoiced sound is produced. These sounds are expand into the vocal tract of the resonance system. The vocal tract itself consist of pharynx, nasal cavity, and oral cavity. Its shape can be changed by several muscles resulting in different transmission properties to articulate different tones. Finally these tones are emitted through the mouth's radiation model and the nose (cf. [Wendemuth 2004]).

The different systems can be modelled as filters with specific transfer functions (e.g. $E(z)$, $H(z)$, and $R(z)$). Based on the findings of Fant, the vocal tract can be modelled as a series of tubes with similar length but different areas, [Fant 1960]. Within a short-time range the filters have invariant properties, thus enabling an estimation of the filter parameters within this short time range. Fant called this technical description the “source-filter model”. The resonance frequencies of the vocal tract are commonly called format-frequencies or “formants” shortly denoted as F_1, F_2, or F_3, for instance. For a more detailed explanation of the respiratory systems and its characteristics I refer the reader to [Benesty et al. 2008; Schukat-Talamazzini 1995; Wendemuth 2004].

The human speech production is controlled by both Autonomic Nervous System (ANS) and Somatic Nervous System (SNS). The influence of these systems is normally ignored for modelling speech production. But, as the they are affected by appraisals (cf. Section 2.2), the “emotional state” is decoded in the acoustics, too. For instance, the resonance of the vocal tract is influenced by the production of saliva and mucus. Their production is regulated by parasympathic and sympathic activity. Johnstone et al. argue that the evaluation of specific (emotional) events can change these activity regardless of their usually task [Johnstone et al. 2001] and contribute to an increased production of salvia and mucus. This changes the vocal tract resonance. Thus, the use of features related to speech recognition is also promising for affect recognition, even if the deeper interrelationship is still unclear.

The acoustic characteristics are distinguished in short-term segmental acoustics also called LLDs, often carrying linguistic information, and longer-term supra-segmental features, carrying a mixture of linguistic, paralinguistic, and non-linguistic information. Also spectral, prosodic and paralinguistic features are distinguished (cf. [Schuller et al. 2010a]). But as the concrete assignment is not always clear, I distinguish between short-term segmental and longer term supra-segmental features.

4.2.1 Short-Term Segmental Acoustic Features

Most feature extraction methods for speech and emotion recognition are based on the analysis of the short-time spectrum of an acoustic signal, since these are allocated to specific tones as well as emotional reactions (cf. [Johnstone et al. 2001]). The resulting features are denoted as “spectral features” and are used to recover the resonance frequencies generated by the vocal tract and motivated by the human auditory system (cf. [Honda 2008]). By way of introduction, I refer to the following books [Benesty et al. 2008; Young et al. 2006; Rabiner & Juang 1993; Wendemuth 2004] that also serve as foundation of this section.

The activity of single auditory nerves is depending on the allocated frequency bands. Furthermore, the resonance frequencies, influenced by the different shapes of the vocal tract, are extractable easily within the spectral domain. As the filter parameters, which are responsible for different tones, are invariant within a short time range of 15 ms to 25 ms, this period serves as segment for short-term segmental features [Mporas et al. 2007]. It is still unclear, whether the same time range applies for emotional characteristics as well (cf. [Paliwal & Rao 1982]), but in several experiments this range showed promising recognition results.

Windowing

To be able to process the short-time spectrum of a speech signal, the analoge signal $s(t)$ is transferred into a discrete signal $s(n)$, by sampling with the sampling frequency f_s. By further applying a quantization the analoge value ranges of the signal are merged into several single discrete values. The Fast Fourier Transformation (FFT) is applied, to afterwards transform the signal from time-domain to spectral domain. The Discrete Fourier Transformation (DFT) as well as the FFT only produces correct results when they are applied to a periodic function. A speech signal is obviously not a periodic function, but its characteristics are stable within a short period of time. Thus, by windowing the original signal and periodically continuing the windowed signal, the application of a FFT ensures correct values. The window length is in the range of 15 ms to 25 ms. As the extracted and continued window causes jump discontinuities at the window edges, leading to high frequencies after transformation, special window functions are used for windowing the speech signal (cf. [Young et al. 2006]). In speech recognition as well as in emotion recognition a Hamming window is typically used, see Eq. 4.18, where n is the input's index and N the length of the window:

$$\omega(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), \qquad 0 \leq n \leq N \tag{4.18}$$

The resulting windowed signal value $\tilde{s}_k(n)$ is then calculated by multiplying the speech signal $s(n)$ with the window $\omega(n)$. This approach results in a sequence of weighted time-discrete signals (cf. Eq. 4.19), each representing one short-term segment (frame), which is continuously moved with a certain number of time steps.

$$\tilde{s}_k(n) = \omega(n) \cdot s(k\tau_0 + n) \tag{4.19}$$

Disposing of Glottis Waveform and Lip Radiation

Furthermore, a simplified model of acoustic speech production is used which neglects the lip radiation and the glottis waveform. Both lip radiation, which normally causes a decrease of the magnitude of higher frequencies, and glottis waveform, which changes the phonotation type, have a substantial influence on the perceived speech signal [Chasaide & Gobl 1993]. To compensate these effects, a further filtering is applied to the speech-signal before the coefficients are calculated (cf. Eq. 4.20). The parameter μ of this first-order high pass filter is normally set in the range of 0.9 to 0.99.

$$s(n) = s(n) - \mu \cdot s(n-1) \tag{4.20}$$

Reducing Channel Influence

To reduce the channel influence two methods can be used, Cepstral Mean Subtraction (CMS) and RelAtive SpecTrAl (RASTA)-filtering. CMS works in the log-cepstral domain. In this domain, the channel transfer-function becomes a simple addition in the same way as the excitation. As it is supposed that these channel changes are much slower than the changing of the phonetic speech content itself, the long-term average of the cepstrum is subtracted from the cepstrum of each windowed frame [Atal 1974]. Environmental noise and channel changes are thereby eliminated. This method is applied after the cepstrum is calculated for several associated segments.

The RASTA-filter (cf. Eq. 4.21) removes slow and very fast spectral changes which do not appear in natural speech or are not needed for ASR [Hermansky & Morgan 1994]. The background noise results in slowly varying spectral elements, the speaker-generated high-frequency modulations convey little information. Thus, by applying a band-pass filter prior to the computation of coefficients, very low spectral components are suppressed and very high spectral components are normalised across the speakers:

$$H(z) = 0.1 * \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.982z^{-1})} \tag{4.21}$$

Although RASTA-filtering is commonly applied when using PLP-coefficients [Hermansky et al. 1992], it can be used for MFCCs, as well (cf. [Kockmann et al. 2011]).

Both techniques minimise the influence of channel characteristics. As recommendation these methods might be applied to speech material, that where recorded under varying conditions. In particular, RASTA-filtering improves the recognition, when the environmental and speech properties are quite different [Veth & Boves 2003].

Spectral (Acoustic) Features

Due to the vocal tract's resonance properties, specific frequency ranges are increased with respect to other frequencies. The frequency ranges with the highest relative amplification are denoted as "Formants" and are manifested as peaks in the spectral domain. Thus, a signal analysis within the spectral domain is able to identify these peaks [Gold & Morgan 2000]. To extract the underlying vocal tract parameters, the resonance frequencies have to be uncoupled from the excitation frequency. Two methods are commonly used, spectral deconvolution and linear prediction.

Mel-Frequency Cepstral Coefficients Spectral deconvolution separates the impulse response $h(n)$ from the excitation $u(n)$. According to Fant's source-filter model the response signal corresponds to the resonance frequencies constituted by the vocal tract while the glottis generates the excitation (cf. Eq. 4.22). The signal is transformed into spectral domain by an FFT $\mathcal{F}(\cdot)$ (cf. Eq. 4.23). But even in spectral domain, the speech signal spectrum is formed by a multiplication of excitation and response signal spectrum. Thus, in spectral deconvolution a logarithmic function is then applied to the signal to convert the multiplication into a summation (cf. Eq. 4.24). Afterwards the Discrete Cosine Transformation (DCT) $\mathcal{F}^{-1}(\cdot)$ is applied to the magnitude spectrum, to obtain the inverse transformation (cf. Eq. 4.25). As the utilised logarithmic function is still included, the inverse transformation does not lead back to the time domain, instead it leads to an artificial domain, the "cepstrum" with its unit "quefrency". Both are neologisms arising from "spectrum" and "frequency" (cf. [Bogert et al. 1963]).

$$
\begin{aligned}
s(n) &= u(n) * h(n) && (4.22)\\
\mathcal{F}\{s(n)\} &= \mathcal{F}\{u(n)\} \cdot \mathcal{F}\{h(n)\} && (4.23)\\
\log \mathcal{F}\{s(n)\} &= \log \mathcal{F}\{u(n)\} + \log \mathcal{F}\{h(n)\} && (4.24)\\
\mathcal{F}^{-1}\{\log |\mathcal{F}\{s(n)\}|\} &= \mathcal{F}^{-1}\{\log |\mathcal{F}\{u(n)\}|\} + \mathcal{F}^{-1}\{\log |\mathcal{F}\{h(n)\}|\} && (4.25)
\end{aligned}
$$

The excitation frequency can be found as a cepstral peak at the inverse excitation frequency. This peak can be filtered out very easily, this method is called "liftering". The remaining cepstral peaks describe the resonance quefrencies of the vocal tract.

Although the cepstral analysis is designed to deconvolute the vocal tract resonances from excitation, which means voiced speech, it can also be used for unvoiced speech. In both cases the cepstral analysis creates a smoothed signal, whose peaks are used as cepstral coefficients [Wendemuth 2004]. The strong relation between the spectral peaks and the formant frequencies has already been examined by Pols et al. They state that the first two principal components of the spectrum result in a pattern similar to the vowel triangle of F_1 and F_2 (cf. [Pols et al. 1969]). To furthermore incorporate the auditory perception of humans, the Mel frequency warping is applied beforehand to the transformation of the signal into spectral domain (cf. [Stevens et al. 1937]). This warping corrects the human loudness perception of the frequencies f (cf. Eq. 4.26). The made-up word "Mel" comes from melody, to indicate that this warping is based on pitch comparisons.

$$mel(f) = 2\,595\,Hz \cdot \log_{10}\left(1 + \frac{f}{700\text{Hz}}\right) \tag{4.26}$$

In principle, the mel-spectrum can be obtained from the DFT-spectrum ($\mathcal{F}(s(k), k = 0, \ldots, N-1)$. But as the frequencies are decoded by the index k and dependent from the window length N, Eq. 4.26 cannot be used directly. Instead, the Mel-frequencies are computed by using a filter bank with triangular filters, where the unit-pulse response becomes broader with increasing frequency (cf. [Wendemuth 2004]). Afterwards, the Mel-cepstrum is computed by applying the DCT on the Mel scaled logarithmic spectrum (cf. [Davis & Mermelstein 1980]). The resulting coefficients are called MFCCs:

$$c(q) = \sum_{m=1}^{M} mel(f) \cos\left(\frac{\pi q(2m+1)}{2M}\right), \qquad q = 1 \ldots \frac{M}{2}, \tag{4.27}$$

where M is the desired number of cepstral coefficients and $mel(k)$ is the Mel spectrum gained by the filter bank. In speech recognition normally the first twelve to thirteen coefficients are used[13] [Mporas et al. 2007]. This turn out, to be also a sufficient number for emotion recognition [Schuller et al. 2008b; Böck et al. 2010]. The main steps of the MFCC calculation algorithm are given in [Sahidullah & Saha 2012]:

```
Window the speech signal
Perform an FFT of the windowed excerpt
Compute the absolute spectrum
Perform a Mel frequency warping
Quantise frequency band by utilising triangular filter banks
Apply logarithmic function
Compute the DCT to obtain the cepstrum
Extract the amplitudes of the resulting spectrum.
```

Perceptual Linear Predictive Coefficients The basic idea of linear prediction is to model the vocal tract by a LPC model whose parameters are comparable to the enhanced frequency bands produced by the vocal tract. This technique relies on the source-filter model by Fant, as the important features are the resonance frequencies generated by the vocal tract's characteristic. The model of acoustic speech production

[13] It seems that the use of 12 to 13 coefficients is due to historical reasons. It mainly depended on early empirical investigations. When using Dynamic Time Warping (DTW) with cepstral coefficients it quickly became obvious that very high cepstral coefficients are not helpful for recognition but their calculation was very complex and time consuming. Thus, the MFCCs were optimised by special "liftering" methods. Thereby it turns out that this weighting ended up close to zero when reaching the 12th or 13th coefficient [Tohkura 1987].

(Figure 4.8 on page 63) is considered in a very simplified manner. This model neglects the nasal cavity as well as the lip's radiation model. For the speech production of $s(n)$ only the excitation $u(n)$ with an amplification σ and the vocal tract's transfer function, represented by its coefficients a_i, are considered (cf. Eq. 4.28).

$$s(n) = \sigma u(n) + \sum_{i=1}^{P} a_i s(n-i) \tag{4.28}$$

$$V(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{P} a_i z^{-i}} \tag{4.29}$$

where P is the order of the model. The coefficients a_i are then converted into spectral domain, by concatenating them into a vector and applying a DFT with the length N afterwards. As the model order P is much smaller than N, a zero-padding, where the remaining parts are filled with zeros, is necessary. The peaks in the spectrum represent the formants of the vocal tract. The Z-Transformation $V(z)$ of the transfer function shows the characteristics of an autoregressive model, also called all-pole model since the function only has poles (cf. Eq. 4.29). As the order P of the linear predictive model and thus the number of sampling points is much smaller than the sample length N, this is equal to a smoothing, by which the most prominent spectral peaks – the forming resonance frequencies – are accentuated.

The difficulty is to accurately estimate the coefficients a_i. It is based on the consideration that the actual signal $\hat{s}(n)$ can be estimated from a superposition of P weighted previous signal values, with estimated coefficients $\hat{a}_i$. Afterwards the error $e(n)$ between the estimated signal and the original signal can be calculated:

$$\hat{s}(n) = \hat{a}_1 s(n-1) + \hat{a}_2 s(n-2) + \ldots + \hat{a}_p s(n-p) \tag{4.30}$$

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^{P} \alpha_i s(n-i) \tag{4.31}$$

The relation of Eq. 4.30 can be used to coefficients optimal coefficients by using the speech signal $s(n)$ and its history $s(n-i)$ and minimising the mean square error E gained over a given sample length N, so that $\frac{\partial E}{\partial \alpha_i} = 0$.

$$E = e(n)^2 = \sum_{n=0}^{N-1} \left(s(n) - \sum_{i=1}^{P} \alpha_i s(n-i) \right)^2 \tag{4.32}$$

The partial derivation results in a covariance matrix, which can be estimated either by a covariance approach or an autocorrelation approach (cf. [Wendemuth 2004]). The covariance approach calculates the mean square error over a fixed range and suppresses transient effects and decay processes. A common method using the autocorrelation approach is the Levinson Durbin Recursion. This iterative method determines the coefficients a_i without performing a matrix inversion [Rabiner & Juang 1993]. Both methods have some flaws, as the autocorrelation approach is not unbiased due to transient effects and the covariance method does not produce a minimal-phase solution.

To overcome these problems, Burg presented a method that minimises both, forward and backward prediction errors (cf. [Burg 1975]). The gained LPC coefficients can directly be used as features for speech recognition, describing the smoothed spectrum of the speech signal. The following empirical formula to determine the necessary order of P has been proven to be practically (cf. [Wendemuth 2004]):

$$P = \frac{f_a}{1kHz} + 4 \tag{4.33}$$

The PLP coefficients are deduced from this method by taking into account the human auditory perception (cf. [Hermansky 1990]). A frequency and intensity correction is applied to the spectrum by either a Mel-frequency warping, weighted by an equal-loudness curve and afterwards compressed by taking the cubic root [Young et al. 2006] or a similar technique consisting of spectral resampling, equal loudness pre-emphasis and intensity loudness conversion as suggested by Hermansky. The PLP approach leads to a better noise robustness in comparison to the cepstral approach. The main steps of the PLP calculation algorithm by Hermansky are as follows:

```
Window the speech signal
Perform an FFT of the windowed excerpt
(Spectral resampling of human auditory perception)
(Compute equal loudness pre-emphasis)
(Perform intensity loudness conversion)
Compute the inverse DFT
Solve the linear equation system (either with Levinson
   Durbin Recursion or Burg)
```

The method of linear predictive coding is also used for signal compression, since only the predicted coefficients and the predictive error have to be transmitted rather than the whole audio signal (cf. [Atal & Stover 1975]).

Formant Position and Formant Bandwidth Starting from the description of the vocal tract by Fant, formants are the most descriptive characteristic of different tones. Most often the two first formants are sufficient to disambiguate vowels, as they describe the dominant characteristics of speech production. The first formant F_1

determines open and closed vowels, whereas the second formant F_2 determines the front or back vowels. The third and fourth formant mainly characterise the anatomy of the vocal tract and the timbre of the voice (cf. [de Boer 2000]). To indicate different vowels, they are can be plotted in the F_1-F_2 space. Besides the vowel-formant relation there is also an affect-formant relation (cf. [Vlasenko et al. 2014]). An emotional reaction leads to a shift of the formants, which can be used to recognise the emotion [Scherer 2005b; Vlasenko 2011]. As the formant positions are influenced by the gender of a speaker, this analysis has to be performed for both groups individually. The position of F_2 is an especially good indicator to distinguish male and female speakers [Vlasenko 2011]. Another Investigation further suggest that the position of F_1 (decreasing) and F_3 (raising) is influenced by the age of the speaker (cf. [Harrington et al. 2007]).

To estimate the formants' location and bandwidth, the LPC-smoothed spectrum is used according to Eq. 4.29 on page 69. Mostly the Burg algorithm is applied to determine this spectrum. By performing a DFT, a signal in spectral domain is obtained. This signal comprises a smoothed version of the original signal with spectral peaks at the formant positions. To extract these positions, which are the frequencies at the spectral peaks, these peaks have to be identified. The formant positions F and the formant bandwidth BW can be determined by either a "peak-picking" method on the smoothed spectral curve or by solving the complex root pairs $z = r_0 e^{\pm\theta_0}$ of the LPC-filter equation in the case $A(z) = 0$ [Snell & Milinazzo 1993]. This algorithm is, for instance, implemented in PRAAT [Boersma 2001].

$$F = \frac{f_s}{2\pi}\theta_0 \tag{4.34}$$

$$BW = -\frac{f_s}{\pi}\ln r_0 \tag{4.35}$$

where θ_0 is the angle in rad of the complex root, r_0 is the absolute value of z and f_s and is the sampling frequency in Hz, BW is defined as the frequency range around the formant with a $-3\,\mathrm{dB}$ decrease of the formant's power [Snell & Milinazzo 1993].

Short-Time Energy The energy feature is used to represent the loudness (i.e. energy) of a sound. For speech analysis mostly the short-time energy is calculated. Thus, the sound energy is computed for each speech frame individually as the log of the signal energy over all speech samples s_n within a window:

$$E = \log \sum_{n=1}^{N} s_n^2 \tag{4.36}$$

Furthermore, energy measures of several adjoint segments can be normalised in the range of $-E_{min}$ and 1.0. Therefore, from each energy measure the maximum energy value of the corresponding investigated segment is subtracted and afterwards a 1 is added (cf. [Young et al. 2006]).

All presented spectral features can be calculated regardless of the excitation. Although MFCC and PLP-coefficient extraction was designed to work for voiced parts of speech (vowels), the coefficients gathered for unvoiced parts (consonants) have been successfully used for speech recognition tasks (cf. [Schuller et al. 2009a; Hermansky 2011]). Their applicability for emotion recognition has been investigated in [Dumouchel et al. 2009; Zeng et al. 2009; Böck et al. 2010]. Popular tools for extracting these features are HTK [Young et al. 2006] or openSMILE [Eyben et al. 2010].

4.2.2 Longer-Term Supra-Segmental Features

In contrast to spectral features, prosodic features appear when sounds are concatenated, which goes beyond the short-term segmental parts of speech. In linguistics, the prosodic information covers the rhythm, stress, and intonation of speech. This information is also important to model the transition from one tone or phoneme to another. But they also transmit the utterance's type (e.g. question) or the emotion of the user (cf. [Scherer 2005b]). The typical characteristic of prosodic features is their supra-segmentality. They are not bounded by a specific segment but depict identifiable chunks in the speech [McLennan et al. 2003]. It is still an open debate, which is the right chunk level, especially for emotional investigations. The statements vary from phoneme-level over word-level up to utterance-level chunking (cf. [Batliner et al. 2010; Bitouk et al. 2010; Vlasenko & Wendemuth 2013]).

Thus, for automatic extraction of these features two approaches are commonly used. Either longer-term and long-term (statistical) features of extracted spectral features are calculated, called supra-segmental modelling [Schuller et al. 2009a]. On the other hand, prosodic cues on specific chunk levels are extracted and used to describe the supra-segmental evolvement [Devillers et al. 2006].

For emotional speech analysis the prosodic features are highly important, as emotional feelings are transported by different tones and intensities, which already Darwin found out [Darwin 1874]. But one problem for automatic extraction of prosodic features is their mixing with contextual information, since also vowels are produced by different tones [Cutler & Clifton 1985]. So a variation in the tone can be caused from either different contextual information or different emotions of the speaker, or even a mixture of both types. The following books [Cutler et al. 1983; Fox 2000] serve as a foundation of the current section.

Longer-Term and Long-Term (Statistical) Features

Short-term segmental features only contain information on the currently windowed speech signal, but inferring contextual characteristics gives additional information about the evolvement of speech and the tone-composition. As a first consideration, frame-wise differences could act as additional features to transport longer-term contextual characteristics. This approach could increase the recognition ability for both speech [Siniscalchi et al. 2013] and emotion recognition [Kockmann et al. 2011].

Inferring Contextual Information To infer contextual information, a common method is to include delta (Δ) and double delta ($\Delta\Delta$, also called acceleration) regression coefficients. These coefficients represent the difference between the coefficient of the actual frame and the coefficient of the previous or succeeding frame. The regression coefficients can be computed by using the formula presented in [Young et al. 2006]:

$$d_t = \frac{\sum_{l=1}^{L} l(c_{t+l} - c_{t-l})}{2\sum_{l=1}^{L} l^2} \tag{4.37}$$

with d_t being the regression coefficient for the time-frame t of the static coefficient c_t and the shift length L. To obtain the $\Delta\Delta$ coefficients, the formula is applied to the delta coefficients. THus, when using the double coefficients, the delta coefficients have to be calculated as well. This results in two more coefficients per static feature. These techniques can be implied to all short-term features. In HTK, the commonly used value for L is 2 [Young et al. 2006]. To be able to apply Eq. 4.37 at the beginning and end of the speech frame, the first or last coefficient is replicated as often as needed.

In [Torres-Carrasquillo et al. 2002] the authors presented the “Shifted Delta Cepstra (SDC) coefficients”. These coefficients utilise a much broader contextual information that lead to an improved language identification performance. The authors of [Kockmann et al. 2011] adopted this method for emotion recognition. The basic idea is, to stack delta coefficients, which are computed across a longer range of speech frames. According to [Torres-Carrasquillo et al. 2002], three parameters are defined, L, P and i; L represents the window shift for the regression coefficient's calculation, i denotes the number of the blocks whose coefficients are concatenated, and P is the time shift between the consecutive blocks. The SDC coefficients are calculated according to:

$$\mathrm{sdc}_t = c_{(t+iP+L)} - c_{(t+iP-L)} \tag{4.38}$$

The authors of [Kockmann et al. 2011] suggest an index i in the range of $[-3, 3]$, a shift factor of $P = 3$, and a shift length of $L = 1$. This adds seven coefficients per static coefficient and results in a temporal incorporation of ±10 frames (cf. Figure 4.9).

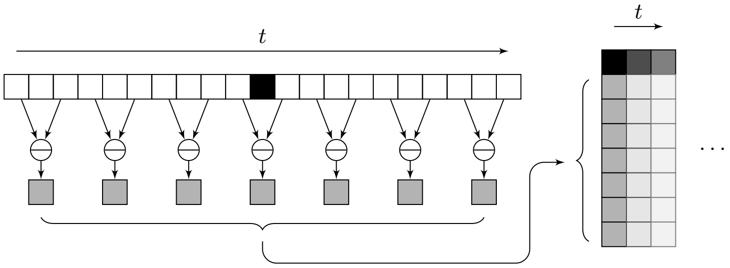


Figure 4.9: Computation Scheme of SDC features. By using the defined values $i = [-3, 3]$, $P = 3$, and $L = 1$, a 7-dimensional SDC vector (▨) is gathered from a temporal context of ten consecutive frames around our actual frame (■).

Inferring functional descriptions To incorporate the general supra-segmental characteristic of speech, specific functionals are utilised on frame-wise extracted features (cf. [Patel 2009; Schuller et al. 2009a]). These functionals describe the shape of the speech signal mathematically. Table 4.2 lists some generally used functionals.

Table 4.2: Commonly used functionals for longer-term contextual information (cf. [after Schuller et al. 2009a, p. 556]).

Functionals	Number
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean	2
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Centroid	1
Zero-crossing rate	1
Linear regression coefficients and corresp. approximation error	4
Quadratic regression coefficients and corresp. approximation error	5

A common tool for extracting these features is the openSMILE toolkit [Eyben et al. 2010]. Hereby, the high order features represent either statistical characteristics of the frame-level coefficients, describing the width or range of the distribution or specific regression coefficients (cf. [Albornoz et al. 2011]).

Prosodic Cues as Features

As stated before, specific prosodic values are distinguishable for different emotions. Especially a change of the prosodic characteristics indicates a change of the user's emotion. Although these characteristics are extracted on short-term speech segments, only the analysis of the long-term evolvement indicates a changed emotion. Common prosodic features are intensity, fundamental frequency and pitch, jitter and shimmer, and speech rate, which will be introduced in the following.

Intensity Apart from the slight increase in loudness to indicate stress, it generally indicates emotions such as fear or anger. The intensity I is a measure for the amount of transported energy E past a given area A per unit of time t (cf. [Fahy 2002]).

$$I = \frac{E}{t \cdot A} \tag{4.39}$$

Since $\frac{E}{t} = P_{ac}$ (sound power), the intensity is commonly denoted as:

$$I = \frac{P_{ac}}{A} \tag{4.40}$$

The unit of I is given in W/m^2. In general, the greater the amplitude of vibrations, the greater the rate of the transported energy. Furthermore, a more intense sound wave is observed.

Since the human auditory perception is very sensitive with a large dynamic range, the distinguishable intensity varies from $1 \times 10^{-12}\,\frac{\mathrm{W}}{\mathrm{m}^2}$ to $1 \times 10^{4}\,\frac{\mathrm{W}}{\mathrm{m}^2}$, normally the decibel dB scale is preferred [Fahy 2002]. Therefore, the intensity is measured in relation to the reference level I_0. Commonly the "threshold of hearing" ($1 \times 10^{-12}\,\mathrm{W/m^2} = 0\,\mathrm{dB}$) is used as I_0. This metric is then called sound intensity level L_I:

$$L_I = 10 \log_{10} \left(\frac{I}{I_0} \right) \tag{4.41}$$

Assuming a spherical propagation of sound pressure around the sound source, the intensity is reduced quadratically with the distance r. Even small changes in the distance between sound source and drain cause quite high changes of intensity. From this it follows that intensity measures the acoustic pressure in dependency of the distance from the sound source. As the distance of the speaker and the recording device cannot be controlled within naturalistic environments, this metric is not suitable as a meaningful feature.

Fundamental Frequency and Pitch In contrast to the presented methods in Section 4.2.1, where the resonance frequencies are in the focus, the pitch estimation tries to determine the excitation frequency. A common method related to pitch detection is the estimation of the fundamental frequency F_0. In this thesis, I will distinguish the fundamental frequency as the excitation frequency within a short-term segment and the pitch as the course of F_0 in a supra-segmental context.

It is known that different pitch levels indicate different meaning, for instance the way in which speakers raise the pitch at the end of a question. However, pitch patterns of rise and fall can indicate such feelings as astonishment, boredom, or puzzlement [Scherer 2001; Patel 2009]. Besides the emotional influence, F_0 also differs for male and female speakers as well as for children (cf. Table 4.3). Also, ageing changes the fundamental frequency. The average F_0 of female voices remains stable over a long period of time and declining only in the alternate years about 10 Hz to 15 Hz [Linville 2001]. In addition, [Hollien & Shipp 1972] noticed for male speakers a continuous decrease of F_0 until the age of 40 years to 50 years, which is accompanied with a drastic increase in further ageing up to 35 Hz with a maximum at 85 years.

Table 4.3: Averaged fundamental frequency for male and female speakers at different age ranges [after Linville 2001].

Age	Male	Female
< 10	260	
20	120	200
40	110	200
60	115	190
80	140	180

To extract F_0, the voiced speech segments are located and the F_0-period is measured within these segments. According to [Rabiner et al. 1976] three categories of F_0-detection algorithms can be distinguished utilising different signal properties: 1) time domain, 2) spectral domain, or 3) time and spectral domain.

Time domain related methods rely on the assumption that a quasi-periodic signal can be suitably processed to minimise the effect of the formant structure. These methods often implement an event rate detection by counting either signal peaks, signal valleys, or zero-crossings [Gerhard 2003]. The number of these events within a specific time range is counted to calculate F_0. A further widely used method is the correlation of shifted waveforms either as autocorrelation $s_{XX}(\kappa)$ (cf. Eq. 4.42) or cross-correlation $s_{XY}(\kappa)$ (cf. Eq. 4.43) [Gerhard 2003]. This method is implemented in PRAAT [Boersma 2001].

$$s_{XX}(\kappa) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{\kappa=-N}^{+N} x[n]x[n+\kappa] \tag{4.42}$$

$$s_{XY}(\kappa) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{\kappa=-N}^{+N} x[n]y[n+\kappa] \tag{4.43}$$

The first peak of $s_{XX}(\kappa)$ or $s_{XY}(\kappa)$ corresponds to the period of the waveform. To be able to calculate $s_{XX}(\kappa)$ the signal must fulfil ergodicity, which can be assumed for short-term speech signals (cf. [Wendemuth 2004]).

In spectral domain analysis, it is taken advantage of the property that a periodic signal in time-domain will have a series of impulses at the fundamental frequency and its harmonics in spectral-domain. These methods normally use a non-linear transformed space and locate the spectral peaks. The cepstral pitch detector computes the cepstrum of a windowed signal, locates the highest peak, which is the signal period, and uses the zero-crossing rate to make a voiced- / unvoiced decision [Noll 1967]. An block diagram is given in Figure 4.10.

The third class of methods utilises a hybrid approach which for instance will spectrally flatten the waveform and afterwards uses time-domain measurements to extract F_0. Thus, the Simplified Inverse Filtering Technique (SIFT) filter uses a 4th order LPC filter to smooth the signal, for instance. Afterwards the fundamental frequency is obtained by interpolating the autocorrelation function in the neighbourhood of the autocorrelation function's peak (cf. [Markel 1972]).

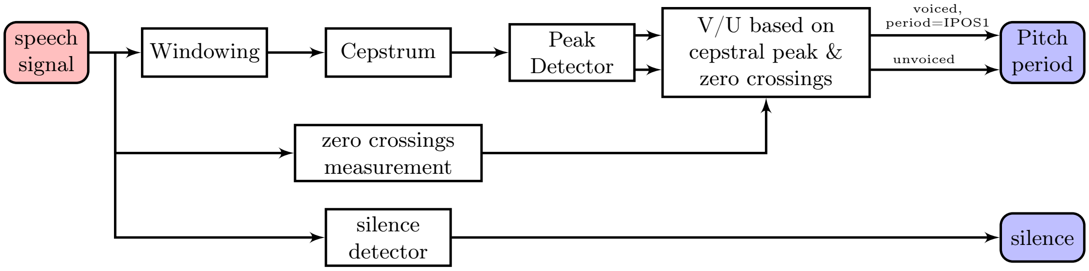


Figure 4.10: Block diagram of a cepstral pitch detector [after Rabiner et al. 1976].

According to [Rabiner et al. 1976], an accurate and reliable F_0-detection is often quite difficult, as the glottal excitation waveform is not a perfect pulse sequence. Although, the interference of the vocal tract and the glottal excitation complicates the measure. Furthermore, the determination of the exact beginning and end of each pitch during voiced segments complicates the reliable measurement.

Jitter and Shimmer Jitter and Shimmer measure slight variations of either the fundamental frequency or the amplitude. These measures are a kind of voice quality features but are also encounted to micro-prosody analysis, as they comprise microscopic changes of the speech signal. The analysis of these effects is used for speaker recognition as well as an early diagnosis of larynxal diseases [Teixeira et al. 2013].

The absolute shimmer $\mathit{Shimmer}_{abs}$ (cf. Eq. 4.44) is defined as the difference in decibel of the peak-to-peak amplitudes:

$$\mathit{Shimmer}_{abs} = \frac{1}{N-1} \sum_{i=1}^{N-1} |20 \log(A_{i+1}/A_i)| \tag{4.44}$$

where A_i is the extracted peak-to-peak signal amplitude and N is the number of extracted F_0 periods. The average value for a healthy person is between 0.05 dB and 0.22 dB [Haji et al. 1986]. The *relative shimmer* measures the absolute difference between the amplitudes, normalised by the average amplitude. The three-point, five-point, and 11-point amplitude perturbation quotient calculates the shimmer within a neighbourhood of three, five, or eleven peaks (cf. [Farrús et al. 2007]).

The absolute jitter Jitter_{abs} (cf. Eq. 4.45) measures fluctuations of F_0. It is calculated by averaging the absolute difference between consecutive periods, T_i are the extracted F_0 period lengths and N is the number of extracted periods [Michaelis et al. 1998].

$$\mathit{Jitter}_{abs} = \frac{1}{N-1} \sum_{i=1}^{N-1} |(T_i - T_{i+1})| \tag{4.45}$$

The *average jitter* measures the absolute difference between two consecutive periods normalised with the average period time. As for shimmer, two further measures include a broader temporal context. The Relative Average Perturbation takes into account the neighbourhood of three periods, while the five-point Period Perturbation calculates the jitter within a five period neighbourhood (cf. [Farrús et al. 2007]).

Praat is commonly used to extract jitter and shimmer measures. Although these two micro-prosodic measures are related to voice quality and therefore can indicate a vocal disease, Scherer listed them as emotional bodily response patterns (cf. Section 2.2 and [Scherer 2001]). These features are also related to stress-indications [Li et al. 2007]. Several studies identify a correlation of ageing and shimmer, which is increasing for elderly speakers (cf. [Ptacek et al. 1966; Ramig & Ringel 1983]), whereas for jitter no such correlation could be identified (cf. [Brückl & Sendlmeier 2005]).

Speech Rate A prosodic measure taking into account the whole utterance is the speech rate, sometimes also called speaking rate. Although a quite obvious measure and, at least, on a subjective level (slow, normal fast) easily assessable by humans, it is only rarely taken into account for automatic emotion recognition. Murray & Arnott described qualitative results for emotional voices, which documented the usability as a feature also for emotion recognition (cf. [Murray & Arnott 1993]).

Several measures exist for the estimation of the speech rate. Most of them calculate the speech rate from samples of connected speech per time unit, as in Words Per Minute (WPM) or Syllables Per Minute (SPM) and Syllables Per Second (SPS). But WPM is not language independent, as words itself can be quite short or quite long. SPM has the problem of ambiguities in syllable estimation [Cotton 1936]. Thus, a measure called Global Speech Rate (GSR) defines the speech rate as a ratio of the overall duration of a target sentence and the overall duration of a reference sentence with equal phonetic content, according to [Mozziconacci & Hermes 2000]. But the phonetic content has to be known.Thus, a robust and reliable automatic estimation of the speech rate is still a challenging task.

The Phonemes Per Second (PPS) metric estimates the number of uttered phonemes based on broad phonetic class recognisers (cf. [Yuan & Liberman 2010]). These recognisers combine acoustically similar phonemes into 6-8 classes, which provides robustness with respect to different languages and speech genres [Yuan & Liberman 2010]. As for speech rate measurements, only the number of uttered phonemes is of interest, these broad phonetic class recognisers are perfectly suited for this task.

Table 4.4: Comparison of different speech rate investigations for various emotions, A = [Mozziconacci & Hermes 2000], B = [Philippou-Hübner et al. 2012], C = [Braun & Oba 2007], and D = [Murray & Arnott 1993]. SR denotes Syllable Rate and AR denotes Articulation Rate. In cases where a reference has been used, this reference is highlighted.

Investigation	A (SR)	B (AR)	C (SR)	C (AR)	D
Measure	GSR	PPS	SPS	SPS	qualitative
fear	1.11	16.5	4.9	5.5	much faster
neutral	*1.00*	15.3	5.3	5.5	*reference*
joy	1.01	14.5	4.7	6.2	slower or faster
anger	0.94	14.0	5.1	5.5	slightly faster
boredom	0.82	13.5	n.s.	n.s	n.s.
disgust	n.s.	10.5	n.s.	n.s	very much slower
sadness	0.92	9.9	4.5	6.0	slightly slower
indignation	0.85	n.s.	n.s.	n.s	n.s.

Although speech rate investigations utilise different time-units and duration ratios, they come to the same results for emotional speech rates (cf. Table 4.4). Furthermore, when investigating the speech rate characteristics, the pauses within an utterance have a large influence on the calculated speech rate. Therefore, a second rate has to be distinguished, the Articulation Rate (AR). While the speech rate is calculated from the whole utterance including pauses, the AR calculation omittes the pauses. Especially, to emphasise specific utterance parts, or to indicate expressions of high cognitive load, or to represent emotional statements, pauses are an important part of spontaneous speech [Rochester 1973]. The difference of Syllable Rate (SR) and AR is investigated for instance in [Braun & Oba 2007]. Their results are given in Table 4.4.

4.3 Classifiers

For emotion recognition from speech, as well as for general pattern recognition problems, several classification methods are established (cf. Chapter 3). In particular, already utilised classifiers for speech recognition are also used for emotion and affect recognition. As already stated in Section 1.2, the community mostly relies on supervised approaches. For these approaches, classifiers are trained from examples, consisting of an input feature vector and its true output value. The utilised learning algorithm analyses the training data and produces an inferred function, which is afterwards used for mapping new (unknown) examples.

Depending on which property is highlighted, there are different orderings of classification approaches. I would like differentiate the type of class assignment. Therefore, I introduce GMMs and HMMs (cf. Section 4.3.1) as production models. As mentioned before (cf. Section 3.2) other approaches are utilised, mainly Multi Layer Perceptrons (MLPs), SVMs or Simple Recurrent Neural Networks (SRNs). In my research I focus on GMMs and HMMs. Therefore, a detailed introduction of other classification approaches is neglected, but I refer the reader to [Benesty et al. 2008; Glüge 2013].

4.3.1 Hidden Markov Models

The human speech production can generate different variants of the same acoustics. These variants can be either stretched or shrinked. This results in an acoustic observation which basically consists of the same characteristics, but varies in the temporal occurrence of the observation. However, in the case of a stretched acoustic, these observation can be seen as consisting of repeating sub-parts whereas for the shrinked case some sub-parts are very small or even non-existent. This kind of observation causes troubles in the recognition since the same meaning can be produced with different (temporal) variants. To overcome this problem, HMMs are utilised. An HMM is constituted by a twofold production process: 1) a temporal evolution, to decode the temporal stretching and shrinking 2) and an output production, to decode the observed acoustic sub-part. This architecture enables the HMM to uncouple the temporal resolution of the speech signal from the observed features. Thus, the HMM produces at first the most possible sequence of states, whereupon a repetition of one or mores states is possible. Afterwards, for each selected state the most likely output is produced. The basic unit of a sound represented by an HMM is either a word, a phoneme, or a short utterance [Young 2008].

HMMs have long been used successfully in speech recognition as well as emotion recognition, thus I only depict the most important parts of this modelling technique. For further details, I refer to the corresponding literature: (cf. mathematical description of HMMs [Eppinger & Herter 1993; Wendemuth 2004; Young 2008], HMMs and emotions [Schuller & Batliner 2013; Vlasenko et al. 2014], parameter optimization [Böck et al. 2010], fusion architecture [Glodek et al. 2012]).

An HMM is a finite state machine $hmm = \{S, K, \pi, a_{ij}, b_{jk}\}$ where $S = \{s_1, \dots, s_n\}$ denotes the set of states, $V = \{v_1, \dots, v_n\}$ denotes the output alphabet, π_s denotes the initial probability of a state s, $\{a_{ij}\} = P(q_t = s_j | q_{t-1} = s_i)$ are the transition probabilities, and $\{b_{jk}\} = P(O_t = v_k | q_t = s_j)$ are the production probabilities. Figure 4.11 shows the graphical representation of a commonly used left-to-right HMM.

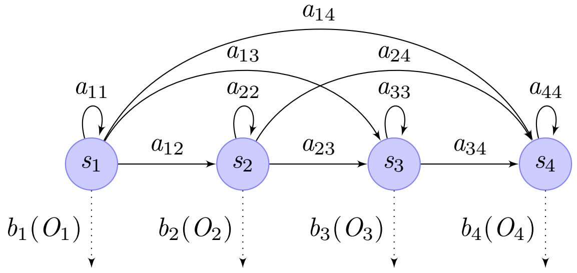


Figure 4.11: Workflow of a four states HMM, a_{ij} is the transition probability from state s_i to state s_j and $b_j(O_v)$ is the probability to emit the symbol O_v in state s_j.

The HMM produces for every time step $t = 1, \dots, T$ one observable output $O_t \in K$ and passes through an unobservable sequences of states. In speech and emotion recognition it is common to use mixtures of Gaussians as output observation probabilities:

$$b_{jk} = \sum_{m=1}^{M} c_{jm}^{k} \mathcal{N}(o, \mu_{jm}^{k}, \sigma_{jm}^{k}) \tag{4.46}$$

where $\mathcal{N}$ denotes a normal distribution with the parameters mean (μ_{jm}) and covariance (σ_{jm}). The parameter M denotes the number of Gaussians and is determined by the length of the feature vector, i.e. the number of used features. Diagonal variance matrices are used to reduce the effort of variance estimation (cf. [Wendemuth 2004; Young et al. 2006]). This restriction has the consequence that the estimated distributions are oriented according to the pre-defined coordinate axes. This constraint can be circumvented by using mixture distributions (cf. [Wendemuth 2004]).

Production modelling tries to find the sequence of words or emotional patterns $W = \{w_1, \dots, w_k\}$ that most likely have generated the observed output sequence $\mathbf{O}$:

$$\hat{W} = \arg\max_{W}[P(W|\mathbf{O})] \tag{4.47}$$

As $P(W|\mathbf{O})$ is difficult to model directly, the Bayes' rule is used to transform Eq. 4.47 into the equivalent problem:

$$\hat{W} = \arg\max_{W}[P(\mathbf{O}|W)P(W)] \tag{4.48}$$

The likelihood $P(\mathbf{O}|W)$ is determined by acoustic modelling, namely the HMM. The prior probability $P(W)$ is defined by a language modelling. These terms show the strong connection to speech recognition. The language model indicates how likely it is that a particular word was spoken or a certain affect occurred given the current context. Therefore, empirically obtained scaling factors are used, for instance n-gram modelling [Brown et al. 1992]. For emotion recognition, the use of a language model is shortly discussed in [Schuller et al. 2011c]. It is stated that due to the data sparseness mostly uni-grams have been applied and they serve as linguistic features (salient words), to define the amount of information a specific word contains about an emotion category [Steidl 2009]. There are no findings yet, on proper "language models" for emotions.

In acoustic modelling, two issues have to be solved: 1) calculate the probability for each model λ generating the observation sequence $\mathbf{O}$, and 2) find the best state sequence matching the given observation. To solve the first issue, the produced observations over all possible state sequences are summarised and multiplied with the likelihood that these state sequences are generated by this model (cf. Eq. 4.49). Therefore, all possible state sequences $1 \dots N$ and all possible output sequences $1 \dots T$ have to be considered. These calculations can be further simplified (cf. Eq. 4.51) and by making use of the consideration that the likelihood of the actual state is only depending on the previous state[14]. The corresponding algorithm is called forward-backward algorithm [Rabiner & Juang 1993].

$$P(\mathbf{O}|\lambda) = \sum_{q} P(\mathbf{O}|\mathbf{q},\lambda) \cdot P(\mathbf{q}|\lambda) \tag{4.49}$$

$$P(\mathbf{O}|\lambda) = \sum_{q} \pi_{q_i} \prod_{t=1}^{T} a_{q_{t-i},q_t} b_{q_t O_t} \tag{4.50}$$

To solve the second issue, finding the most likely path of states $\mathbf{q}_{max}$ through the model, the sequence of states with the highest likelihood have to be calculated:

$$\mathbf{q}_{max} = \max_{\mathbf{q}} P(\mathbf{q}|\mathbf{O},\lambda) \tag{4.51}$$

$$P(\mathbf{O},\mathbf{q}^*|\lambda) = \max_{\mathbf{q}\in Q^T} P(\mathbf{O},\mathbf{q}|\lambda) \tag{4.52}$$

Eq. 4.52 is evaluating efficiently with the Viterbi algorithm [Viterbi 1967] by taking advantage of the Markov property. The Viterbi algorithm iteratively calculates the maximum attainable probabilities for a sub-part of the observation under the additional condition to end in a certain gradually increasing state s_j and at the same time storing the requested sequence by a backtracking matrix (cf. [Wendemuth 2004]).

But before the the most likely path can be calculated, the HMM's parameters $\{a_{ij}\}$ and $\{b_{jk}\}$ have to be estimated. To calculate these parameters, a training corpus with acoustic examples and pre-defined labels have to be utilised. For an efficient estimation the Baum-Welch (BW) algorithm is used (cf. [Wendemuth 2004]). This algorithm uses the forward-backward algorithm and is an instance of the Expectation-Maximization (EM) algorithm [Dempster et al. 1977]. The iterative EM algorithm consists of an E-step to compute state occupation probabilities and an M-step to obtain updated parameter estimates utilising maximum-likelihood calculations (cf. [Young 2008]).

As a special case, GMMs are distinguished from HMMs by having only one emitting state[15]. GMMs are used to capture the observed features within one state without inferring transitions. It is assumed that these models will better capture the emotional content of a whole utterance without comprising the spoken content, which is varying within the utterance. The same methods as for HMMs are used for training and testing. The only difference is that due to the self-loop all observations in a GMM are mapped to the same state. When considering an HMM the number of Gaussian mixture

[14] This is called first order Markov property. The actual state only depends on the previous state and not a sequence of states that preceded it.

[15] In the literature (cf. [Vlasenko et al. 2007a]) these classifiers are also denoted as HMM/GMM, as for training and testing the GMM is seen as an one-state HMM with a self-loop. Thus, different lengths of utterances results in different numbers of self-loops.

components is normally between 10 and 20 [Young 2008], for GMMs commonly many more mixture components are used, 70-140 for emotion recognition [Vlasenko et al. 2014] and up to 2 048 for speaker verification [Reynolds et al. 2000]. To increase the number of mixture components, a technique called mixture splitting is mostly applied (cf. [Young et al. 2006]). Hereby, the mixture component with the highest corrected mixture weight[16] ("heaviest" mixture) is copied, the weights are divided by 2, and the mean is perturbed by ±0.2 of the corresponding standard deviations (cf. [Young et al. 2006]). Afterwards, all parameters are re-estimated by applying the BW algorithm.

4.3.2 Defining Optimal Parameters

When using classifiers an initial problem is an optimal selection of the model parameters. For a GMM classifier these are the number of mixtures and the number of iteration steps. For HMMs an additional parameter, namely the number of hidden states, has to be defined. Furthermore, the choice of utilised feature sets also has an effect on the classification performance. Afterwards, the classifier can be trained to determine the values of the parameters, accordingly.

Optimal Parameters for HMMs

The number of hidden states for emotion recognition was investigated by [Böck et al. 2010], for instance. In a comparative experiment with three different databases the number of states was changed step-wise from one state to four states. As an optimal number, three states were identified. In the case of very short utterances consisting only of a few phonemes, even one state, leading to a GMM classifier system, was identified as sufficient [Böck et al. 2010].

The second parameter, the number of iterations, was also investigated in [Böck et al. 2010] for HMMs. This number specifies the iterations for the BW algorithm and was changed between 1 and 30. The authors concluded that three iterations provide the best recognition performance utilising a three-state HMM on simulated material, whereas on naturalistic material five iterations provide the best performance utilising the same classifier. The use of more iterations results in a decreased performance. Thus, it can be concluded that the models lose their capability to generalise, which is comparable to the over-fitting problem for ANNs (cf. [Böck 2013]).

[16] The corrected mixture weight is calculated by subtracting the number of already performed splits in the actual step from the corresponding mixture component. This method assures that repeated splitting of the same mixture component is discouraged (cf. [Young et al. 2006]).

Also, the influence of different spectral features sets was analysed in [Böck et al. 2010] and [Böck 2013]. The difference of the zeroth cepstral coefficient (C0), which represents the mean of the logarithmic Mel spectrum and thus closely related to the signal energy (cf. [Marti et al. 2008]), and the short-term energy (E) itself were investigated. To this end, two different spectral feature sets, MFCC, PLP, their temporal information (Δ and $\Delta\Delta$), are compared once utilising the C0 and once using E. These investigations are pursued on both simulated and naturalistic material. Böck et al. stated that for simulated material the performance of the feature sets according to the additional term is quite similar. For naturalistic material, the performance utilising short-term energy degrades [Böck 2013]. This is attributed to the fact that in naturalistic material this energy term is influenced by several factors (distance speaker to microphone, different loudness of speakers). In comparison of PLP and MFCC features, the author concluded that MFCCs should be preferred. This is supported by observations of the INTERSPEECH 2009 Emotion Challenge [Schuller et al. 2011c]. The importance of temporal information for HMMs using Δ and $\Delta\Delta$ coefficients are confirmed in, for instance, [Glüge et al. 2011], by comparing the classification results for emotion recognition of SRNs, having temporal information by design.

Another study by Cullen & Harte compared five different feature sets to classify various dimensional affects on a naturalistic affect corpus using HMMs. The utilised feature sets are (1) energy, spectral, and pitch related features, (2) pure spectral features (MFCC), (3) glottal features, (4) Teager Energy Operator (TEO) features, and (5) long term static and dynamic modulation spectrum (SDMS) features . The authors compared the performance of these feature sets for different emotional dimensions, as `activation`, `valence`, `power`, `expectation`, and `overall emotional intensity`. Cullen & Harte concluded that for different emotional dimensions, different feature sets gain an optimal performance. Feature set (1) gains the best performance on `activation` and also captures `power` and `valence`. These findings are also approved by [Schuller et al. 2009a]. Feature set (2) provides the best results for `power` and `valence`. Using glottal features, the classifier performance decreases for all dimensions. An HMM trained with TEO features gains high performance for `expectation` and `valence`. The long-term SDMS features perform well on `expectation` and it is assumed that this affect may vary quite slowly (cf. [Cullen & Harte 2012]).

Optimal Parameters for GMMs

In contrast to HMMs, only two parameters have to be investigated for GMMs. Applying GMMs for emotion recognition gives better classification results than HMMs, as shown in [Vlasenko et al. 2014]. The optimal number of mixtures and iterations depends largely on the type of material. Especially in [Vlasenko et al. 2007b] and [Vlasenko et al. 2014] the number of mixtures needed for GMMs utilising simulated material (emoDB with `low` and `high arousal` emotional clustering, cf. Section 5.1.1) and naturalistic material (VAM, cf. Section 5.2.2) was investigated. To this end, the authors varied the number of mixtures in the range of 2 to 120 and concluded that the optimal number of mixtures to gain stable and robust results is 117 for the simulated (emoDB) and in the range of 77 to 90 for the used naturalistic affect database (VAM) when applying their phonetic pattern independent classifiers. As features they used the first 12 MFCCs and the zeroth cepstral coefficient (C0) with Δ and $\Delta\Delta$ coefficients. The authors used five iteration steps for their experiments. The authors of [Vlasenko et al. 2014] and [Vlasenko et al. 2007b] could furthermore show that the results gained with GMMs are more stable and robust in comparison to HMMs with two to five states. The gained UAR on HMMs was roughly 10 % lower than the UAR gained with GMMs.

My own experiments on the influence of different features, the effect on over-fitting when incorporating investigations about the number of iterations and the optimal number of mixtures can be found in Section 6.2.1.

4.3.3 Incorporating Speaker Characteristics

As emotional expressions are very individual, it would be the best to utilise individualised classifiers or adopt the classifers onto the emotional reaction of a specific user. But these methods are not always feasible since the material for each emotional reaction of a user has to be present. However, the problem of speaker variability has been already addressed for ASR systems (cf. [Burkhardt et al. 2010; Bahari & Hamme 2012]).

In ASR, the problem of inter-speaker variability caused a performance degradation while recognising many different users [Emori & Shinoda 2001]. This is due to different speaker characteristics, where gender is the most significant influence. This gender effect is caused by different sizes of the vocal tract between male and female users. The vocal tract of male users is approx. 18 cm long and generates a lower frequency spectrum, whereas female users' vocal tract is approx. only 13 cm long, resulting in higher frequencies [Lee & Rose 1996]. These differences affect the spectral formant positions by as much as 25 % (cf. [Lee & Rose 1998]). Apart from these anatomical reasons, different speaking habits also have an effect on speech production, as for instance the speaking rate or the intonation (cf. [Ho 2001]). The authors in [Dellwo et al. 2012] also argue that speech is a highly complex brain-operated series of muscle movements allowing to a certain degree an individual operation. This is called an "idiosyncratic motion" and also affects the speech signal [Dellwo et al. 2012].

Therefore, two different approaches, to deal with these inter-user variabilities, have been used successfully in speech recognition: Either the speaker variabilities are normalised or speaker-group dependent models are used. Vocal Tract Length Normalisation (VTLN) normalises the speaker variabilities by estimating a warping factor to correct the different vocal tract lengths of the speakers [Emori & Shinoda 2001], which is either compressed (female users) or expanded (male users). Therefore, a piecewise linear transformation of the frequency axis is pursued (cf. [Zhan & Waibel 1997]):

$$f' = \begin{cases} \beta^{-1} f & \text{if } f < f_0 \\ bf + c & \text{if } f \geq f_0 \end{cases} \tag{4.53}$$

where f' is the normalised frequency, β is the user-specific warping factor, f_0 is a fixed frequency to handle the bandwidth mismatching problem during the transformation, b and c can be calculated with a known f_0 [Wong & Sridharan 2002]. The warped MFCCs are then created for all files with warping factors in a range of 0.88-1.22.

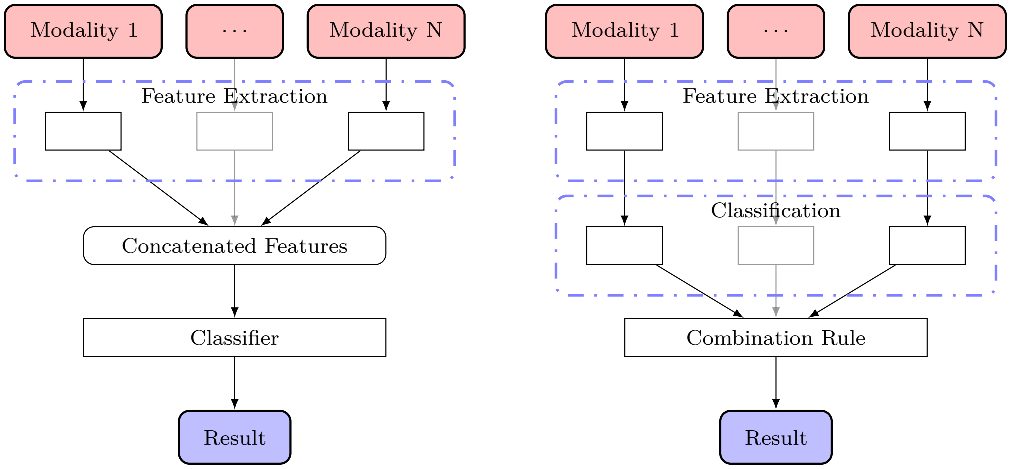
An important condition for VTLN is the estimation of the warping factors. A rough rule for these factors can be deduced from the application of VTLN. As it is used to normalise anatomical differences of the vocal tract for different speakers, the factor should e.g. “reduce” the length of the vocal tract for male speakers and “stretch” the vocal tract length for female speakers. Childrens’ vocal tract should be stretched even more [Giuliani & Gerosa 2003]. Here, an investigation is presented, showing that VTLN also has an age dependency for children. In the age of 7 to 13 the characteristics of speech changes drastically. It can be assumed that these age-dependent changes also apply to adults, but in much longer ranges of years.

For speaker-group dependent modelling, different speaker groups are defined and group-specific models are trained with each utilised speaker group and emotion. For this, the corresponding speaker group has to be identified in advance. Unnormalised features are used with the group-specific models [Vergin et al. 1996]. In order to achieve recognition, the speaker group of the actual speaker has to be known, either a priori or, for instance, by upstreamed gender-recognition. Then the acoustics are recognised applying the selected speaker-group dependent model.

Recognising age and gender automatically is well known in ASR systems. This can be done in the very first beginning of a dialogue, by using just a few words of the subject. Typical architectures to distinguish age and gender uses SVMs, MLPs or HMMs [Bocklet et al. 2008; Burkhardt et al. 2010]. An advanced method is to utilise an UBM together with a GMM to take advantage of the adjustable threshold [Bahari & Hamme 2012; Gajšek et al. 2009]. The authors of [Burkhardt et al. 2010; Li et al. 2010] utilise a decision-level fusion to combine several age and gender detectors. Typically spectral and prosodic features are used, as PLPs and MFCCs, F_0, jitter and shimmer [Meinedo & Trancoso 2011]. They can be enriched by their first order regression coefficients to incorporate contextual information. The application of functionals can be used to generate long-term statistical information (cf. [Bocklet et al. 2008; Li et al. 2010]). Automatic approaches clustering the user regarding age and gender reach accuracies of approx. 96 % (cf. [Lee & Kwak 2012; Mengistu 2009]).

4.3.4 Common Fusion Techniques

Multimodal approaches are an arising topic for emotion recognition, especially in the case of naturalistic interactions. This approach copies the human way of understanding emotions by inferring information from several modalities simultaneously. In general, two types of fusion approaches are distinguished (cf. [Wagner et al. 2011]): feature level fusion (cf. Figure 4.12(a)) and decision level fusion (cf. Figure 4.12(b)).



(a) Sketch of a feature level fusion architecture. Features of each modality are concatenated. The final decision is generated by a classifier on the concatenated features.

(b) Sketch of a decision level fusion architecture. The features of each modality are classified separately. The final decision is generated by any kind of combination rule.

Figure 4.12: Overview of feature and decision level fusion architectures.

In the first case, the different modalities are concatenated directly on feature level into a single high-dimensional feature set (cf. [Busso et al. 2004]). For this, it is assumed that this resulting feature set contains a larger amount of information than single modalities and thus, achieves a higher classification performance. One constraint that has to be considered here is that the features of all involved modalities are extracted on the same time scales. Thus, it has to be secured that the emotional characteristics present, for instance, in acoustics are matching the expressed facial expressions. This means in other words that the involved multimodal response patterns are present at the time of the investigation.

In decision level fusion, the contrary approach is used. Specific feature sets on single classifiers for each modality are applied. The final decision is gained afterwards, by combining the single results using rules like for instance, Bayes' Rule or Dempster's Rule of Combination (cf. [Paleari et al. 2010]). The decision level fusion has many benefits over the use of a feature level fusion. Different time scales of single modalities can be adjusted in the individual classifiers. Besides the obvious training efficiency attainable by using several small feature vectors instead of one high-dimensional one, the resistance against fragmentary data of real-time data is rising. Especially, when different classifiers for different modalities are used, the malfunction of one sensor device will only result in a malfunction of the corresponding classifier and just marginally influence the final decision [Wagner et al. 2011]. Additionally, also combinations of both approaches, called "hybrid fusion" are investigated (cf. [Kim 2007; Hussain et al. 2011]). In this case, both feature level fusion and decision level fusion are pursued and the final decision is achieved by combining all single decisions using a third fusion level.

Most works in emotion recognition use a bi-modal approach and focus on audiovisual information [Busso et al. 2004; Zeng et al. 2009]. There, most fusion approaches utilise either feature level fusion or decision level fusion. Surprisingly only rarely other modalities such as body gestures [Balomenos et al. 2005] or physiological information [Kim 2007; Walter et al. 2011] are utilised. These studies mostly rely on decision level fusion, as the time scales of the modalities are quite different and thus difficult to combine on feature level. Just a few studies try to integrate more than two modalities (cf. [Wagner et al. 2011]).

The Markov Fusion Network

To perform the fusion of several modalities under the constraint of fragmentary data, a late fusion approach utilised by colleagues at the Ulm University (cf. [Glodek et al. 2012]) should be shortly introduced. The Markov Fusion Network (MFN) (cf. Figure 4.13) reconstructs a non-fragmented stream of decisions $\mathbf{y}$ based on an arbitrary number of fragmented streams of given decisions $\mathbf{x}_{tm}$ where $m = 1, \dots, M$ and $t = 1, \dots, T$. In this case, M is the number of different modalities and T is the timepoint a decision is available. In an MFN, the relationship of the reconstructed decisions over time is represented by a Markov chain, whereas the decisions of the modalities (input decisions) are connected to the Markov chain of final decisions whenever they are available (cf. Glodek et al. 2012). The model is originated from the application of Markov random fields in image processing.

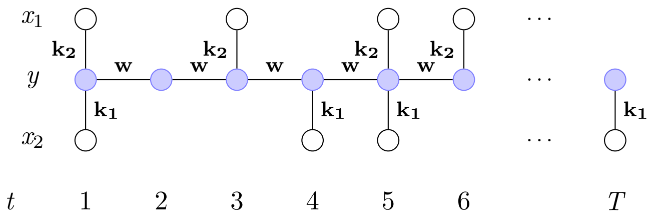


Figure 4.13: Graphical representation of an MFN. The estimates y_t are influenced by the available decisions x_{tm} of the source m at time t and the adjacent estimates y_{t-1}, y_{t+1}.

Once the input decisions and parameters are determined, the most likely stream of final decisions needs to be estimated. The most important parameters of a MFN are $\mathbf{k}$ and $\mathbf{w}$. The parameter vector $\mathbf{k}$ defines the strength of the influence of each single modality. Thus, in the presented approach, we distinguish between $\mathbf{k_v}$ for the visual modality, $\mathbf{k_a}$ defining the acoustics' modality influence, and $\mathbf{k_g}$ adjusting the gesture influence. The parameter vector $\mathbf{w}$ weights the cost of a difference between two adjacent nodes of the MFN. Due to the limited number of dependencies, it is sufficient to perform a gradient descent optimization. More details about the training alorithm can be found in [Glodek et al. 2012].

4.4 Evaluation

The main goal of evaluation is to assess the performance of the investigated method, for instance in affect recognition or prediction. This infers to choose between several feature sets, classifiers, and training algorithms. For this, at first the data samples have to be prepared in such a way that data bias and overfitting can be avoided. Second, the classification performance or the prediction error has to be estimated and the classifier minimising this criterion has to be selected. A good survey on model selection procedures is given in [Arlot & Celisse 2010].

A validation is utilised to be able to estimate the classifier performance. For such a set the assignment of classes to the data samples is a priori known. This allows the indication of the performance of a chosen model or classifier. Therein, common statistical quality criteria are used (cf. [Olson & Delen 2008; Powers 2011]). In speech recognition as well as in emotion recognition several methods exist how training and test sets are arranged and how the performance of a classifier utilising different emotional classes and speakers is calculated. The most common types are shortly descibed in the following.

4.4.1 Validation Methods

The most common validation method splits the dataset randomly into j mutually exclusive subsets or folds. The $j-1$ partitions provide material for the training and the remaining partition is used for testing. This method is called j-fold cross-validation (cf. [Kohavi 1995]). The j indicates the number of partitions. This procedure is repeated j times, to compensate a possible influence of the selection. The individual results are then averaged over the number of runs. This method assumes that the data is identically distributed and training and test data samples are independent. These two assumptions are usually valid and, thus cross validation can be applied to almost any learning-algorithm within almost any framework. The question of choosing j is discussed extensively in [Arlot & Celisse 2010]. In speech recognition as well as in affect recognition from speech commonly a 10-fold cross-validation is used [Schaffer 1993; Kohavi 1995]. Audio data from several speakers is usually used when applying cross-validation. This method utilises samples of all speakers for training and testing. Thus the learning algorithm has been faced with samples of every speaker. Hence, the classifier should learn a general characteristic that is valid for all speakers, although training and validation set are disjoint subsets of the data material.

An extension of this approach is the stratified j-fold cross-validation (cf. [Diamantidis et al. 2000]). In this case, each part contains the same portions of all classes or labels as the complete dataset. This approach is applied in cases, where the assumption of equally distributed data samples cannot be ensured.

Another extension, intended to represent a realistic speech and emotion recognition scenario, separates the subsets according to the number of speakers within the dataset. This method takes into account the subject-to-subject variation. The users of speech recognition systems are mostly not known before and thus, their characteristics can differ from the characteristics of the speakers used for training. To test the overall generalization ability of the implemented methods, the dataset is split into the n different speakers. One speaker is reserved for testing and the remaining $n-1$ ones are used for training, which is then repeated n times. Afterwards the average is used to describe the performance of the classification. Thus, the speech characteristics of the particular test speaker is never seen by the classifier. This approach is derived from the Leave-One-Out (LOO) cross-validation procedure (cf. [Picard & Cook 1984]), where j corresponds to the number of samples in the dataset and denoted as LOSO[17]

[17] In literature, especially in fMRI studies, also the term Leave-One-Subject-Out is common.

to clearly indicate the connection to the involved speakers (cf. [Schuller et al. 2008a]). This approach is similar to cross-validation, except that the characteristics of the utilised test pattern are not included into training. In terms of conformity with reality, LOSO is a more accurate approach, but also a more exhaustive splitting method, as the number of speakers n is normally greater than 10, which is the common number of folds for j-fold cross-validation. But for LOSO, the researcher have to take into account that different each speaker could have different amounts of data and especially a different distribution of data for each class. As for the j-fold cross-validation the LOSO validation has a stratified version, to adjust the different amount of data for each speaker. In this case, the distribution of different classes in the training data is artificially aligned, by presenting under-represented data more often.

To avoid the high number of test runs for LOSO, sometimes a so-called Leave-One-Speaker-Group-Out (LOSGO) is utilised. [Schuller et al. 2009a]. In there, a fixed number of speakers are omitted during training but used for testing. This method uses the advantages of the LOSO-training method that the characteristics of the test-speakers have not been seen during training, but it avoids the required high number of training folds. The size of the speaker group is commonly chosen in such a way that the number of cycles does not exceed ten (cf. [Schuller et al. 2009a]). Both methods, LOSO and LOSGO do not guarantee the same class distribution in training and test data. Therefore, the chosen performance measure has to correct this case.

4.4.2 Classifier Performance Measures

When evaluating a classifier, there are different ways to measure its performance. The simplest measure would be the percentage of correctly classified instances. But this measure provides no statements about failed classification which must be taken into account to compare the results of different classifications. Therefore, in information retrieval and pattern recognition several other measures have been established (cf. [Olson & Delen 2008; Powers 2011]). These measures are mostly based on the confusion matrix (cf. Table 4.5). Each column represents the predicted items from classifier output, while each row represents the true items in the classes. This visualisation highlights the classification confusion between classes; see Table 4.5 for a binary classification problem. In this example, the two classes are denoted as `Positive` and `Negative`. Furthermore, the values in the individual cells are commonly denoted as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [Powers 2011]. TP denotes the examples which are correctly predicted as positive, TN is the number of items correctly classified as negative. FP indicates the number of items wrongly predicted as positive, which are originally from the negative class. The same applies to FN, this value specifies the wrongly negative predicted items, whose true class is positive.

Table 4.5: Confusion matrix for a binary problem. Additionally the marginal sums are given. N is the total number of samples in the dataset.

	Predicted Class		
True Class	Positive	Negative	
Positive	TP	FN	TP+FN
Negative	FP	TN	FP+FN
	TP+FP	FN+TN	N

The most commonly used evaluation measure is the accuracy rate (Acc). It measures the percentage of correct predictions:

$$Acc = \frac{TN + TP}{FN + FP + TN + TP} = \frac{TN + TP}{N} \tag{4.54}$$

The error rate (Err) is the complement of Acc. It evaluates the percentage of incorrect predictions. Both measures are measures that can be directly applied to multiclass classification problems:

$$Err = \frac{FN + FP}{FN + FP + TN + TP} = 1 - Acc \tag{4.55}$$

Further measures are used to estimate the effectiveness for each class in the binary problem. The recall Rec (cf. Eq. 4.56) measures the proportion of items belonging to the positive class and being correctly classified as positive. This measure is also known as sensitivity or true positive rate. The specificity Spe (cf. Eq. 4.57) measures the percentage of correctly predicted negative items.

$$Rec = \frac{TP}{TP + FN} \tag{4.56}$$

$$Spe = \frac{TN}{FP + TN} \tag{4.57}$$

The precision Pre, also called positive predictive value (PPV), estimates the probability that a positive prediction is correct, whereas the inverse measure, the negative predictive value (NPV) denotes the probability that a negative prediction is correct.

$$Pre = \frac{TP}{TP + FP} \tag{4.58}$$

$$NPV = \frac{TN}{TN + FN} \tag{4.59}$$

It should be noted that it is not possible to optimise all measures simultaneously. In particular, recall and specificity are negatively correlated with each other [Altman 1991]. Hence, combined measures are used to have a single value judging the quality of a classification and balancing this effects. The F-measure (cf. Eq. 4.60) combines precision and recall using the harmonic mean. Therein, a constant β controls the trade-off between both measures. For most evaluations the F_1-measure is used, which weights precision and recall equally (cf. Eq. 4.61).

$$F_\beta = \frac{(1+\beta^2) \cdot Pre \cdot Rec}{(\beta^2 \cdot Pre) + Rec} \tag{4.60}$$

$$F_1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec} = \frac{2 \cdot TP}{2TP + FP + FN} \tag{4.61}$$

But mostly, the emotion classification is not binary, since several classes are utilised (cf. Section 3.1). Therefore, the confusion matrix spans several classes. In this case, the performance measures introduced become limited to “per class rates”. Thus, an overall classifier evaluation measure is needed. We denote the confusion matrix A where each element A_{ij} indicates the number of items belonging to class i assessed to class j. The dimension of A is $c \times c$, where c is the number of classes. An example of a “per class rate” is given in [Olson & Delen 2008] as an extension of the recall definition. Olson & Delen define a true classification rate as the number of correctly assessed items to a class divided by all samples assigned to the corresponding class:

$$\text{per class rate}_i = \frac{a_{ii}}{\sum_{j=1}^{c} a_{ij}} \tag{4.62}$$

A commonly used overall classifier evaluation measure that can be deviated from Olson & Delen’s definition, is described in [Rosenberg 2012]. It applies an averaging over the number of classes summing each “per class rates”.

$$\text{AvR} = \frac{1}{c} \sum_{i=1}^{c} \frac{a_{ii}}{\sum_{j=1}^{c} a_{ij}} \tag{4.63}$$

Rosenberg called this measure Average Recall (AvR), but most of the researchers in the speech community call this measure Unweighted Average Recall (UAR) (cf. [Schuller et al. 2009a; Schuller et al. 2010a; Schuller et al. 2011c]). When a LOSO validation is performed, the average over all UARs of all folds is reported as UAR.

As a further consideration, the number of samples for each class could be taken into account to correct highly unbalanced class distributions. Thus, the ratio of all samples per class to the overall number of samples N is utilised as a weighting factor. Hence, the Weighted Average Recall (WAR) can be calculated as

$$\mathrm{WAR} = \sum_{i=1}^{c} \frac{\sum_{j=1}^{c} a_{ij}}{N} \frac{a_{ii}}{\sum_{j=1}^{c} a_{ij}} = \sum_{i=1}^{c} \frac{a_{ii}}{N}. \tag{4.64}$$

Given Eq. 4.64 it is obvious that WAR is equivalent to the accuracy Acc. Thus, Schuller et al. argued “the primary measure to optimise will be unweighted average (UA) recall, and secondly the weighted average (WA) recall (i.e. accuracy)” [Schuller et al. 2009c].

4.4.3 Measures for Significant Improvements

Reporting and comparing the classifier’s performance and its improvement alone, is not sufficient to derive a statement that the performance enhancement is caused by the investigated method. The results of the performed experiment itself may be subject to fluctuations, thus further tests are necessary. These tests are known as statistical test methods. In the following, required terms and methods will be explained shortly. For a detailed introduction I refer the reader to [Bortz & Schuster 2010; NIST/SEMATECH 2014], which also serve as a foundation of this section.

The classifier’s improvement serves as a starting point and will be denoted as the hypothesis H_1. This hypothesis describes a phenomenon θ which is to be confirmed by an experiment. For this, two types of hypotheses are distinguished: a directional and a non-directional hypothesis. A directional hypothesis assumes that the phenomenon differs either positively or negatively from a given phenomenon θ_0 (cf. Eq. 4.65). The non-directional hypothesis just assumes a difference of θ and θ_0 (cf. Eq. 4.66). The certain kind of hypothesis has an influence on the later choice of the test statistic.

$$H_1 : \theta > \theta_0 \text{ or } H_1 : \theta < \theta_0 \tag{4.65}$$

$$H_1 : \theta \neq \theta_0 \tag{4.66}$$

To prove the correctness of H_1, the method called “reductio ad impossibilem” is used (cf. [Salmon 1983]). By this method it is tried to prove H_1 by showing that the experimental results are incompatible with a H_0, assuming the opposite of H_1. As H_1 is the opposite of H_0 this hypothesis must then be valid.

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta > \theta_0 \text{ or } H_1 : \theta < \theta_0 \tag{4.67}$$

$$H_0 : \theta = \theta_0 \leftrightarrow H_1 : \theta \neq \theta_0 \tag{4.68}$$

Despite this procedure, a correct decision cannot be guaranteed as it is still possible to choose the wrong hypothesis (cf. Table 4.6) because the results of the experiments are just samples. Thus, a test statistic T representing an estimated true distribution has to be calculated.

Table 4.6: Types of errors for statistical tests.

	H_0 is true	H_1 is true
H_0 was accepted	right decision	wrong decision Type II Error
H_0 was rejected	wrong decision Type I Error	right decision

Furthermore, a boundary needs to be specified to define the threshold between “still with H_0 compatible” and “already incompatible with H_0”. This boundary α is called “level of significance”. By adjusting α, the Type I error can be controlled. A conventional value for α is 0.05 that is the probability of producing a Type I error at 5 %. I further use 0.01 and 0.001 as level of significance.

Usually the p-value is given additionally on the rejection of the null hypothesis at specified α. This value specifies the actually observed level of significance and it corresponds to probability of α at which the test result is barely significant under the assumption that H_0 is true [NIST/SEMATECH 2014].

Afterwards, a test statistic T can be estimated based on the data. The calculation assumes that the null hypothesis is true. The level of significance determines the region which leads to rejection of the null hypothesis. This region of rejection now depends on the type of hypothesis. For a directed hypothesis the region is limited by a “critical value (T_{critic})”:

$$H_1 : \theta > \theta_0,\ H_0 : \theta = \theta_0 \tag{4.69}$$

$$\hookrightarrow T \geq T_{critic} \rightarrow \text{reject } H_0 \tag{4.70}$$

$$\hookrightarrow T < T_{critic} \rightarrow \text{keep } H_0 \tag{4.71}$$

In the case of a non-directional hypothesis both a too small and a too large value of T leads to a rejection of the hypothesis. Therefore, the region of rejection consists of two intervals with a halved level of significance (cf. Figure 4.14).

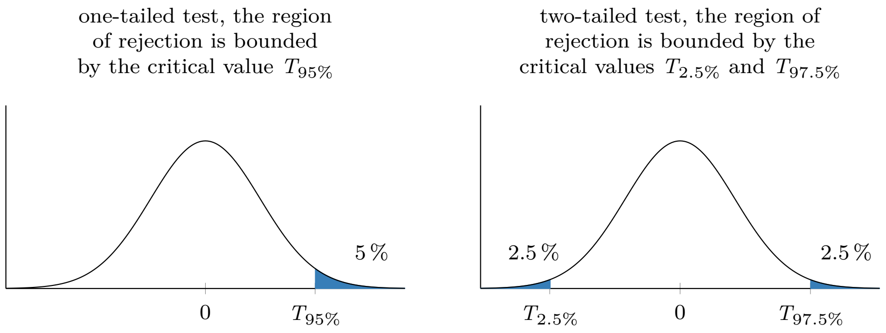


Figure 4.14: Scheme of one- and two-sided region of rejection for the directional and non-directional hypothesis $H_1 : p > p_0$ for $\alpha = 0.05$.

The specific configuration of the test statistic T depends on the intended use, number of samples, dependence or independences of the samples, and assumptions about the underlying samples' population(s) (cf. [NIST/SEMATECH 2014]). In my experiments, I am interested on testing the influence of an improved method (factor) on the recognition performance having different samples sizes in the range of 5 to 80. Therefore, I use an one-way Analysis of Variance (ANOVA)[18] as test statistic.

The basic concept of an ANOVA is to investigate whether the impact of a dependent variable is caused by specific factors. Mostly, the dependent variable describes an effect, whereas the factors are used to group the samples. The null hypothesis is that the effects present in the different groups are similar since the groups are just random samples of the same population. Hence, it is now investigated if the variance between these groups is bigger than the variance within the groups. The calculation of the test statistic for ANOVA assumes independent observations, normal distributed residuals, and homogeneous variances (homoscedasticity) between the groups. As for my later application neither the group samples nor the models used for classification depend on each other, I can state that the observations are independent.

Another assumption for ANOVA which needs to be tested is normal distributed, although it is reported in the literature that the ANOVA is quite robust against the violation of this assumption (cf. [Khan & Rayner 2003; Tan 1982]). The Shapiro Wilks test W_{SW} (cf. [NIST/SEMATECH 2014]), is a very robust test of normal distribution. It utilises H_0 assuming normal distributed samples (F_0) and tries to prove[19] H_0 for a

[18] The ANOVA is a generalization of the t-test along the number of groups. For two groups the test outcome is identical to the t-test [Bortz & Schuster 2010].

[19] Test on assumptions of statistical tests are proved against H_0, to avoid Type I Errors. Thus, these tests are calculating a probability of error that the assumption is right. In this case, an α of 0.1 is preferred.

given α (cf. Eq. 4.74). The calculation of the test statistic is given in Eq. 4.73.

$$H_0 : F_0 = \mathcal{N} \quad \text{and} \quad H_1 : F_0 \neq \mathcal{N} \tag{4.72}$$

$$W_{SW} = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{4.73}$$

where $x_{(i)}$ is the ith-smallest number in the sample, $\overline{x}$ is the sample mean and a_i are constants generated from the means, variances, and covariances of the order statistics of a sample of size n from a normal distribution $\mathcal{N}$. The exact calculation of a_i is described in [Pearson & Hartley 1972]. This test is proven to have a high statistical power especially for small test samples $n < 50$ [NIST/SEMATECH 2014]. The ANOVA is quite robust against the violation of the assumption of normal distribution.

To test homoscedasticity, usually the Levene-test W_L is used[20]. This test assumes normal distributed data but is quite robust against a violation of this assumption [Bortz & Schuster 2010]. As null hypothesis, it is assumed that the k groups have a similar variance, the alternative hypothesis assumes the opposite (cf. Eq. 4.74). To prove H_0, for all groups the deviations from the mean are calculated. If the groups have different variances, the average mean deviation should be different. The calculation of the test statistic is given in Eq. 4.75.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{and} \quad H_1 : \sigma_1^2 \neq \sigma_2^2 \tag{4.74}$$

$$W_L = \frac{(N-k)}{(k-1)} \frac{\sum_{i=1}^{k} N_i (Z_i - Z)^2}{(k-1)\sum_{i=1}^{k}(Z_{ij} - Z_i)^2}, \tag{4.75}$$

where N is the total number of cases in all groups and N_i is the number of cases in the ith group. Furthermore, Z is the mean of all Z_{ij} and Z_i is the mean on the Z_{ij} for the ith group. In this case, Z_{ij} is defined as $Z_{ij} = |\,Y_{ij} - \overline{Y}_i|$, where $\overline{Y}_i$ is the mean of the ith subgroup and Y_{ij} is the value of the measured variable for the jth case from the ith group[21]. Afterwards, the calculated W_L value is compared to the critical value $W_{L_{critic}} = F(\alpha, k-1, N-k)$ derived from statistics tables [Bortz & Schuster 2010], where F is a quantile of the F-test distribution.

The ANOVA uses an H_0 where the samples in the groups are derived from popula-

[20] As a rough rule of thumb a simplification of the F-Test (Hartley's test) can be utilised: The ratio of the largest group variance and the smallest group variance should not exceed 2.

[21] An extension of Levene's test was proposed by [Brown & Forsythe 1974] to use either the median or the trimmed mean in addition to standard mean and showed that hereby the robustness increased for non-normal data [NIST/SEMATECH 2014].

tions with the same mean values, whereas the H_0 assumes the opposite:

$$H_0 : \mu_1 = \mu_2 \tag{4.76}$$

$$H_1 : \mu_1 \neq \mu_2 \tag{4.77}$$

To calculate the test statistic, the ratio of the variance between the groups (SS_F) and the variance within the groups (SS_E) is calculated. Both variances are calculated as sums of squares. For the comparison of both values, the degrees of freedom are considered, which k being the number of groups and N the total number of samples:

$$F_{ANOVA} = \frac{\frac{SS_F}{k-1}}{\frac{SS_E}{N-k}} \tag{4.78}$$

The gained value is then compared to the F-distribution. If the calculated F_{ANOVA}-value is greater than the value of the F-distribution for a chosen α and a given degree of freedom, the differences between the groups are significant [Bortz & Schuster 2010].

When the assumptions of normality distributed samples or homoscedasticity cannot be fullfilled, a non-parametric verson of ANOVA can be used, although the standard ANOVA is quite robust against the violations of its assumptions [Bortz & Schuster 2010]. The Kruskal–Wallis one-way analysis of variance by ranks (cf. [Kruskal & Wallis 1952]) is a non-parametric method for testing whether samples originate from the same distribution. To calculate the test statistic F_{RW}, the samples are ranked by their value regardless of their group and a rank sums is calculated:

$$F_{RW} = \frac{12}{N(N-1)} \sum_{i=1}^{g} n_i \overline{r}_i^2 - 3(N+1) \tag{4.79}$$

where N is the total number of all samples across the groups g, n_i is the number of samples in group i and $\overline{r}_i^2$ is the average rank of all samples in group i. The F-value is appromimated by by $\Pr(\chi^2_{g-1} \geq K)$, which follows a chi-squared distribution, the degree of freedom is one less than the group size (cf. [Bortz & Schuster 2010]). For my significance tests, I used the Statistical Package for the Social Sciences (SPSS) software developed by IBM to calculate W_{SW}, W_L, F_{ANOVA} and F_{RW}.

In addition to the tests on statistical significance, hypotheses are also a common scientific method generally based on previous observations. According to Schick & Vaughn, when proposing a scientific hypotheses the following considerations have to be taken into account:

Testability The hypothesis must have properties that can be tested.

Parsimony Avoid the postulation of excessive numbers of entities[22].
Scope The evidently application to multiple cases of phenomena.
Fruitfulness A hypothesis may explain further phenomena.
Conservatism The degree of "fit" with existing knowledge.

In empirical investigations, most hypotheses can be seen as "working hypotheses". A working hypothesis is based on observed facts, from which results may be deduced that can be tested by an experiment. Working hypotheses are also often used as a conceptual framework in qualitative research (cf. [Kulkarni & Simon 1988]).

4.5 Summary

The successful recognition of affective states needs several steps, which were depicted and discussed in this chapter. First, the material has to be annotated using suitable methods. This includes the evaluation on the reliability of the utilised annotation. Both aspects were discussed in Section 4.1. Afterwards, adequate features describing the emotional characteristics of the underlying acoustics have to be extracted. Commonly used features and their extraction were presented in Section 4.2. These features and the previously gained annotation serve as input-pairs for the classifiers. In general several kinds of classifiers can be used for affect recognition. In my thesis I mainly utilise HMMs and GMMs. They are introduced in Section 4.3.1, in which optimal parameter sets are discussed as well. Finally, evaluation strategies, validation methods, performance measures and significance tests are presented in Section 4.4.

All these methods presented will serve as a basis for my own experiments, which are presented in Chapter 6. As discussed in Section 4.3.2, the performance of classifiers is highly depending on the applied corpora. Thus, in the following chapter the datasets used for my investigations are introduced in greater detail.

[22] This property is known as "Occam's razor": among competing hypotheses, the one with the fewest assumptions should be selected (cf. [Smart 1984]). Although this maxime represents the general tendency of Ockham's philosophy, it has not been found in any of his writings. (cf. [Flew 2003]).

Chapter 5

Datasets

Contents

In this chapter, I describe all datasets that were used in the experiments presented in this thesis. A broader overview of emotional speech databases can be found in Section 3.1. Furthermore, I recommend the following survey articles [Ververidis & Kotropoulos 2006] as well as [Schuller et al. 2008a].

I distinguish between simulated and naturalistic datasets. For simulated material the emotions are either posed by actors or non-professionals or induced by external events. These types of databases consist of recordings representing several isolated utterances. The observed emotions are mostly very expressive.

In contrast, naturalistic databases attempt to reproduce a more naturalistic reaction to events. Thus, these recordings consist of longer interactions within a naturalistic setting and represent less expressive emotional utterances. Therein, an external inducement can be persued, but a more subtle method (e.g. simulating malfunctions) is predominantly used to elicit emotional reactions. In these types of databases, it is assumed that the emotional reactions are less expressive and reproduce a broader variety of human emotional reactions.

5.1 Datasets of Simulated Emotions

In the beginning of emotion recognition from speech most databases were quite small and the recordings were based on recording experience for speech recognition corpora generation. Thus, the recordings were conducted under controlled conditions and contained short, acted emotional statements. The emotional output was known beforehand and no emotional assessment was needed. By using perception tests, the most recognisable and natural utterances were selected.

Since the emotions were acted, the expressiveness of these emotions is quite high (cf. [Batliner et al. 2000]). Although, this kind of utterance will most likely not occur within naturalistic interactions, these type of corpora served as a good starting point as the way the different emotions can be characterised by acoustic features was still unknown. The assumption made in the emotion recognition community was that experiences made with simulated material can be transferred to naturalistic material.

Additionally, as the conditions of simulated material are under control, some of these simulated databases served as a benchmark test to compare and evaluate new methods (cf. [Schuller et al. 2009a]). I have chosen one well know simulated corporus for method development and evaluation, namely emoDB.

5.1.1 Berlin Database of Emotional Speech

One of the most common emotional acoustic databases is the Berlin Database of Emotional Speech (emoDB) [Burkhardt et al. 2005]. Although this database is nowadays used widely for automatic emotion recognition, its intention was to generate suitable material to investigate and evaluate emotional speech synthesis [Burkhardt et al. 2005]. Especially due to the high recording quality, this corpus served as a benchmark of acted emotional speech, enabling the comparison of different methods for feature extraction, feature selection, and emotion classification (cf. [Schuller et al. 2009a; Schuller et al. 2007a; Ruvolo et al. 2010]).

This corpus contains seven emotional states: `anger`, `boredom`, `disgust`, `fear`, `joy`, `neutral`, and `sadness`, comparable to Ekman's set of basic emotions (cf. [Ekman 1992]). The content is pre-defined and spoken by ten actors (five male, five female). The age of the actors is in the range of 21 to 35, with a mean of 29.7 years and a standard deviation of 4.1 years.

The recordings were done in an anechoic cabin using a Sennheiser MKH 40 P 48 microphone at 48 kHz and a Tascam DA-P1 portable DAT recorder. Afterwards, the audio recordings were downsampled to 16 kHz. The distance from the speaker to the microphone was fixed to about 30 cm. Thus, the energy and intensity can be seen as a reliable measure related to the acoustic expression rather than representing a changed recording distance. Additionally, the electro-glotto-grams were recorded using a portable laryngograph.

Each emotion is uttered utilising ten different German sentences. The mean length of the recordings is 2.76 s and the standard deviation is 1.01 s. The content of the utterances is not related to the emotional expression, decoupling literal meaning from the acoustics. This procedure enables researchers to investigate the emotional acoustic variations detached from the acoustics of the content [Burkhardt et al. 2005]. This resulted in about 800 recorded utterances.

In a perception test, conducted by the corpus creators, all utterances below 60 % naturalness and 80 % recognizability of emotions were discarded, resulting in 494 phrases with a total length of 22.5 min. Unfortunately, due to the removal of several recordings, the gained distribution of emotional samples is unbalanced. The mean accuracy of this perception test for human listeners is reported as 84.3 % [Burkhardt et al. 2005]. An overview of available material per speaker-group and emotion is given in Figure 5.1. Due to the conducted perception test, the gained distribution of emotional samples is unbalanced.

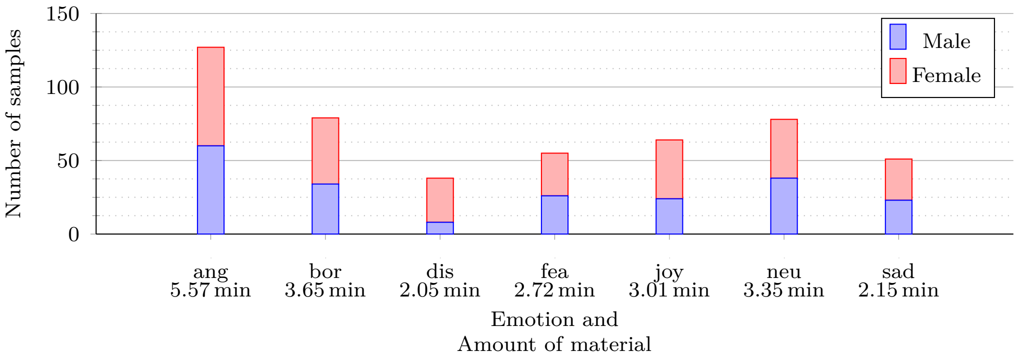


Figure 5.1: Distribution of emotional samples for emoDB. The mean length of the samples is 2.76 s, with a standard deviation of 1.01 s.

In addition to the original defined set of emotions, [Schuller et al. 2009a] generated a two-class emotional set on `arousal` and `valence` dimensions. The authors combined `boredom`, `disgust`, `neutral`, and `sadness` as `A-` (`low arousal`) and `anger`, `fear`, and `joy` as `A+` (`high arousal`). `V-` (`negative valence`) is clustered by `anger`, `boredom`, `disgust`, `fear`, and `sadness`, whereas `V+` (`positive valence`) is clustered by `happiness` and `surprise`. The reordered available material is given in Table 5.1. Thus, Schuller et al. are able to compare the results of several databases that do not cover exactly the same emotional categories but can be grouped into such kind of clusters.

Table 5.1: Available training material of emoDB clustered into A− and A+.

	A−	A+
samples	249	246
length	12.16 min	10.34 min

5.2 Datasets of Naturalistic Emotions

Batliner et al. argued for the need of real live material to be able to utilise emotion recognition from speech and pointed out recognition difficulties with realistic material [Batliner et al. 2000]. In [Grimm & Kroschel 2005] the difference between simulated and naturalistic material is investigated and it is stated that the emotional expressiveness in realistic material is much lower than in simulated material. Furthermore, in Section 6.1.2 I will show that for naturalistic material the bandwidth of emotions is expected to be much broader than in simulated material and that the set of basic emotions by Ekman is not sufficient to cover all observed emotional variations.

The research community employs several ways of generating naturalistic corpora. I refer to Section 3.1 for an overview of naturalistic corpora. For instance researchers used excerpts of human to human interactions, which are expected to contain emotional episodes, as TV-shows for VAM. But this type of database generation has the disadvantage that neither the recording conditions nor the interaction can be controlled by researchers. The material has to be taken “as is”.

Another method for collecting emotional speech data is to conduct a so-called Wizard-of-Oz (WOZ) scenario, where the application is controlled by an invisible human operator, while the subjects believe to talk to a machine, and investigate the emotional statements within an HCI. Whith this method, the researcher is in the control of the experiment. Such emotional inducement and the progress of the experiment can be defined in advance. Furthermore, the scenario can be planned in such a way that the interactions only consist of specific dialogue barriers. These are specific pre-defined breakpoints where a user reaction can be expected, but the specific type of reaction is not determined. This enables the researchers to study the whole variety of HCI. This procedure has been applied for the LAST MINUTE corpus (LMC).

Both methods have the advantage that a complete interaction can be observed, which is helpful for the needed annotation process. As it can be assumed that the labeller can consider the contextual information of the interaction, this leads to an increased reliability (cf. Section 6.1.3 on page 125). An annotation has to be performed as a proper emotional labelling of the material is missing.

Switching from simulated to naturalistic material increases variability of the occurring emotions but at the same time variations in term of acoustics, individuality, and recording conditions are increased as well. The impact of this development on the classification performance was already discussed in Section 3.3.

5.2.1 NIMITEK Corpus

The NIMITEK Corpus (cf. [Gnjatović & Rösner 2008]) was designed to investigate emotional speech during HCI. It comprises emotionally rich audio and video material that was gathered during a WOZ setup. The conducted experiments used a hybrid approach to elicit emotionally coloured expressions: On the one hand, a motivating experiment was conducted – the user was told to participate in an intelligence test. On the other hand, different strategies of the wizard were pursued to increase the stress level of the user to induce negative emotions. For example, for a short period in which the user gets to know the system, the wizard recognised the user's input correctly, and the system performed the right actions and provided useful comments and answers. However, this strategy changed in the second part of the experiment. In this part, the wizard began to simulate malfunctions of the system which provoked emotional reactions of the user by inappropriate system behaviour.

The language of the corpus is German. Ten native German speakers, three male and seven female with an average age of 21.7 (ranged from 18 to 27) participated in the experiments. Thus, the participants are belonging to the young adults age group. None of the participants had a background with spoken dialogue systems and they were not aware of the wizard at any time.

The corpus consists of ten sessions with an approximate duration of 90 min per session. Since the task of the experiment, namely simulating an intelligence test with special questions to solve, is very specific, the vocabulary is limited. However, the comments of the users were recorded as well, and since the wizard stimulated the user to express emotions verbally too, the corpus is emotionally rich [Gnjatović & Rösner 2008]. The first emotion labelling experiments, presented by Gnjatović & Rösner, showed a majority of negative emotions, as it was intended by the study's design. This labelling task was performed with German as well as Serbian native labellers, to test the influence of the lexical meaning on the annotation process. Four randomly selected sessions (approx. 5 h) were chosen from the whole corpus. As evaluation unit one dialogue turn or a group of several successive turns were defined. Each unit had to be labelled with one or more labels. The labellers had the opportunity to chose from a basic set of emotional terms, comparable to Ekman's set of basic emotions [Ekman 1992], but were allowed to extend this set. It is not clear, how the additional labels are combined into the final labels of `nervousness`, `contentment`, and `boredom`. As a result of this labelling study, it can be stated that this corpus contains several emotional utterances with a broad variability and a shift to negative emotions. As I use this corpus for labelling purposes only, I only report the emotional distribution (cf. Table 5.2) given in [Gnjatović & Rösner 2008]. For my investigations on emotions, I refer to Section 6.1.

Table 5.2: Reported emotional labels gathered via majority voting for two different labeller groups for four randomly selected sessions of the NIMITEK corpus (cf. [Gnjatović & Rösner 2008]). Total denotes the total assignments, weak denotes a majority of two labellers, strong a majority of all three labellers.

Labels	German speaker			non-German speakers		
	total	weak	strong	total	weak	strong
Anger	77	46	31	18	12	6
Nervousness	8	8	–	224	131	93
Sadness	6	7	1	1	1	–
Joy	17	14	3	1	1	–
Contentment	12	12	–	4	4	–
Boredom	9	5	4	13	10	3
Fear	–	–	–	–	–	–
Disgust	–	–	–	–	–	–
Neutral	205	124	81	54	45	9

5.2.2 Vera am Mittag Audio-Visual Emotional Corpus

The Vera am Mittag Audio-Visual Emotional Corpus (VAM) contains spontaneous and unscripted discussions from a German talk show [Grimm et al. 2008]. The creators mentioned two reasons for this kind of data source. First the discussions are quite spontaneous and thus reflect naturalistic emotions for both audio and video channel. Second, by using TV-show recordings the authors were able to collect a sufficient amount of data material for each of the 47 speakers. The chosen talk show (Vera am Mittag) assures that the guests were not paid to perform as lay actors [Grimm et al. 2008]. Ten broadcasts of this show were used to create the corpus. Each of them showed a guided discussion led by the moderator. The discussions consists of dialogues between two to five persons. This material can be seen as a collection of spontaneous naturalistic data. The acoustic recordings are available as 16 bit wav-files at 44.1 kHz downsampled to 16 kHz. Due to the origin of the data, the emotional content covers a huge range of expressiveness (cf. Figure 5.2), which is a bit contrary to commonly expected low expressiveness of naturalistic emotional data (cf. [Batliner et al. 2000]). Thus, this corpus is seen as a (special kind) of naturalistic corpora.

The broadcasts were manually segmented into associated discussions and afterwards into single utterances. These utterances mostly contain complete sentences, but also contain exclamations, "affect bursts"[23] or ellipses. Furthermore, the speakers were

[23] According to [Schröder 2003], affect bursts are defined as short emotional expressions of non-speech acoustics interrupting regular speech, for instance laughter or interjections. A more general investigation can be found in [Scherer 1994].

roughly separated into four classes denoting the expected emotional content of the speakers in terms of the amount of sentences and the spectrum of emotions. This results in 1 018 emotional sentences, 499 of "very good quality" speakers and 519 of "good quality" speakers. But the authors also stated that many utterances have to be skipped because of background noise hindering a further acoustical analysis [Grimm et al. 2008]. Due to this approach, the interaction course gets lost as the remaining utterances depict only some speakers within the ten broadcasts.

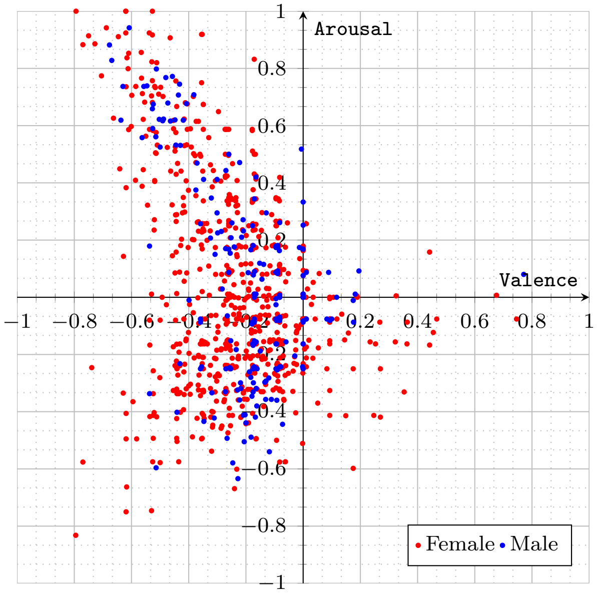


Figure 5.2: Distribution of VAM samples distinguished for male and female speakers within the `valence-arousal` space.

After this preselection, the sentences of the very good and good quality speakers were labelled using Self Assessment Manikins (SAM) (cf. [Morris 1995]). Each dimension is divided into a five-point scale in the interval of $[-1, 1]$. The labelling was done in two rounds, at first only the sentences of the very good quality speakers were labelled by 17 human annotators. As this selection was quite unbalanced in terms of emotional content, in a second round also the sentences from the good quality speakers were labelled. Unfortunately only six annotators were still available [Grimm et al. 2008]. The single evaluations for each sentence were combined afterwards using an evaluator weighted estimator[24] (cf. [Grimm et al. 2007]).

[24] The evaluator weighted estimator averages the individual responses of the labellers. For this purpose, it is taken into account that each evaluator is subject to an individual amount of disturbance during the evaluation, by applying evaluator-dependent weights. These weights measure the the correlation between the labeller's responses and the average ratings of all evaluators (cf. [Grimm & Kroschel 2005]).

In the end, this database contains 946 sentences with a total length of approx. 48 min with a mean utterance length of 3.03 s and a standard deviation of 2.16 s. This database also gives additional information on the speakers, such as age and gender. The age of participants ranges from 16 years to 69 years with a mean of 30.8 years and a standard deviation of 11.4 years. Eleven speakers are male and 31 are female. The number of samples per speaker is quite unbalanced. In total, this database has 196 samples for male speakers and 750 samples for female speakers. An overview of the available material grouped according to gender within the two-dimensional `valence-arousal` space is given in Figure 5.2.

To guarantee a sufficient number of training material as well as a robust classification, Schuller et al. proposed to consider the samples separately on the `valence` and `arousal` dimension. Therefore, all positive values on the `arousal` axis are clustered as `A+` and all negative values are clustered as `A-` (cf. [Schuller et al. 2009a]). A similar clustering is performed for the `valence` axis, to define `V-` and `V+`. This results in 445 (502) samples for `A+` (`A-`) and 71 (876) samples for `V+` (`V-`), respectively. The resulting duration for the subsequently considered `A+` and `A-` samples is given in Table 5.3.

Table 5.3: Available training material of VAM.

	A−	A+
samples	443	503
length	26.31 min	20.93 min

5.2.3 LAST MINUTE corpus

The LAST MINUTE corpus (LMC) (cf. [Rösner et al. 2012]) contains multimodal recordings of a so-called WOZ experiment with the aim to collect naturalistic user reactions within a defined course of dialogue. The participants are briefed that they have won a trip to an unknown place called “Waiuku”. As background information they were told to test a new natural language communication interface. The experimental setup was strictly designed and follows a manual (cf. [Frommer et al. 2012a]). Trained wizards were used to ensure an equal experimental cycle for all subjects.

This corpus was collected in the SFB/TRR 62 by colleagues of both the knowledge-based systems and document processing group and the department of psychosomatic medicine and psychotherapy at the Otto von Guericke University Magdeburg. Using voice commands, the participants have to prepare the journey, equip the baggage, and select clothing. A visual screen feedback was given depicting the available items per category. The task contains the need for planning, change of strategy and re-planning, and is designed to generate emotional enriched material for prosody, gestures, facial expressions, and linguistic analysis. The entire corpus contains 130 participants with nearly 56 h of material. But as the experiment was set up as an interaction study only few utterances are emotionally. Most of the material is transliterated with additional time-alignments, so that an automatic extraction of utterances is possible. More details on the design can be found in [Frommer et al. 2012b; Rösner et al. 2012].

To ensure comprehensive analyses, the experiments were recorded with several hardware synchronised cameras, microphones, and bio-physiological sensors. Four HD-cameras were utilised to capture the subject from different viewing angles and enable the analysis of facial expressions. Additionally two stereo cameras were used capturing possible gestures. For acoustic analysis, two directional microphone and one neckband headset were employed. They recorded 32 bit wav-files with 44.1 kHz. The hardware synchronization ensured that sound and video streams are synchronous over the entire recording. To enable a further analysis of body reactions, skin conductance, heartbeat, and respiration were measured as well. Exact descriptions and technical specifications can be found in [Frommer et al. 2012b; Rösner et al. 2012].

Table 5.4: Distribution of speaker groups in LMC, (after [Prylipko et al. 2014a]). Distribution of educational level is given in braces: number of subjects with higher education (first number) and others (second number). One missing data point for educational level (elderly male).

	Male	Female	Total
Young	35 (22/13)	35 (23/12)	70 (45/25)
Elderly	29 (14/14)	31 (13/18)	60 (27/32)
Total	64 (36/27)	66 (36/30)	130 (72/57)

A remarkable advantage of this corpus is the large quantity of additionally collected information on user characteristics. First of all, the experiment was conducted with several opposing speaker-groups, young vs. elderly speakers, male vs. female speakers, and subjects with higher education against others. The younger group ranges from 18-28 years with a mean of 23.2 years and a standard deviation of 2.9 years. The elder group consists of subjects being 60 years and older, the mean for this group is 68.1 years and standard deviation is 4.8 years. It was aimed to gain an equal distribution of the opposing groups on age, gender, and educational level, the resulting distribution can be found in Table 5.4.

Furthermore, the participants had to answer several psychometric questionnaires to evaluate psychological factors such as personality traits. The correlation of these traits with interaction cues is presented in Section 7.4. A brief discussion of the influence of personality traits and their utilization for emotion recognition is given in Section 2.3. The following questionnaires were used for LMC:

- Attributionsstilfragebogen für Erwachsene (Attributional style questionnaire for adults) (ASF-E) [Poppe et al. 2005]
- NEO Five-Factor Inventory (NEO-FFI) [Costa & McCrae 1995]
- Inventory of interpersonal problems (IIP-C) [Horowitz et al. 2000]
- Stressverarbeitungsfragebogen (stress-coping questionnaire) (SVF) [Jahnke et al. 2002]
- Emotion Regulation Questionnaire (ERQ) [Gross & John 2003]
- Questionnaire on the bipolar BIS/BAS scales (BIS/BAS) [Carver & White. 1994]
- Questionnaires for the attractiveness of interactive products (AttrakDiff) [Hassenzahl et al. 2003] and Questionnaire for the assessment of affinity to technology in electronic devices (Fragebogen zur Erfassung von Technikaffinität in elektronischen Geräten) (TA-EG) [Bruder et al. 2009]

In addition to these psychometric instruments, socio-demographic variables such as marital status and computer literacy are collected.

The experiment is composed from two modules with two different dialogue styles: personalisation and problem solving (cf. [Prylipko et al. 2014a]). The personalisation module, being the first part of the experiment, has the purpose of making the user familiar with the system and ensure a more natural behaviour. In this module the users are encouraged to talk freely. During the problem solving module the user is expected to pack the suitcase for his journey from several depicted categories, for instance Tops or Jackets & Coats. The dialogue follows a specific structure of specific user-action and system-confirmation dialogues. This conversation is task focused and the subjects talk more command-like. Thus, this part of the experiment has a much more regularised dialogue style. The sequence of these repetitive dialogues is interrupted by pre-defined barriers for all users at specific time points [Frommer et al. 2012a]. These barriers are intended to interrupt the dialogue-flow of the interaction and provoke significant dialogue events in terms of HCI. Four barriers are of special interest for this thesis (cf. [Panning et al. 2012; Prylipko et al. 2014a]):

Baseline[25] After the second category (Jackets & Coats) it is assumed that the first excitation is gone and the subject behaves naturally.

Listing After the sixth category (Accessories), the actual content of the suitcase is listed verbally. This cannot be interrupt by the user.

Challenge During the eighth category (Sporting Goods) the system refuses to pack further items, since the airline's weight limit is reached. Thus, the user has to

[25] This part of the experiment does not represent a barrier but serves as a "interaction baseline" from which the other barriers are distinguished. Thus, it is written in italics.

unpack things. The weights of items, the actual weight of the suitcase, and the distance to the weight limit is neither mentioned nor presented to the subject and cannot be inquired.

Waiuku At the end of the tenth category (Drugstore Products) the system informs the participant about the target location. Most subjects assumed a summer trip. But the final destination is in New Zealand, so it is winter in Waiuku at the time the scenario is settled. Thus, the users have to repack their suitcase.

Furthermore, nearly half of the subjects get an empathic intervention, after the `waiuku` barrier (cf. [Rösner et al. 2012]).

As the collection of this corpus was an ongoing procedure, there are several subparts generated from different development stages of the corpus which made use of more and more transliterations, annotations, and speakers of this database. I will distinguish them by the number of speakers: 1) the “20s” set (Gold-Standard), 2) the “79s” set, and 3) the “90s” set. In terms of acoustic analysis, these sets concentrate on the speaker turns uttered shortly after the occurring barriers. In its currently largest set, this database contains nearly 2 500 samples with a total length of about 1 h with a mean utterance length of 1.67 s and a standard deviation of 0.86 s.

The first and smallest part, the “20s” set was generated with the intention to have a number of subjects undergoing several experiments related to the SFB/TRR 62[26]. Thus it is also denoted as the Gold-Standard. It only contains 20 subjects selected with the aim, to cover a nearly equal distribution among age and gender groups. But, as only thirteen of them are usable for acoustic analysis, due to some technical problems occurred during the experiments, the achieved distribution is quite imbalanced. At least for the considered acoustic analysis. The “79s” and “90s” sets comprises all subjects where the acoustic information from the directional microphones can be utilised. The enlargement to the “90s” sets is mainly due to the evaluation of personality factors. The results will be presented in Section 7.1. The age and gender distribution of these larger sets is not as balanced as the “20s” set. In general, it can be stated that male speakers are underrepresented. The distribution of the two age-related groups is nearly equal for male speakers, especially in the “90s” set, whereas elderly female speakers are overrepresented in comparison to younger female speakers. Figure 5.3 gives an overview of the available amount of samples for each set.

[26] The other experiments are Emo Rec, conducted by the Medical Psychology Group from the University Ulm [Walter et al. 2011], and an experiment investigating the effects of delayed system response time, conducted by the Special-Lab Non-Invasive Brain Imaging at the Leibniz Institute for Neurobiology [Hrabal et al. 2013].

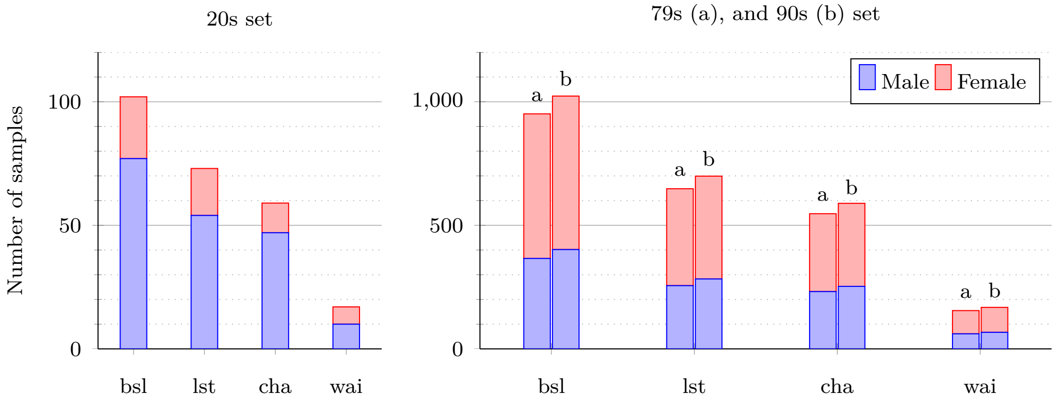


Figure 5.3: Number of samples for the different dialogue barriers separated for male and female speakers in LMC. The mean length of the samples is 1.67 s the standard deviation is 0.86 s. The term bsl denotes `baseline`, lst `listing`, cha `challenge`, and wai `waiuku`.

5.3 Summary

The presented emotional corpora are representatives of simulated and naturalistic databases. They cover a broad variety of acoustic and emotional qualities, leading to different analytical methods and specific results in the course of this thesis.

EmoDB supplies very expressive emotions that can be clearly identified by humans as well as automatic classification systems. Furthermore, the recording quality also assures an easy further processing. This database serves as a benchmark for several emotion recognition methods. The NIMITEK Corpus comprises emotional reactions of subjects within a WOZ scenario. It is a representative of databases aiming to provoke emotional reactions and to investigate the emotional speech during an HCI. VAM on the other hand represents more naturalistic emotions. But due to the different emotional annotation and the different origin, this corpus is not fully comparable to emoDB. At least the first restriction can be circumvented, when specific emotions are clustered together in a dimensional representation (cf. [Schuller et al. 2009a]). The last presented corpus, LMC, focusses more on significant communication patterns that are related to emotions. But the observable emotions within this material are even less expressive than in VAM. The recording quality of this material is very good and supplies a naturalistic participant's behaviour. Furthermore, the data is recorded directly in an HCI context and thus covers a broad number of cues and characteristics which influence the interaction.

In the following chapters, the presented corpora are used to investigate, adapt and improve methods for the speech-based emotion recogition and user behaviour in HCI.

Chapter 6

Improved Methods for Emotion Recognition from Speech

Contents

THE state of the art in automatic affect recognition from speech has been depicted in Chapter 3. I demonstrated that the recognition results decreased due to the transition from simulated material to naturalistic interactions. Thus, the research community has to increase their efforts and new methods also have to be introduced, which is still ongoing. In Chapter 4 common methods for emotion representation, emotional labelling, and recognition has been introduced. Based on this, the current chapter serves to illustrate my own work for an improved affect and emotion recognition. Therein, I rely strongly on the general processing steps of pattern recognition. Towards this goal, I investigate methods for an improved affect recogntion, namely the proper annotation of the investigated phenomena, a pre-processing step to reduce variability within the data and finally, the utilization of this variability reduction within a multimodal fusion approach. As already mentioned in Section 1.3 the following chapters represent my own work, which is clearified by proper references.

Section 6.1 presents methodological improvements for emotion annotation. First, a tool to support the transcription and annotation process is introduced. Afterwards, suitable emotional labelling methods are investigated and the influence of several contextual information on the reliability of the emotional annotation is debated.

In Section 6.2, an improvement of speaker-group specific training is introduced. This method is transferred from speech recognition using the knowledge about acoustical speaker grouping to gain an improved level of acoustic recognition. To this end, several groupings are investigated on different databases demonstrating the general applicability of this method for emotion recognition. Furthermore, this method is compared with other methods adjusting the acoustical variations.

Thereafter, in Section 6.3 I discuss my contribution, the speaker group dependent modelling appoach of Section 6.2, to multimodal emotion recognition. The focus is on naturalistic interactions, where especially the problem of fragmentary data arises.

6.1 Annotation of Naturalistic Interactions

As stated in Section 3.1, the need for datasets with naturalistic affects will be more and more at the center of focus. These naturalistic interactions are needed to provide successful HCI, but due to their origin, they contain less expressive emotional statements within a longer lasting interaction. Furthermore, this comes along with the disadvantage of a missing annotation. Annotation was given by design for simulated affects corpora, as the emotions were either acted or induced, see Section 3.1. In naturalistic interactions, the occurring affects as well as interaction cues are not known a-priori. Hence, the first step before one can train the classifier to recognise emotions on naturalistic material is the annotation of this material. This step should ensure a valid and reliable ground-truth.

Unfortunately, annotation is both a quite challenging issue and a quite time consuming task. This process mostly has to be done fully manually by well-trained labellers or experienced non-professionals, which means persons who are familiar with assessing human behaviour like psychologist but are not labellers by profession. To increase the validity in the latter case, a large number of labellers is required to judge the material. Afterwards, the resulting label is gained by majority voting. Finally, by calculating the reliability the labelling quality can be evaluated (cf. Section 4.1.3).

The pre-processing of a given dataset for a later automatic emotion recognition, is usually divided into three steps (cf. Section 4.1.1), the literal transcription, the optional linguistic annotation, and the emotional labelling. By literal transcription the spoken utterances are transferred into a textual notation. Hereby, only what has been said is written down, for instance with mispronunciations and elliptic utterances. The annotation of a text afterwards means adding prosodic and paralinguistic signs to a text, for instance, using GAT [Selting et al. 2009]. Finally, in the labelling step, further information like emotions and interaction cues are appended to the material description. As stated before, in the speech community usually the terms annotation and labelling are used synonymously.

As is common for literal transcription and linguistic annotation, either text-editors or systems, which are adapted to professionals (Folker, cf. [Schmidt & Schütte 2010]), are utilised. But for an emotional or multimodal labelling these systems cannot be used, as they do not support the specifica of emotional labelling methods (cf. Section 4.1). Hence, in Section 6.1.1 I present a tool which allows non-professionals to transcribe, annotate, and label all in one, given datasets using audio recordings, which was developed in cooperation with Ronald Böck (cf. Section 6.1.1).

The emotional annotation of naturalistic material raises two further questions. Firstly, which emotional representation should be applied for the annotation of a naturalistic interaction? Are basic emotions sufficient, should a more complicated concept as the GEW be used, or is it sufficient to apply SAM for a dimensional representation? These will be answered in Section 6.1.2. Secondly, how should the annotation be conducted? Is it sufficient for the annotation, to observe short snippets from the whole interaction? Which modalities are needed or necessary? Those questions will be discussed in Section 6.1.3.

6.1.1 ikannotate

To conduct the emotional labelling, the annotators should be supported by a tool assisting them. Several tools exist to support the literal transcription, for instance Exmaralda (cf. [Schmidt & Wörner 2009]) or Folker (cf. [Schmidt & Schütte 2010]), but for emotional labelling such tools are rare. For content analysis, the tools Anvil (cf. [Kipp 2001]) for video analysis and ATLAS (cf. [Meudt et al. 2012]) for multi-video and multimodal analysis can be used. But none of them provides the possibility to transcribe and annotate the material in a continuous way and supports the emotional annotation by depicting the corresponding annotation schemes. Therefore, Ronald Böck and I developed a tool called *i*nterdisciplinary *k*nowledge-based *anno*tation *t*ool for *a*ided *t*ranscription of *e*motions (*ikannotate*). This tool was released in 2011 and is hosted at the Otto von Guericke University Magdeburg. It was published in the ACII 2011 proceedings [Böck et al. 2011b], and demonstrated at the ICME 2011 [Böck et al. 2011a]. *ikannotate* supports both a literal transcription enhanced with phonetic annotation and emotional labelling using different methods. Thus, the tool is able to support the labelling process and analyse the emotional content of new arising naturalistic affect databases. In the following, I will present the tool briefly and focus mainly on the annotation part, as this tool has been used for the annotation experiments of this thesis (cf. Section 6.1.2 and Section 6.1.3).

The main advantage of *ikannotate* is that for each processing step – literal transcription, prosodic annotation, and the emotional labelling – the same data structure, namely XML, is used to store the relevant information. To date, the literal transcription was done by using standard text editors or tools, which are focused on these specific tasks, for instance, “Folker” provided by the Institute for the German Language [Schmidt & Schütte 2010]. They provide a more comfortable handling of the process, but are intended for users who are familiar with the specifications of annotation systems. In contrast, the tool *ikannotate* can be utilised by non-professional users as well. Both transcription and annotation is done on utterance level, which allows the user to focus on one utterance at a time. Furthermore, transcription and annotation are divided into two modules to separate both tasks.

In the first version, *ikannotate* was focused on audio materials. Thus, it handled two types of recordings: i) WAV or MP3 coded recordings of a whole session can be continuously processed and split afterwards; ii) already split recordings can be handled utterance by utterance. In the current version *ikannotate* also supports the processing of audio-visual data. The tool is written in QT4, which is a programming environment, based on C++, and can be thus used with many different operating systems. Versions for GNU/Linux, Microsoft Windows are provided. More technical details can be found in [Böck et al. 2011b].

Literal Transcription The literal transcription is done on utterance level based on audio material of the dataset. T he users have to type in the utterance, which is heard from the build-in audio player. Additionally, the start and end time of each utterance can be set, to enable an optional later audio material splitting.

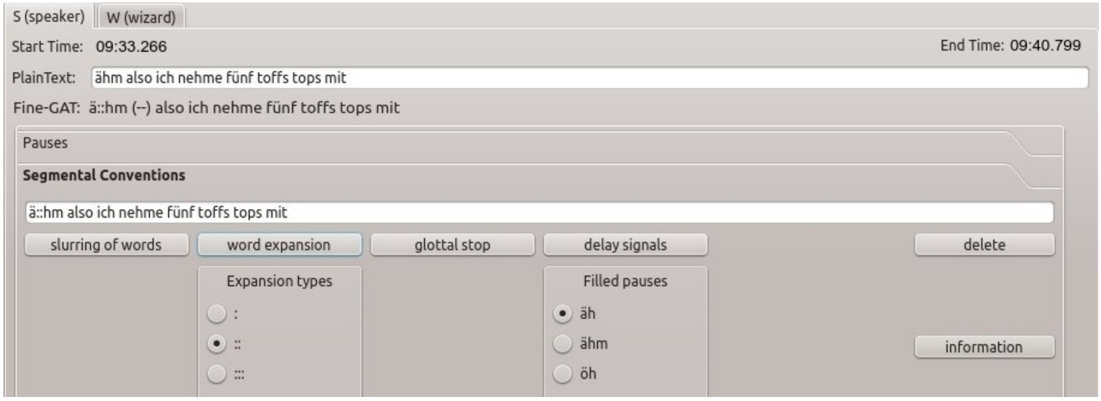
Once the processing of a sentence is finished, the current information is automatically stored on sentence level and saved in a corresponding XML file. Furthermore, it is possible to load already transcribed material and thus, stop and continue the transcription process as such. Each sentence is hence the base unit for the next steps in the process of data preparation, namely annotation.

Paralinguistic Annotation The paralinguistic annotation is an important aspect for the pre-processing of a corpus. Speech recognition and emotion recognition from speech mutually benefit especially from this information. The annotation module of *ikannotate* is based on GAT [Selting et al. 2009] (cf. Section 4.1.1).

The GAT system tries to offer a standard for the prosodic annotation of spoken speech that is already transcribed. The main ideas of GAT are derived by analysing German utterances. The system has been developed with several criteria, for instance, i) expandability, ii) readability, iii) unambiguousness, and iv) relevance. The expandability means that it is possible to work with three levels of detail. The readability ensures that also non-linguists are able to read and understand the system. The system defines exactly one sign for each linguistic phenomenon, to be unambiguous.

Furthermore, the prosodic characteristics depicted by GAT are important to interpret and analyse verbal interaction (cf. [Selting et al. 2009]).

According to GAT, *ikannotate* distinguishes three levels of granularity: 1) minimal (fewest information, usable for interaction analysis), 2) medium (enhanced information, to avoid misunderstandings within conversation), and 3) fine (containing detailed information especially about prosody). Through this concept, the annotation process can either be persued bottom up (from minimal to fine), by focussing first on most important aspects or top-down (from fine to minimal), by utilising a reduction towards specialised analyses which need only a few entities (cf. [Selting et al. 2009]).



Start Time: 09:33.266

End Time: 09:40.799

PlainText: ähm also ich nehme fünf toffs tops mit

Fine-GAT: ä::hm (--) also ich nehme fünf toffs tops mit

ä::hm also ich nehme fünf toffs tops mit

Expansion types

○ :
● ::
○ :::

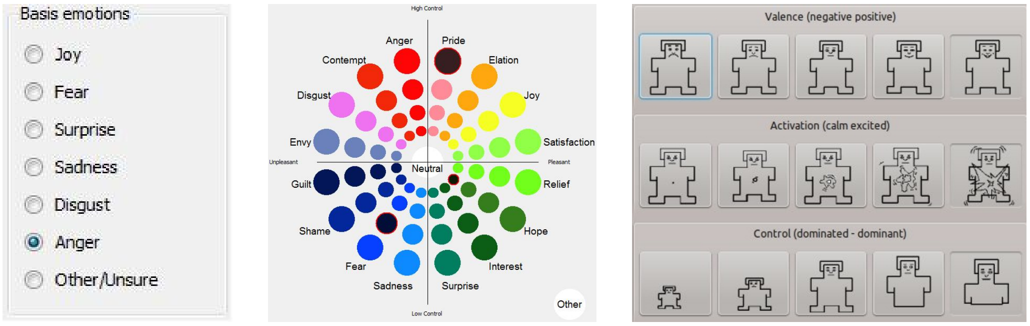
Filled pauses

● äh
○ ähm
○ öh

Figure 6.1: Excerpt of the annotation module of *ikannotate*, highlighting a word expansion (::) and a pause (- -) for the sentence “uhm so I take five toffs tops”.

The main advantage of *ikannotate* is that the annotator is supported in using the specialised signs of GAT, which are defined to mark the corresponding linguistic characteristics (cf. Figure 6.1). These signs are inserted automatically in *ikannotate* according to selected characteristics by clear words. For this, even untrained annotators, or those experienced in other annotation systems can utilise GAT.

Emotional Labelling A rather important step for pre-processing of material is the emotional labelling. As during annotation, the user is supported by *ikannotate* while labelling. That means in contrast to other tools, where the user can select an emotional label from a list of terms, *ikannotate* directly implements three emotional labelling systems to support the labeller. According to the emotional labelling methods, discussed in Section 4.1.2, the following methods are implemented (cf. Figure 6.2): 1) the list of basic emotions according to [Ekman 1992], 2) the GEW as proposed by Scherer (cf. [Scherer 2005b]), and 3) the SAM according to [Lang 1980]. Details of the implementation are given in [Böck et al. 2011b].



(a) Ekman's basic emotions (b) Scherer's GEW [Siegert et al. 2011] (c) Lang's SAM [manikins after Lang 1980]

Figure 6.2: The three emotional labelling methods implemented in *ikannotate*.

Additional Functions In addition to the introduced main components of *ikannotate*, the tool provides helpful functions for the meta-analysis of the corpora.

As already stated for the literal transcription, the utterance level is also used for emotional labelling. This approach is reasonable because the investigated emotions usually change slowly enough to be covered by one utterance [Sezgin et al. 2012]. It is additionally assumed that emotional expressions are not equally distributed over all words in a sentence [Picard 1997]. Hence, a maximum of emotional intensity could be found within an utterance. This assumption was already successfully implemented in another tool, I co-authored (cf. [Scherer et al. 2010]). Therefore, we added a supplementary module to *ikannotate* that allows the labeller to define the maximum of emotion intensity within an utterance. For the sake of convenience, users can adjust the intensity in units of words only in position and width. As it can be assumed that the height (i.e. the intensity of the perceived emotion) is already coded in the assigned emotional label, this is especially true in the case of GEW and SAM.

A further feature is the possibility that the labeller can specify his certainty about the emotional assessment. For every labelling step, the user is asked to assign the degree of uncertainty of his assigned emotional label given in the current step. This approach has the advantage that the gathered assessments can be evaluated afterwards. The recording of labellers' uncertainty makes it possible to incorporate this knowledge when labels are combined and enable the usage of the Dempster-Schafer Theory for instance for this task (cf. [Böck et al. 2013b]). To the best of my knowledge, *ikannotate* is the only tool which provides this feature, to date.

Furthermore, the tool allows to export the gathered transcription, annotation, and labels into various formats, which are directly processable by common machine learning tools, for instance HTK (cf. [Böck et al. 2011b; Böck et al. 2011a]).

6.1.2 Emotional Labelling of Naturalistic Material

As stated earlier (cf. Section 4.1), the difficulty of naturalistic emotional data is finding appropriate labels for the included expressions. Thus, beside the support of the labellers with suitable emotional annotation tools as *ikannotate*, valid and well founded emotion labelling methods also have to be utilised (cf. [Cowie & Cornelius 2003; Grimm et al. 2007]), as the annotation is a notably complex task.

Such annotation should, on the one hand, cover the wide range of observed emotions and on the other hand establish a clear and easy labelling process. The emotion recognition community is aware of these difficulties but no investigation of the effects of different labelling methods itself has been conducted to date. For the several emotional annotations used so far in databases, task or scenario specific emotional terms were preselected and used to label the whole material, for instance in the SAFE [Clavel et al. 2006] or the UAH corpus [Callejas & López-Cózar 2008] (cf. Section 3.1). Some emotional databases use a data driven approach to get their final labels a posteriori (cf. [Batliner et al. 2008; Wöllmer et al. 2009]). From the wide range of emotional labels assessed by human annotators, broader clusters are build by either a non-metrical clustering or Long Short-Term Memory Networks (LSTMNs). These clusters are than used as “classes” to train the emotional classifiers. A further approach is to merge samples that do not occur frequently enough in a remaining category mostly called `other` [Batliner et al. 2004; Lee & Narayanan 2005].

These labelling approaches have the danger that not all emotions that occur and are present in the material may be covered. This will either result in emotional classes subsuming various emotional characteristics or a quite large number of samples merged as `other` [Batliner et al. 2004; Lee & Narayanan 2005]. The classifier training has to rely strongly on the labelled material. I investigated whether it is possible to use well-founded emotional labelling methods from psychology for the labelling of simulated emotional speech data. To do so, I formulate the following hypotheses:

Hypothesis 6.1 *The application of well-founded emotional labelling methods from psychology results in a proper emotion coverage with broader emotional labels and a decreased selection of categories like* `other`*.*

Hypothesis 6.2 *The application of well-founded emotional labelling methods from psychology results in the possibility to get a proper decision for all samples.*

Furthermore, by applying these methods, it is possible to determine a relation between different emotional labels, which allows a later “informed” clustering, for rare labels. The presented investigation is published in [Siegert et al. 2011].

Methods

To test these hypotheses, I selected two categorial emotion representations (a Emotion Word List (EWL) based on Ekman's basic emotions [Ekman 1992] and Scherer's GEW [Scherer 2005b]) as well as a primitives-based representation (Lang's SAM [Lang 1980]). These three labelling methods cover a broad variety of available methods in terms of different emotional representations, number of emotional categories or range, as well as difficulty. A detailed discussion on emotional labelling methods can be found in Section 4.1.2. The investigation was conducted using *ikannotate* (cf. [Böck et al. 2011b; Böck et al. 2011a]).

To survey these studies, the NIMITEK Corpus [Gnjatović & Rösner 2008] was used as underlying database. This corpus comprises emotionally rich audio and video material and was recorded using a WOZ scenario to elicit emotional user reactions. This corpus is introduced in detail in Section 5.2.1. For the labelling task, excerpts of two different sessions where the subject should solve a tangram puzzle were chosen. Each of the two chosen parts is about 30 min long and is taken from two different persons to attenuate the influence of different speaker characteristics. For each method the labellers have to give exactly one assessment.

For the labelling of emotions, ten psychology students were employed, who studied in the first to third semester, having basic knowledge about emotion theories. None of them had ever participated in WOZ experiments or emotion labelling sessions. To compare the different emotion labelling methods, each student labelled both session parts with all three methods. To obtain an unbiased result, the methods were presented to each student in different order. Furthermore, the labelling was done on utterance level, resulting in 581 samples to be labelled.

Results

To see which differences were obtained between the labelling methods used, first the total number of all labels was analysed. For this purpose, the distribution of chosen labels for each method is compared. Hereby, I relied on the categorial labels. To allow a comparison of the resulting labels for SAM with the two categorial emotion representations, I divided the PAD-space into eight octants (cf. [Bradley & Lang 1994]) with a neutral centroid placed in the centre of the space[27]. Additionally, emotions located on a boundary area between octants are identified as "mixed emotions" and counted proportionally for all corresponding octants. The resulting distributions for

[27] The neutral centroid uses 1/5 of every dimension.

each method are given in Figure 6.3, Figure 6.4, and Figure 6.5, respectively, by calculating the mean for each emotional label between both sessions.

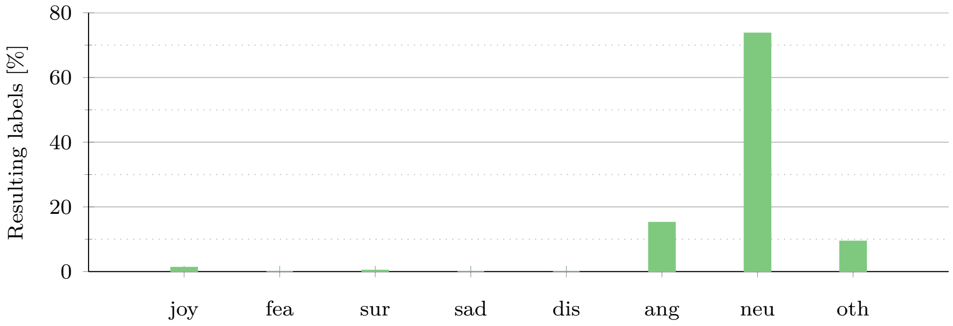


Figure 6.3: Resulting distribution of labels using a basic emotion EWL.

Regarding Figure 6.3, it can be noted that when using the basic emotion EWL, only a few terms out of all available emotional terms are chosen by the labellers to describe the observed emotions. Mostly the label `neutral` is used followed by `anger` and `other`. `Joy` and `surprise` occur only sporadically. The other three emotions (`fear`, `sadness`, and `disgust`) do not occur at all. Thus, utilising the gained emotional labels for the following classifier training, only four of six emotions can be used. Even if one accepts that in the chosen corpus the subjects spoke most of the utterances in a neutral state, the distribution of emotions is very unbalanced. Especially if it is additionally taken into account that `other` is labelled with 9.4 %, it can be assumed that this method is not suitable for labelling emotions in naturalistic HCI.

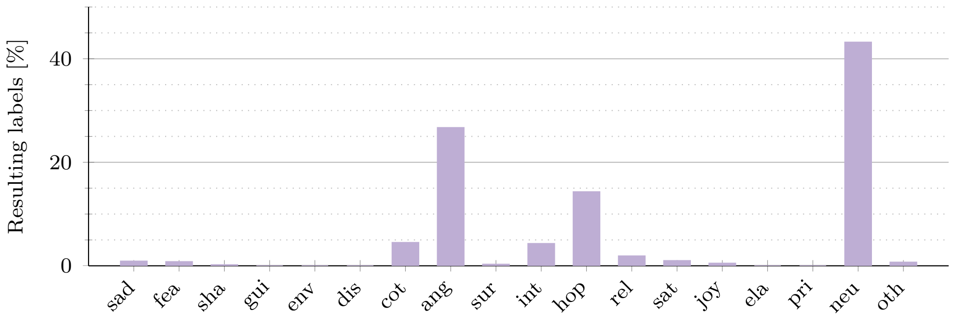


Figure 6.4: Resulting distribution of labels utilising the GEW.

Investigating the labels obtained with GEW (see Figure 6.4), a different distribution is received. First, it is noted that the label `other` is used much less than with the basic emotion EWL with only about 0.7 %. Although `neutral` is again labelled frequently, it is followed closer by `anger` and `hope`. Also, a high number of labels for `contempt` and interest and a small number for relief and fear is obtained. With GEW many more emotions beside anger can be found in the same excerpts. This observation supports the assumption that due to the richness of labels available and with their arrangement in GEW, the labellers could distinguish more emotional observations. In total, 13 of 18 emotional labels were chosen with this method.

Analysing the labels gathered by SAM, it can be noticed that only the octant where valence, arousal, and dominance are positive (+V+A+D) was not used. This may be related to the experimental design, as the interaction was mainly controlled by the system. Most labels are given for neutral, which is also represented by the EWL and GEW. But in terms of numbers, the octants +V-A-D and +V+A-D are close to neutral. All other octants are labelled with circa 2 % to 3 % of the total labels. It should be noted that due to the absence of the label other in this method, the labellers were forced to choose a category. Therefore, the results are not completely comparable to the used categorial methods (a basic emotion EWL and GEW).

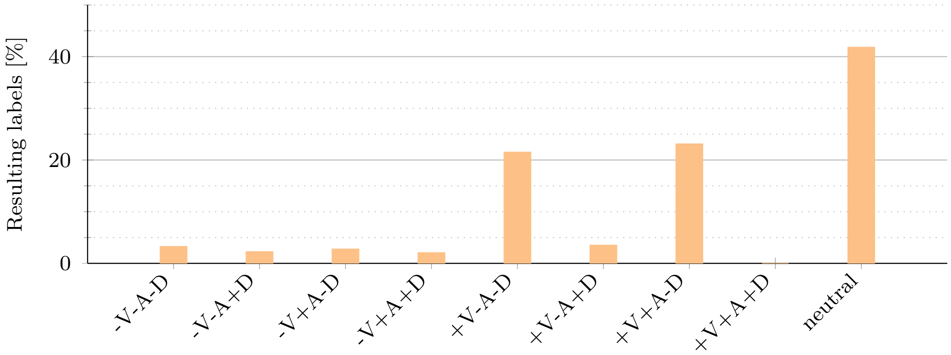


Figure 6.5: Resulting distribution of labels using SAM.

The results from this investigation can already be used to testify their usefulness. The basic emotion EWL cover far to less and just extreme cases. Thus, this method is not very useful for the emotional labelling. SAM has the problem that the interpretation of every observation is up to the labeller. This can cause some confusion which can be seen in the results, as the resulting octands are very different from the distribution expected, when comparing SAM with EWL and GEW. The results obtained with GEW let expect a high usefulness, the number of category labels as well as the expressiveness is much more suited for the annotation of naturalistic interactions than the two other methods.

The next aspect I examined is the possibility of finding a proper emotion label for each utterance of this database of naturalistic interaction. While former figures only analyse the total number of labels, now it is analysed whether it is possible to come up with a decision using Majority Vote (MV) for an emotion label for each utterance. It can be assumed that in the case of GEW, where the labeller can chose from 16 emotional terms plus `neutral` and `other`, a majority decision can hardly be reached. Whereas for basic emotions, where only six emotional terms plus `neutral` and `other` are used, the chance to get a majority decision is much higher, as the number of choices for basic emotions is smaller and well defined. Therein, the winner-take-all criterion is chosen for decision making: The emotion which is labelled by most labellers for an utterance is chosen as the observed emotion for this utterance. In comparison with the SAM ratings, the same clustering as used for the analysis of the distribution of emotional labels is applied. Thus, the ratings are analysed on the basis of octants, including the neutral octant.

Figure 6.6 depicts the results of this investigation. Therein, I distinguished the number of resulting votes for each sample, extending the standard "winner takes all" method. Obtaining an emotional label means that more than five labellers decided for the same emotional term – a single "majority vote" label could be gathered. Two groups of four or five labellers that chose the same two emotional labels is denoted as two resulting labels. In the same way, three labels and four labels are specified. This group is denoted as "multiple consensus". In the case of more than four equally rated emotions, this utterance is denoted as "undecided" in terms of emotional label.

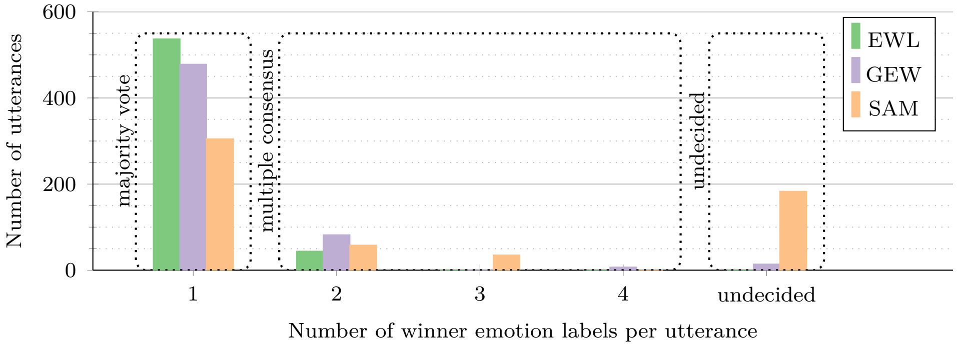


Figure 6.6: Number of resulting labels for each utterance utilising each labelling method. The total number of utterances is 581.

According to Figure 6.6, in more than 90 % of all cases, a clear decision is possible using the EWL comprising basic emotions. Multiple consensus can only be observed for a few utterances. In a detailed view of these "multiple concensus" labels, it can be found that for all of them at least one label is either `other` or `neutral`.

Examining GEW, a clear decision for most utterances (over 80 %) is possible. For a small amount of utterances ($\sim 16\,\%$) multiple concensus labels were chosen. In comparison to multiple concensus labels labels of the EWL method, a slightly different composition of these labels can be found for GEW. Although a large number of them again consists of `neutral`, the additonal emotions always have a very low intensity. Additionally, only very few of these multiple concensus labels contain the label `other`. Another type of the multiple concensus labels consists of emotions, either from the same quadrant of the GEW, mostly direct neighbours, like `contempt` and `anger`, or emotions from the same semi-circle, like `hope` and `joy` (cf. Figure 6.2 on page 118). For very few utterances (2.6 %), the labellers revealed undecided.

Taking SAM into account, only for about 50 % of all cases a clear Majority Vote decision is possible. Although a quite broad clustering into the eight octants has been utilised, for a large amount of ca. 30 % of all utterances, no decision is possible. This group consists mostly of labels with a very `low arousal` distributed closely around the `neutral` centroid. Moreover, some labellers rated a high `dominance` on samples where others did not. So it can be stated that for situations with small `arousal` or `valence`, the labelling of `dominance` is more difficult. This observation goes along with investigations by [Bradley & Lang 1994], their comparison of a Semantic Differential Scale and SAM just found a quite small correlation on the `dominance` dimension. In addition, also the assessment of `dominance` in other subjects is quite challenging, as a decision on all three emotional dimension have to be pursued.

In addition, the distribution of emotions within the utterances having a MV was also investigated. Therein, it appears that the resulting MVs for each utterance in the case of basic emotions and GEW are similar to the distribution of the overall distribution (cf. Figure 6.3 and Figure 6.4 on page 121). The SAM majority labels differ from the overall distribution. The `neutral` centroid share is approx. 83 %, the `+V-A+D` share is around 6 % and the `+V-A-D` share is approx. 4 %. The remaining 7 % of labels are distributed over the remaining octants.

Conclusion

As a result, for Hypothesis 6.1 on page 119 I can state that basic emotions are not sufficient to label emotions in naturalistic HCI, as much more variations are observable than covered by this method. Additionally, this method does not cover weaker emotions, as occurring in realistic interactions. With SAM it is possible to cover all of these variations. But for this method labellers have to identify values on three different emotional dimensions and it is difficult for non-trained labellers to assess particularly `dominance`. Furthermore, a later clustering into meaningful emotions has to be implemented. By using GEW, labellers could cover nearly all variations. This method allows a mapping of emotional categories into a two dimensional (`valence-`

`dominance`) space, as well. Thus, regarding my Hypothesis 6.1 on page 119, I can state that just SAM and GEW providing a proper emotional coverage.

Considering the second hypothesis (Hypothesis 6.2 on page 119), GEW is to be preferred as well. Although basic emotions guarantee many majority labels, the insufficiency of emotional covering suspend them from further applications. The only possibility to utilise emotional basic emotions is the generation of broader EWLs with proper emotional terms. Labelling with SAM does not provide enough majority labels, especially for the low expressiveness expectable in naturalistic HCI. This is due the non-lexical design of this method. Labellers tend to interpret the five scales more divergently than with given emotional terms (cf. [Cowie et al. 2000]).

6.1.3 Inter-Rater Reliability for Emotion Annotation

In addition to a good emotional coverage, a valid and reliable ground-truth is needed to train classifiers and detect emotional observations robustly. Therefore, the material has to be annotated to obtain adequate labels that cover important issues as well as a better system control on the interaction. But the pure annotation alone does not guarantee correct labels. Thus, at first it has to be shown that the obtained annotation is reliable. Reliability is assumed, if independent coders agree to a determined extent on the categories assigned to the samples. Then, it can be inferred that these coders have "internalized a similar understanding" [Artstein & Poesio 2008].

The Inter-Rater Reliability (IRR) has been proven to be a measurement for the quality of a given annotation. Good surveys of different reliability measures are given in [Artstein & Poesio 2008; Gwet 2008b]. The calculation of different utilised coefficients for the IRR was presented in Section 4.1.3.

In the following, I first discuss actually gained IRR-values for different databases with naturalistic affects. Here, the selection is limited to databases, where either a reliability measure is reported, or the reliability can be calculated as the particular labels of the individual annotators are given. To gain comparability, I utilise Krippendorff's α_K (cf. Section 4.1.3). Additionally, the databases are selected to cover a broad variety of annotation methods (cf. Section 4.1.2). Afterwards, I present my own studies on methods to achieve a better IRR and thus increase the reliability of the emotional labels and I therefore raise the following hypotheses:

Hypothesis 6.3 *Due to the subjective perception of emotions, the achieved IRR is generally lower than for other assessment objects.*

Hypothesis 6.4 *Utilising visual as well as context information improves the IRR.*

Hypothesis 6.5 *Preselecting emotional episodes of the interaction circumvents the second kappa paradox.*

Furthermore, I will answer the questions, which emotional representation should be applied for the annotation of a naturalistic interaction and how the annotation should be conducted. The results of these investigations are published in [Siegert et al. 2012b; Siegert et al. 2013d; Siegert et al. 2014b]

Reliability Values for Different Emotional Databases

Since the IRR is a good measure for quality of the observation, it first would be useful to get a feeling for the attainable IRRs for emotional labelling. Therefore, I present the reliability of corpora where either a given or computable IRR is on hand. As the shift towards naturalistic interaction studies was performed only recently, just a few databases fulfilling this demand are available. Even fewer reported an IRR-value for their emotional annotation. Here, different coefficients, multi-κ for VAM (cf. [Grimm et al. 2007]) and Cronbach's α_C for SAL (cf. [McKeown et al. 2010]), are used, which complicated the comparison. Therefore, the reliability values for gained annotations are re-calculated where needed, according to Section 4.1.3. The selected corpora and their applied labelling method are given in Table 6.1.

Table 6.1: Utilised emotional databases regarding IRR.

Corpora	Labelling	Specification
UAH	EWL	3 emotional terms + neutral
VAM	SAM	3 Dimensions á 5 steps
SAL	FEELTRACE	5 Dimensions
NIMITEK	EWL	6 emotional terms + neutral, other
	GEW	16 emotions + neutral, other
	SAM	3 Dimensions á 5 steps

Reliability utilising Emotion Word Lists, the UAH corpus The authors of [Callejas & López-Cózar 2008] calculated Krippendorff's α_K for the UAH. This corpus contains 85 dialogues from a telephone-based information system spoken in Andalusian dialect from 60 different users. They used emotional EWLs to discern four emotions (`angry`, `bored`, `doubtful`, and `neutral`). The annotation process was conducted by nine labellers assessing complete utterances. To infer the relation between the emotional categories, these were arranged on a 2D `activation-evaluation` space with self-defined angular distances, in the range of 0° to 180°. The authors reported an α_K of 0.338 for their "angle metric distance" [Callejas & López-Cózar 2008]. When evaluating this IRR-value with the agreement interpretations of [Landis & Koch 1977] (cf. Figure 4.7 in Section 4.1.3), only a slight agreement can be determined.

Reliability utilising SAM, the VAM corpus The VAM corpus contains spontaneous and unscripted discussions between two to five persons from a German talk show [Grimm et al. 2008]. The labelling is performed using SAM and each dimension is divided into a five-point scale in the interval of $[-1, 1]$. This database contains 499 items derived from very good quality speakers (denoted as VAM I) evaluated by 17 labellers and 519 items from good quality speakers evaluated by only six labellers (denoted as VAM II). The authors of this corpus do not provide a reliability measure. But the original labelling assessment is included, so that the inter-rater agreement using α_K (cf. Eq. 4.12 on page 60) with nominal and ordinal distances can be calculated (cf. Section 4.1.3). The resulting IRRs are given in Table 6.2.

Table 6.2: Calculated IRR for VAM, distinguishing Nominal (nom) and Ordinal (ord) Metric for each Dimension and Part.

Part	IRR		
	`Valence`	`Arousal`	`Dominance`
	nom/ord	nom/ord	nom/ord
VAM I	0.106/0.189	0.180/0.485	0.176/0.443
VAM II	0.086/0.187	0.210/0.431	0.137/0.337
VAM (I+II)	0.108/0.199	0.194/0.478	0.175/0.433

The resulting IRRs for each dimension are quite poor. When evaluated with the agreement interpretations suggested by [Landis & Koch 1977] (cf. Figure 4.7 on page 62 in Section 4.1.3), the nominal values are *poor* to *slight*, whereas the ordinal values are *fair* to *moderate*. But with a smallest value of 0.086 and a highest value of 0.478, they are far away from a *good* or *substantial* IRR of 0.6, which is expected for content analysis.

Reliability utilising FEELTRACE, the SAL corpus The SAL corpus is built from emotionally coloured conversations. With four different operator behaviours, the scenario is designed to evoke emotional reactions. To obtain annotations, trace style continuous ratings were made on five core dimensions (`valence`, `activation`, `power`, `expectation`, overall emotional intensity) utilising FEELTRACE [Cowie et al. 2000]. The number of labellers varied between two and six, the segment length was fixed to about 5 min. An example trace can be found in Figure 4.4 on page 53.

The authors in [McKeown et al. 2012] calculated the reliability using Cronbach's alpha (α_C)[28] (cf. [Cronbach 1951]) on correlation measures applied to automatically

[28] The authors of [McKeown et al. 2012] do not motivate the choice of using α_C, a discussion about the flaws of Cronbach's α_C can be found in [Schmitt 1996] and [Sijtsma 2009].

extracted functionals, for instance mean or standard deviation. I utilised the same parameters to calculate Krippendorff's α_K.

The calculation of α_K with ordinal metric distances considers the intra-clip agreement where each trace is reduced to a list of values averaged over 3 s. Hereby, every value is seen as an "independent category", this results for the FEELTRACE stepsize of $\Delta = 0.0003$ in over 6 000 "categories". But as FEELTRACE is a continuous labelling method the difference between adjacent values is mostly quite small. Additionally, I reduced the number of different categories, by discretising them. Each step size of 0.05 results in a change of the "category", this is denoted by the additional value $\alpha_{0.05}$ in Table 6.3 and reduces the number of categories to 40.

Table 6.3: IRRs for Selected functionals of SAL comparing α_K and $\alpha_{0.05}$ for the traces `intensity` (`I`), `valence` (`V`), `activation` (`A`), `power` (`P`), and `expectation` (`E`).

		I	V	A	P	E
	α_K	0.14	0.12	0.12	0.11	0.09
median	$\alpha_{0.05}$	0.01	0.01	0.01	0.01	0.01
	α_K	0.14	0.14	0.12	0.11	0.09
sd	$\alpha_{0.05}$	0.07	0.07	0.07	0.06	0.05

The results attained again show that the IRR for emotional annotation is poor in comparison to annotations of gestures, head positions, or linguistic turns (cf. [Landis & Koch 1977; Altman 1991]). Krippendorff's α_K and $\alpha_{0.05}$ achieve values lower than 0.14 (cf. Figure 4.7 on page 62 in Section 4.1.3). The substantial decrease for $\alpha_{0.05}$ is due to the increased distance that has a remarkable influence on the calculation, (cf. Eq. 4.12 and Eq. 4.14 on page 60). This emphasises the effect that the same emotion is observed differently by several labellers, as described in [Fragopanagos & Taylor 2005]. Hence, the original FEELTRACE stepsize of $\Delta = 0.0003$ is to be preferred.

Comparison of Different Annotation Methods utilising the same Database
The previous investigated corpora depicted that emotional labelling gives quite poor reliabilities compared with common interpretation intervals (cf. Figure 4.7 on page 62 in Section 4.1.3). But it cannot be ensured that the specific method chosen for each corpus did not influence the IRR. Therefore, I calculated α_K for my annotation method investigation presented in Section 6.1.2. Thus, by comparing the three different labelling methods – EWL, GEW, and SAM – on the same corpora, the resulting inter-rater reliabilities avoid inter-corpora specific issues.

The conducted emotional labelling uses the NIMITEK corpus. The annotation of 581 items utilises 8 classes comprising Ekman's basic emotions for EWL, the 18 classes of GEW, and a 5-item scale for each SAM dimension with ten labellers each. The resulting α_K can be found in Table 6.4. The IRR is calculated with a nominal distance metric (α_n) considering an equal distance of 1 between all labelling pairs (cf. Section 4.1.2). Additionally, for SAM α_o with an ordinal metric difference incorporating the item's scale range as defined in [Krippendorff 2012] is used (cf. Section 4.1.2).

For GEW I also defined a distance metric. As the labels in the GEW are arranged on a circle with different radii, a simple ordinal metric cannot be used. To indicate the distance for GEW labels, I constitute the labels as polar coordinates, using different angles and radii. The distance from one "emotion family" to another is given as the angle $\varphi = \frac{360°}{16} = 22.5°$. The radius r is set to 1 for this investigation, as no intensity measure is inferred. Thus, the distance d_{GEW} between two GEW emotion families j and l can be calculated using the Euclidean distance:

$$\mathbf{d}_{c_j c_l} = \sqrt{(\cos\varphi_{c_j} - \cos\varphi_{c_l})^2 + (\sin\varphi_{c_j} - \sin\varphi_{c_l})^2} \qquad (6.1)$$

To include the labels `neutral` and `no emotion`, their angles are defined as 0° and 180° and the radius is set to 2 for both of them. This is contrary to the graphical presentation given in Figure 4.1 on page 49, but needed as a huge difference between these both assessments is necessary. The resulting IRR utilising this distance metric is denoted as α_{GEW} in Table 6.4.

Table 6.4: Comparison of IRR for EWL, GEW and SAM on NIMITEK. α_n denotes α_K with a nominal distance measure, α_o and α_{GEW} utilises specific distance measures as described in the main text.

Method		α_n	α_o	α_{GEW}
WL		0.208	–	–
GEW		0.126	–	0.336
SAM	V	0.217	0.387	–
	A	0.204	0.399	–
	D	0.165	0.384	–

Comparing the achieved inter-rater reliabilities for the three used annotation methods with the reliabilities on the presented corpora, I can state that these results confirm the first hypothesis (cf. Hypothesis 6.3 on page 125) of a low inter-rater agreement for emotional annotation especially for data of naturalistic interactions. The values for the reliability utilising a nominal metric distance are between 0.165 and 0.217, which means a poor to lower fair agreement when applying the interpretation scheme by [Landis & Koch 1977]. As the values for the different methods are in a similar range, I suppose that the specific method does not affect the inter-rater reliability and therefore the choice is only a matter of the investigated scientific question or the desired emotional labels.

Methods for Inter-Rater Reliability Improvement of Emotional Labelling

The comparison of the so far reported inter-rater reliabilities for emotional annotation reveals rather poor agreement measures for emotional data. Additional efforts are needed to increase the reliability and thus gain an improved annotation quality. I claim that in addition to the pure annotation methods, contextual information is also required. This will lead to an improvemed number of correctly assessed emotions. As sources of relevant contextual information the available modalities (audio and video) and the presence of surrounding information like interaction progress is investigated.

Therefore, I conducted two experiments. In the first experiment, the influence of the perception of audio and video information and the influence of the dialogue course on the reliability is investigated. The second experiment investigates a further improvement of reliability, focussing on certain parts of the emotional material. These parts are preselected due to expectable emotional reactions, utilising a-priori knowledge about the dialogue course.

The studies are conducted using the LMC (cf. [Rösner et al. 2012]) containing multimodal recordings of 130 native German subjects collected in a WOZ experiment. A detailed description is given in Section 5.2.3. As the expected time effort for labelling is up to eight times higher than the length of the material, a subset of approx. 2 h is selected. Furthermore, the events are split regarding each subject's answer as one utterance resulting in 405 snippets with an average length of 11 s ranging from 3 s to 50 s.

Increasing Reliability by Adding Context – Study I In this first investigation, I tested the following hypothesis: A greater IRR can be achieved when both acoustic and visual information are present and the annotators have information about the interaction evolvement (Hypothesis 6.4 on page 125). Therefore, different labelling tasks were designed, varying the different modalities and the interaction course.

To obtain the different labelling sets, two dependent variables and their expressions are defined. The modality consists of the values `audio only`, `video only` and `multimodal`. The interaction is either `random` or `ordered`. This results in six different experimental sets where both variables take all their defined values (cf. Table 6.5).

To receive proper emotional labels, I utilised the results of the study on emotion labelling methods presented in Section 6.1.2 (cf. also [Siegert et al. 2011]). There, the differences between three labelling methods and the observed emotional labels for a similar HCI were investigated. To adopt the emotional labels onto the expected outcome of LMC, a preliminary study were conducted to define a set of suitable emotional terms for the actual corpus by using GEW with the possibility to add additional terms (cf. [Siegert et al. 2012b; Siegert et al. 2013d]). This studies revealed the following labels: `sadness`, `helplessness`, `interest`, `hope`, `relief`, `joy`, `surprise`, `confusion`, `anger`, `concentration`, and `no emotion`. These eleven labels are combined into a EWL as a definition of a dimensional relation between these labels comparable to a GEW is not the focus of this thesis.

The study to increase the IRR by adding context was conducted with ten labellers, all of them with psychological background. To support the labellers during their annotation, a version of *ikannotate* was utilised (cf. Section 6.1.1 and [Böck et al. 2011a]). The labellers could see or hear the current snippet and could choose one or several emotional labels from the presented EWL. Each snippet contains the users' command as well as the wizards' response. The order of presented snippets was predefined. Additionally, the tool forced them to watch the complete snippet and assess it afterwards, a repeated view of the current snippet was possible.

To calculate the IRR, Krippendorff's α_K is utilised (cf. Section 4.1.3). As EWLs do not allow to determine a relation between the labels, the nominal distance metric is used to calculate α_K for all six sets (cf. Table 6.5). Furthermore, a MV is utilised to obtain the resulting label for each item. Therein, only the assessment where more than five labellers agreed on the same emotional state is used as a resulting label. The number of attainable labels is given in Table 6.5.

Table 6.5: Number of resulting MVs and the IRR for the investigated sets. The total number of items is 405.

interaction	modality	MV	α_K
`random`	`audio only`	306	0.195
	`video only`	297	0.183
	`multimodal`	312	0.251
`ordered`	`audio only`	375	0.341
	`video only`	375	0.323
	`multimodal`	393	0.398

Comparing the resulting values for the random and ordered experiments, it should be noted that the IRR is higher for an ordered presentation. The annotators agree more in emotional items when having knowledge about the interaction course. Thereby, the gained α_K is increased by approx. 37 %, from 0.251 to 0.398. This extended the results of [Cauldwell 2000; Callejas & López-Cózar 2008] that an interaction history is needed to give a reliable assessment. Concurrently, the number of MV-labels increases only by 20 % from 312 to 393. Although this behaviour seems a bit strange, increasing the reliability with about 37 % while increasing the number of MV only by about 20 %, it can be noted that the number of labellers involved in the MV is increasing. While in the first case the MV-label consists mostly only of 5-6 labellers, in the latter mostly 8-10 labellers are involved in an MV.

When comparing the single-modality sets with the multimodal sets, it can be noted that multimodal sets reach a higher reliability regardless of the ordering of the presented snippets. This observation is in line with the investigation of [Truong et al. 2008], stating that the influence of multimodal context information (in their case audio plus video) increased the IRR on spontaneous emotion data using a dimensional approach.

When further comparing the ordered sets with their counterparts in terms of modality, I can state that all `ordered` sets have a higher α_K value than the `random multimodal` sets. Furthermore, there is no or only a small difference in the resulting number of labels for the `audio only` and `video only` sets. Rather, the reliability varies substantially. The higher reliability for `audio only` sets suggests that humans can assess audible information better with less confusion and higher agreement among each other. But this may also be caused by additional contextual information as, for instance, the wizard utterances are also audible.

In contrast, [Lefter et al. 2012] reported a large confusion for multimodal annotation compared to sets with limited modality information. These findings could not be reproduced. This can be caused by the type material, as the investigated material in [Lefter et al. 2012] is quite different from LMC used in the actual investigation. LMC provides frontal perspectives showing a single person with high quality acoustics in comparison to a full scene perspective with far range group acoustics for the material used in [Lefter et al. 2012]. Also, [Douglas-Cowie et al. 2005] found that an annotation on video data alone is not sufficient, which is in line with our results.

Considering the `ordered` vs. `unordered` cases, I conclude that the incorporation of context information, especially the interaction course, improves the inter-rater agreement, ss the context allows the annotator to judge the user's speaking style and interaction course for material extracted in the HCI context. Considering both, ordered video and audio recordings of an individual person in a frontal perspective, the reliability can be increased further. Using this method, the reliability in the presented investigation could be increased from *poor* (0.195) to *fair* (0.398) according to [Landis & Koch 1977] (cf. Figure 4.7 on page 62).

In comparison to the investigations of [Callejas & López-Cózar 2008], where a slight decline of α_K from 0.3382 to 0.3220 is observed utilising emotional EWLs on the UAH corpus, in the presented investigation an improvement can be observed. Such an improvement is given only when both audio and video modalities are available. The declined reliability reported in [Callejas & López-Cózar 2008] is due to the high bias for `neutral` labels with more than 85 % of the material. This can be attributed to the "first" and "second Kappa paradox" (cf. [Callejas & López-Cózar 2008; Feinstein & Cicchetti 1990]), as α_K averages over judgement pairs in the same way as Fleiss' K. α_K decreases when the prevalence of one label is rising (first) or the distributions of agreement are not equal (second). The most frequent label in our material has an occurrence of about 50 % of all labels.

Increasing Reliability by Pre-informed Selection – Study II Another consideration how the reliability could be improved is to use additional knowledge to preselect certain parts of the data material where an emotional reaction of the subject is more likely. Hereby, neutral and emotional parts are tried to be rebalanced, which circumvents the "second Kappa paradox" (cf. [Feinstein & Cicchetti 1990]). This investigation should prove Hypothesis 6.5 on page 125. This supports the annotator, as only those parts of the experiment have to be annotated where an emotional reaction of the subject is expected. Furthermore, this decreases the annotation effort, as only a subset, but rich of emotional reactions, has to be regarded. This method was also applied in a framework, I co-authored (cf. [Böck et al. 2013a]). A reliable experimental design is necessary to define such parts. This is is given by LMC's design.

The LMC defines several so-called barriers where the subject is faced with a suddenly arising problem (cf. Section 5.2.3). In this investigation, the "weight limit barrier" is considered, which is called `challenge` (`cha`). During the category "Sportswear", the system for the first time refuses to include selected items because the airline's weight limit for the suitcase is reached. All subjects are faced with this barrier. The barrier has been overcome when the subject could successfully pack something in the suitcase again. For the related utterances an emotional reaction of the subject, indicating `irritation` or `confusion`, is to be expected [Rösner et al. 2012] (cf. Figure 6.7). This results in much less items than regarding all utterances of all subjects. Considering the same items used for the first study, 87 snippets instead of 405 are obtained, but with assured comparability.

Furthermore, the same ten labellers using the same EWL and the same version of *ikannotate* [Böck et al. 2011a] (cf. Section 6.1.1) as for the first study were employed.The results from the previous study were incorporated as well. Thus, the multimodal utterances are presented in an ordered way.

For these 87 utterances, an IRR of 0.461 is attained. A final MV label can be specified for 86 of 87 items (99 %). These percentage values are higher than those from the previous study, where α_K only reached 0.398 and for only 97 % of all items a final label could be determined. This can be attributed to the "second Kappa paradox", which describes the phenomenon that α_K decreases when the distribution of categories (here emotions) is not balanced [Callejas & López-Cózar 2008; Feinstein & Cicchetti 1990]. As most of the interaction's parts are supposed to be neutral, this emotion is over-represented. Hence, a pre-selection of emotional parts helps to balance the classes. The achieved IRR is now considered as *moderate* when using the agreement interpretations from Figure 4.7 on page 62.

Although only a subset is considered, the reliability increased and the available classes are more balanced, which is necessary for classifier training, as the attained reliability guarantees valid emotional labels. This method can be further used for semi-automatic annotation, where automatically pre-classified utterances are manually corrected. This can be even used for multimodal data (cf. [Böck et al. 2013a]).

Resulting Emotions of LMC's basline and challenge barriers

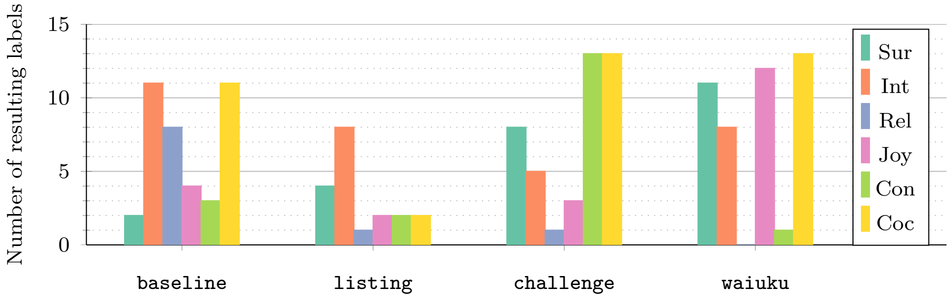


Figure 6.7: Distribution of MV emotions over the events of LMC, taking only labels gathered with wizard responses into account (cf. [Siegert et al. 2012b]).

Finally, I will indicate the outcome of the investigation of this section and depict the resulting distribution of MV over the different barriers of the LMC (cf. Section 5.2.3). The comparison of the emotional labels for the four different barriers is depicted in Figure 6.7. It reveals that interest is nearly equally spread over all barriers. Each of the emotional states relief, joy, and confusion has a maximum at different barriers, namely baseline, waiuku, and challenge. The emotion concentration is labelled for all barriers, except for listing. Interestingly, in listing, where the system lists all packed things, the votes for concentration are quite low. This has to be further investigated. One first guess might be that the system is talking too much and the participant gets bored. To select the barriers worth for later automatic analyses, the amount of the user's speech data has to be taken into account. As for `listing` and `waiuku`, the user is hardly involved as only information is presented, I conducted my further experiments to distinguish `baseline` and `challenge`.

Discussion

Comparing the gained IRRs for the presented corpora in Section 6.1.3, I conclude that for all emotional labelling methods and types of material the reported reliabilities are very distant from the values seen as reliable. Even well known and widely used corpora like VAM and SAL reveal a low inter-rater agreement. Krippendorff's α_K utilising a nominal distance metric is between 0.01 and 0.34. Using an ordinal metric increased the α_K only up to 0.48 at its best. Both cases are interpreted as a *slight* to *fair* reliability (cf. Figure 4.7 on page 62). Thereby, I see Hypothesis 6.3 as proved.

Furthermore, the comparative study of three different annotation methods reveals that the methods themselves only have a small impact on the reliability value. This is supported by my first investigation that influence of various emotional labelling methods on the emotion coverage of naturalistic emotional databases is just small (cf. Section 6.1.2). Hence, it is up to the researcher to choose an adequate method, suitable for the current investigation. Furthermore, I was able to show that interpretation schemes used so far are inappropriate for emotional annotation, as even for well conducted and secured assessments the achieved reliability values are below 0.46, which is only seen as *moderate* (cf. Figure 6.8).

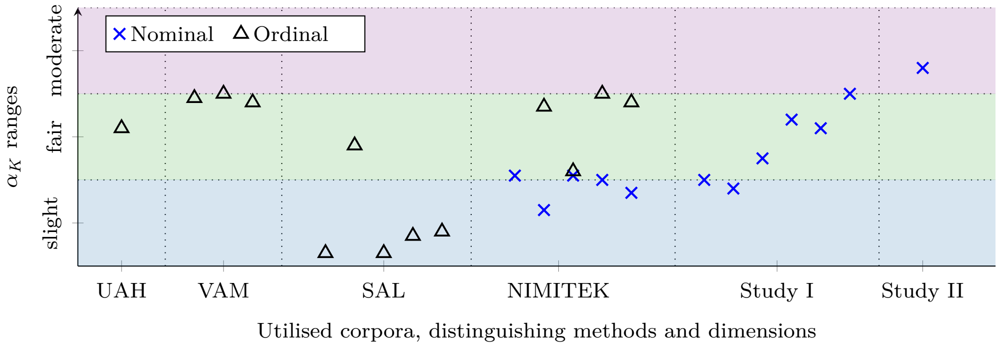


Figure 6.8: Compilation of reported IRRs, plotted against the agreement interpretation by Landis & Koch 1977 (after [Siegert et al. 2014b]).

Afterwards, two of my own approaches were presented, to increase the reliability on LMC as representative of naturalistic emotional speech databases. In the first study, it could be shown that the reliability can be increased by utilising both audio and video recordings of the interaction as well as presenting the interaction in its natural time order, which confirms Hypothesis 6.4. The second study further increased the reliability by preselecting emotional parts. Therein, a method able to circumvent the "second kappa paradox" is presented (cf. Hypothesis 6.5). All reported α_K values of this section are given in Figure 6.8 for comparison and arranged according to the interpretation scheme by Landis & Koch [Landis & Koch 1977].

6.2 Speaker Group Dependent Modeling

As presented in Section 4.3.3, speaker variablities influence the performance of ASR systems. A reduction of speaker variabilities within the recognition process, by suitable pre-processing methods, increases the ASR performance. Emotion recognition from speech utilises the same acoustic features, for instance MFCCs, pitch, and energy as well as derived functionals (cf. Section 4.2 and [Böck et al. 2010; Schuller et al. 2009a]). Moreover, the same classifiers, like SVMs, GMMs, are utilised [Ververidis & Kotropoulos 2006; Zeng et al. 2009]. The incorporation of age and gender differences has also already been used to improve speaker recognition [Kelly & Harte 2011; Kinnunen & Li 2010], but has been only rarely used for emotion recognition (cf. Section 3.4.2). Psychological research has empirically investigated the influence on age and gender on emotional regulation, showing that both characteristics influences the way users are reacting emotionally. Thus, I raise the following two hypotheses:

Hypothesis 6.6 *The age-related change of speakers' acoustics suggests that emotion recognition can be improved by considering age and gender as group characteristics.*

Hypothesis 6.7 *Using speaker group dependent modelling results in a higher improvement than performing an acoustic normalisation where the differences in emotional regulation between the speaker groups is not considered.*

Until now, only the obvious gender dependency has been investigated to some amount. My studies extended these experiments by incorporating age dependency (cf. Section 6.2.2). These investigations are performed on LMC, which is very prototypical in terms of age groups and presented in Section 6.2.3. Afterwards, I extended the investigations to additional databases covering high quality, simulated emotions as well as a different age grouping (cf. Section 6.2.4). Hereby, I will examine Hypothesis 6.6. The different results are compared and shortly discussed in Section 6.2.5. Afterwards, the results of my approach are compared with VTLN, a speaker characteristics' normalisation technique (cf. Section 6.2.6), to prove my second hypothesis.All presented results are published in [Siegert et al. 2013c; Siegert et al. 2014d].

6.2.1 Parameter tuning

As stated in Section 4.3.2, the number of mixtures and iteration steps have to be tuned for GMM-classifiers. Reported results of optimal number of mixtures suggest 80 to 120 mixtures for databases of simulated emotions and around 120 mixtures for material of naturalistic emotions [Böck et al. 2012b; Vlasenko et al. 2014].

Furthermore, a Feature Set (FS) that will be used for all forthcoming investigations also has to be defined. I rely on the investigations performed by [Böck et al. 2012b; Cullen & Harte 2012; Vlasenko et al. 2014], who stated that a spectral feature set (MFCC) is well-suited for emotion recognition from speech. Thus, I utilised a GMM-classifier with twelve MFCCs their deltas and double deltas. Additionally, I use three prosodic characteristics, the fundamental frequency (pitch) (F_0), the short-term energy (E) and zeroth cepstral coefficient (C0). The exact configuration in terms of temporal characteristics and channel normalisation will be investigated further. Meaning and extraction of these features and techniques is described in Section 4.2. For the classifier training and testing, I use HTK (cf. [Young et al. 2006]).

I examined two different kinds of emotional material, emoDB as a representative of databases with simulated emotions and LMC as naturalistic emotional corpus. As validation method, LOSO is chosen (cf. Section 4.4.1). For emoDB a set of six emotions is utilised (`anger`, `boredom`, `fear`, `joy`, `neutral`, and `sadness`), discarding `disgust` as only few speakers provided samples (cf. Section 5.1.1). LMC is utilised with the two dialogue barriers `baseline` and `challenge` (cf. Section 5.2.3). To compare the results, I calculated the UAR (cf. Section 4.4.2) as the samples for each class in the utilised corpora are quite unbalanced. The overall performance for the different LOSO folds is given as mean over all speakers's UARs. Significant improvements are denoted and an ANOVA (cf. Section 4.4.3) is used. The test of pre-conditions and all individual resutls are reported in [Siegert 2014].

Varying the Number of Mixture Components In comparison to other studies investigating an optimal number of mixture components, I conducted my experiments using 1 to 200 components to expose a broader range of mixtures. It can be assumed that due to the larger feature space and emotional variations and due to the naturalness of emotions within the LMC a larger number of mixture components is also needed. Additionally, these experiments could give insights on how GMMs behave if more than an optimal number of mixtures are used. Ti this end, I used a step-width of 1 for the first ten mixture components, afterwards the step-width of 10 is applied. The resulting classification performance (UAR) is depicted in Figure 6.9.

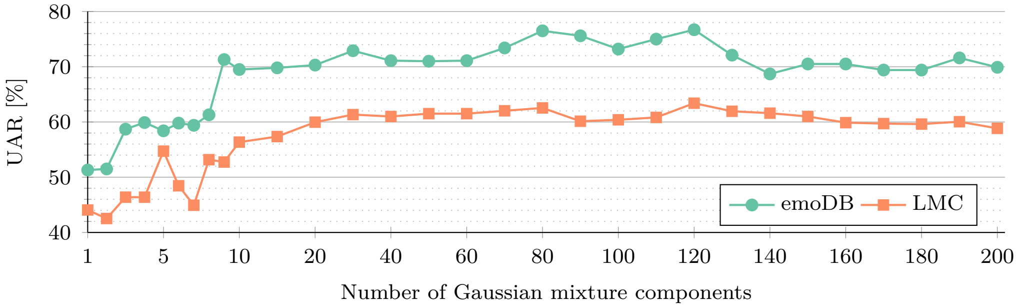


Figure 6.9: UARs for databases of simulated and naturalistic emotions in the range of 1 to 200 Gaussian mixture components. For each added component, 4 iterations were used.

Regarding the results, it can be noted that the classification performances remain quite stable when more than 20 mixtures are used, especially for naturalistic material. The performance only varies in the range of 6 % on emoDB and just 4 % on LMC. Furthermore, the GMMs have two peaks of classification performance – at 80 and at 120 mixture components. This behaviour can be observed independently of the material's type. Due to HTK's splitting of the "heaviest" mixture component, a general prototype model converges more and more in a specialised model representing the acoustic features that characterise one specific emotion.

Employing more than 120 mixture components leads to a model losing its generalisation ability, as the classification performance decreases. This behaviour can be seen as an over-generalization. Due to HTK's training algorithm, the heaviest mixture is perturbed and thus class-specifica are abandoned. This approach leads to smoothed models that will in the end even out the feature differences.

Similarly to [Vlasenko et al. 2014], in my experiments I observed a performance drop when more than 80 mixtures are utilised for a database of naturalistic emotions. But the performance is increasing and outperforming the classification achieved with 80 mixtures, if 120 mixture components are used. As the effect could also be observed when performing own experiments on VAM, I assume that the additional variations inferred by prosodic features cause this effect. But for a simulated database an optimal number of mixture components of 117 is reported by [Vlasenko et al. 2014], which is comparable to my observation. For LMC a number of 120 mixtures also appears to be a boarder, as higher numbers led to a decreased classification performance, although the observed decreasing is not as strong as on emoDB.

Varying the Number of Iteration steps Furthermore, I conducted investigations on number of iterations in the range of 1 to 20. Therein, the two numbers of mixture components having the best performance (80 and 120) are chosen. The experiments are repeated with the same features on the same two databases using a LOSO validation. The results are depicted in Figure 6.10.

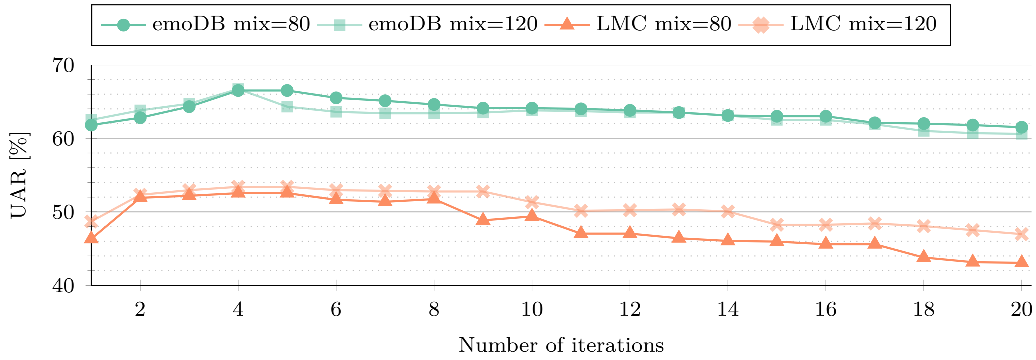


Figure 6.10: Gained classification performance (UAR) for databases of simulated and naturalistic emotions utilising different iterations steps in the range of 1 to 20.

Comparing the different numbers of iterations, the well-known over-fitting problem of can be observed (cf. [Böck 2013]). Applying more than 5 iterations for emoDB decreases the classification performance down to 70.6 % UAR using 20 iteration steps. This is a decrease of about 8 %. In the case of naturalistic emotional material, the number of mixtures has an influence on the recognition using different iterations. Having 120 mixture components, the decrease shows up later, when more than 8 iteration steps are used. Having just 80 mixture components, the performance decreases already for 6 iterations. Furthermore, a remarkable performance drop can be observed when more than 8 iterations are used. It can be assumed that the higher number of mixtures is able to compensate the over-fitting, as the additional components can cover more characteristics. For my investigations, I choose 4 iteration steps, as this is optimal in terms of recognition performance as well as computational load.

Including Contextual Characteristics According to the previous experiments, I choose 120 mixtures and 4 iteration steps as parameters for all further experiments using GMMs. The classification is actually based on short-term segments and thus, utilising only information on the actual windowed speech signal. But it is known that the incorporation of contextual characteristics for emotion recognition increases the recognition ability (cf. [Glüge et al. 2011; Kockmann et al. 2011]). By such an approach, acoustic characteristics of the surrounding frames can be included to evaluate the actual short-term information.

Two methods can be used to incorporate context. At first, delta and double delta regression coefficients (Δ and $\Delta\Delta$) can be employed. Secondly, the SDC-coefficients can be used, utilising a much broader contextual information. SDC-coefficients were proposed in [Torres-Carrasquillo et al. 2002] and led to an improved language identification performance. The applicability for emotion recognition was investigated by [Kockmann et al. 2011]. Both methods are described in Section 4.2.2.

Both approaches cover different ranges of temporal context, the Δ-coefficients incorporate ± 2 frames, the $\Delta\Delta$-coefficients covering ± 4 frames. The employed SDC-coefficients comprise a range of ± 10 frames[29]. These experiments are again conducted on the same databases of simulated emotions (emoDB) and naturalistic emotions (LMC). Therein, I utilise the same features (12 MFCCs, pitch and energy) and model parameters (120 mixtures, 4 iteration steps) as identified above. LOSO is applied as validation strategy. Furthermore, the significance of improvement is tested by using ANOVA (cf. Section 4.4.3), the pre-conditions (Normal distribution, homoscedasticity) are fulfilled (cf. [Siegert 2014]). The results are presented in Figure 6.11.

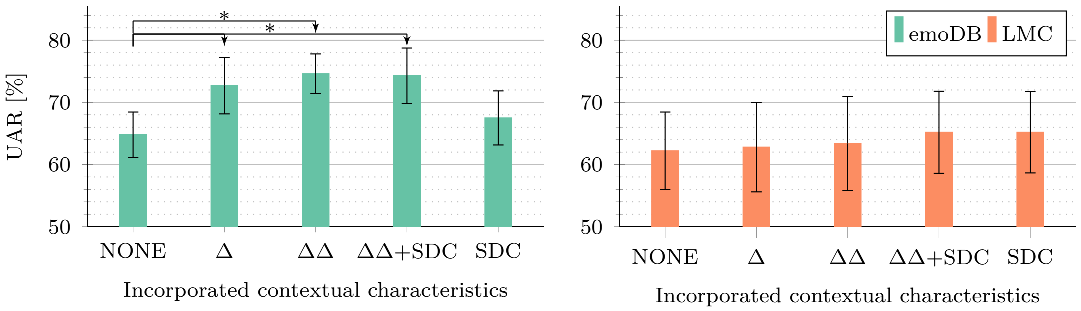


Figure 6.11: Gained UARs (mean and standard deviation) for databases of simulated and naturalistic emotions using different contextual characteristics. The $\Delta\Delta$ coefficients include the Δ coefficients. Stars denote the significance level: * ($p < 0.05$), ** ($p < 0.01$).

For both databases the incorporation of regression coefficients (Δ and $\Delta\Delta$) increases the recognition performance. In case of emoDB, the performance raises significantly when Δ ($F = 4.5856$, $p = 0.0462$) and $\Delta\Delta$ ($F = 10.1899$, $p = 0.0051$) are included. Using only SDC-coefficients on emoDB does not significantly ($F = 0.5652$, $p = 0.4619$) increase the recognition. When applying the same features on LMC, the classification performance does not increased by the same amount as on emoDB. In contrast to emoDB, the incorporation of SDC-coefficients raises the classification performance by about 3 %, but this is not significant ($F = 2.1686$, $p = 0.1429$).

When performing the same experiments with VAM (cf. [Siegert et al. 2014d]) the classification results on VAM show the same behaviour as on emoDB. Thus, I assume

[29] In my application of SDC-features, I rely on the following parameters, suggested by [Kockmann et al. 2011]: i is in the range of $[-3, 3]$, $P = 3$, and $L = 1$.

that the expressiveness and location of emotions within an utterance rather than the type of emotion in terms of simulated or naturalistic is important. The same aspects have been raised by [Wöllmer et al. 2010; Glüge et al. 2011]. From this it can be concluded that for emoDB and comparable datasets, ΔΔ-coefficients should be incorporated, while for LMC it is worth to additionally incorporate SDC-coefficients.

Comparison of two Channel Normalisation Techniques After identifying the optimal number of mixture components, iteration steps, and amount of contextual characteristics, I also investigated the impact of channel normalisation techniques. I concentrated on CMS and RASTA-filtering (cf. Section 4.2.1). Both are applied in ASR-systems to reduce the influence of different channels and noise conditions. The experiments are again conducted on a simulated (emoDB) and a naturalistic (LMC) emotional database. I used the same features (12 MFCCs, F_0, E, C0) and model parameters (GMMs, 120 mixtures and 4 iteration steps) and validation strategy (LOSO) as before. The results are presented in Figure 6.12.

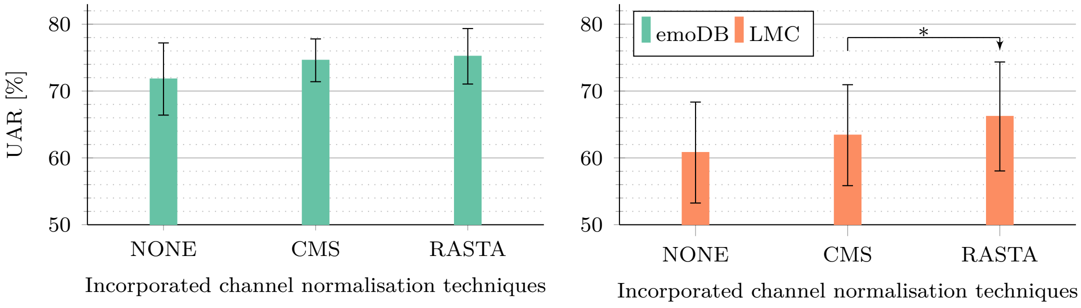


Figure 6.12: Gained UARs (mean and standard deviation) for databases of simulated and naturalistic emotions using different channel normalisation techniques. Stars denote the significance level: * ($p < 0.05$).

The incorporation of channel normalisation techniques increases the classification performance for both emoDB and LMC. The CMS is done by estimating the average cepstral parameter over each input speech file [Young et al. 2006]. This approach compensates long term effects, for instance, from different microphones or transmission channels. An absolute improvement of 2.8 % for emoDB and 2.6 % on LMC can be achieved. But these are not significant.

Adding RASTA-filtering results in an improved recognition for corpora with a high variability in recordings. On LMC an absolute performance increasement of 5.4 % is achieved in comparison to no channel compensation. Performing the RASTA-filtering for studio-recorded corpora as emoDB raised the result by only 3.6 % in comparison to no compensation. At the same time, the standard deviation on emoDB increases from

6.4 % to 8.3 %. This approach leads only to a significant improvement when inferring RASTA-filtering on LMC ($F = 5.1035$, $p = 0.0253$).

Resulting Feature Sets

According to my experiments on parameter tuning, I can now define a standard set of features (FS1), namely 12 MFCCs, C0, F_0, and E, where CMS as channel normalisation technique is applied. The Δ and $\Delta\Delta$ coefficients of all features are used to include contextual information.

Additionally, I tested RASTA-filtering as alternative channel normalisation technique and used SDC-coefficients as further contextual characteristics in different combinations. The four different feature sets are given in Table 6.6.

Table 6.6: Definition of Feature Sets (FSs).

Set	Features			
	Spectral / Prosodical	Context	Channel	Size
FS1	MFCCs C0 F_0 E	Δ, $\Delta\Delta$	CMS	45
FS2	MFCCs C0 F_0 E	Δ, $\Delta\Delta$	RASTA	45
FS3	MFCCs C0 F_0 E	SDC	CMS	120
FS4	MFCCs C0 F_0 E	SDC	RASTA	120

In the next section, I will present my results using speaker group dependent modelling. For this, I first need to define the different age and gender groupings. Additionally, the utilised corpora and their speaker groups are depicted.

Afterwards, the achieved results for each corpus are presented and discussed. Furthermore, the results are compared across the different corpora, as intermediate results. Then I compare the results achieved with my method with the alternative approach of acoustic normalisation.

6.2.2 Defining the Speaker-Groups

There is almost no research on the definition of proper speaker-groups for emotion modelling (cf. Section 4.3.3). Thus, it has to be initially clarified which grouping should be used. Therefore, I rely on research for automatic age and gender detection from speech and will shortly depict the speaker groups utilised in this field of research. This will hopefully lead to speaker groups being able to distinguish the acoustic characteristics for emotion recognition as well (cf. Table 6.7).

Most researchers agree on distinguishing the following age groups: children and teens as well as young, middle aged, and senior adults. But there is no general definition, where to draw the borders. In most cases, it can be stated that young adults are younger than 30 years [Lipovčan et al. 2009], sometimes 35 years is also used as the limit [Meinedo & Trancoso 2011]. Seniors are older than 55 or 60 years [Burkhardt et al. 2010; Hubeika 2006]. The middle aged adults cover the interval between these two groups.

As gender groups, male and female speakers are considered. The children's voice has major differences in comparison to adult voices [Potamianos & Narayanan 2007], thus they should be grouped into a separate gender-group. For children below twelve years of age the grouping can be conducted regardless of their gender, because no statistically significant gender differences exist [Lee et al. 1997]. After undergoing the voice change, boys can be differentiated from girls and both are considered as teens.

Table 6.7: Overview of common speaker groups distinguishing age and gender. The speaker groups written in *italics* are not considered in this thesis.

Age		Gender	
		male speaker (m)	female speaker (f)
children (c)	<12	*young children (yc)*	
teens (t)	<16	*male teens (mt)*	*female teens (ft)*
young adults (y)	<30	young male adults (ym)	young female adults (yf)
middle aged (m̲)	>30	middle aged males (m̲)	mmiddle aged females (m̲f)
seniors (s)	>60	senior male adults (sm)	senior female adults (sf)

Psychological research also identified several user groups in terms of emotion regulation, which is responsible for emotional expressions [McRae et al. 2008; Lipovčan et al. 2009]. The authors of [Butler & Nolen-Hoeksema 1994] investigated differences of male and female college students responding in a depressed mood. Whereas the authors of [Gross et al. 1997] investigated the influence of ageing on emotional responses and state that older participants showed a lesser expressivity. These considerations suggest that in addition to the speakers' gender their age must also be taken into account for a robust emotion classification. When investigating the emotional speech content of children it can be noted that they utter their emotional state differently than adults. Especially when talking to machines, children use an enriched, wordily way of talking [Potamianos & Narayanan 2007], which also encourages a separate grouping from the emotion recognition perspective.

6.2.3 Initial Experiments utilising LMC

As the focus of this thesis is on the improvement of emotion recognition for naturalistic HCI, the Speaker Group Dependent (SGD) modelling approach will be initially applied to LMC, introduced in detail in Section 5.2.3. This corpus has the advantage to contain four roughly balanced groups in terms of age and gender, namely young and old male as well as female speakers (cf. Table 6.7). The age structure of these two age groups is as follows: 18-28 years for the young and over 60 years for the senior adults. Thus, the given speaker groups represent fairly extreme cases in terms of age.

For the classification I concentrated on two key events of the experiment, where the user should be set into a certain clearly defined condition: `baseline` (`bsl`) and `challenge` (`cha`). During `bsl`, the subject feels comfortable and has been adapted to the experimental situation and the first excitement has gone. Within `cha` the system creates mental stress by suddenly claiming to reach a previously fixed luggage limit. This causes a "trouble in communication", which can be seen as a critical point within a dialogue [Batliner et al. 2003]. A detailed description of the corpus is given Section 5.2.3. The emotional assessment is described in Section 6.1.3.

For this investigation, I utilised the "79s" subset (cf. Section 5.2.3). As classification baseline, I used the Speaker Group Independent (SGI) set, which contains all 79 speakers regardless of their age or gender grouping. The different age-gender groupings together with the number of corresponding speakers are depicted in Figure 6.13. To perform my experiments, I rely on a-priori knowledge about the age and gender grouping for each speaker, on the basis of the speakers' transcripts. To reference the specific groupings, I use the following abbreviations: the grouping by age is denoted as age specific Speaker Group Dependent (SGDa), the grouping by gender is denoted as gender specific Speaker Group Dependent (SGDg), and age and gender specific Speaker Group Dependent (SGDag) denotes the simultaneous grouping by age and gender.

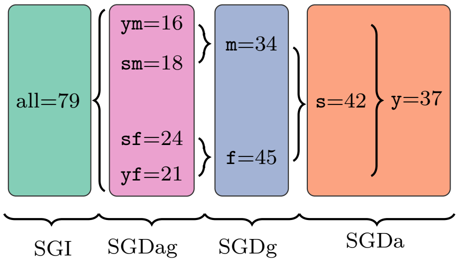


Figure 6.13: Distribution of subjects into speaker groups and their abbreviations on LMC.

To generate the material for training and testing, the associated dialogue turns from each speaker of the utilised subset were extracted automatically on the basis of the transcripts. Afterwards, the resulting parts were manually corrected concerning wizard utterances and unusual noise and the turns are chunked into single phrases. This results in 2 301 utterances with a total length of 31 min (cf. Table 6.8). It can be noticed that the distribution of samples is unbalanced, as a higher amount of samples is available for the `baseline` condition.

Table 6.8: Overview of available training material of LMC.

	bsl	cha
samples	1 449	852
length	18.68 min	12.22 min

According to my experiments on parameter tuning (cf. Section 6.2.1), I use the four defined set of features, comprising the following acoustic characteristics: 12 MFCCs, C0, F_0, and E. The Δ and $\Delta\Delta$ coefficients of all features are used to include contextual information. As channel normalisation technique CMS is applied. Additionally, I tested RASTA-filtering as an alternative channel normalisation technique and incorporated SDC-coefficients as further contextual characteristics, as these features have been proven to be promising for the naturalistic emotional database LMC. As classifiers, GMMs with 120 mixture components utilising 4 iteration steps, are trained applying a LOSO validation strategy. The baseline classification results using the SGI set are given in Table 6.9. The significance is calculated again utilising standard ANOVA (cf. Section 4.4.3) when pre-conditions (Normal distribution, homoscedasticity) are fulfilled, or the Kruskal-Wallis non-parametric ANOVA if the pre-conditions are not fulfilled. Details on the calculation can be found in [Siegert 2014].

Table 6.9: Applied FSs and achieved performance in percent of the SGI set on LMC. The significance improvement against the SGI FS1 result is denoted as $p < 0.01$. Explanations of the FSs, see Table 6.6.

Feature Set	UAR [%]	
	mean	std
FS1	63.4	15.1
FS2	66.2	16.3
FS3	64.9	12.9
FS4	68.7	10.0

Regarding Table 6.9, it can be observed that inferring longer contextual characteristics as well as RASTA-filtering increases the classification performance of the SGI set by 1.5 % to 3 %. Applying both techniques increases the performance by 5.3 % which is significant ($F = 6.7654$, $p = 0.01$).

After defining the achieved UAR on the SGI set, I performed the same experiments using the previously defined sets SGDa, SGDg, and SGDag. Doing so, the speakers are grouped according to their age and gender in order to train the corresponding classifiers in a LOSO manner. The number of available speakers and the speaker groups are depicted in Figure 6.13 on page 144. The results for each speaker group based on the UAR's mean and standard deviation are presented in Table 6.10.

Table 6.10: Achieved UAR in percent using SGD modelling on LMC. The outlier, showing worse results than SGI, is highlighted. The significance improvement against the corresponding SGI results is denoted as follows: $p < 0.05$, $p < 0.01$, $p < 0.001$. Explanations of the FSs, see Table 6.6.

Grouping	UAR [%]							
	FS1		FS2		FS3		FS4	
	mean	std	mean	std	mean	std	mean	std
m	65.2	13.1	72.1	12.2	71.3	8.1	73.7	9.3
f	64.8	11.8	74.1	12.4	72.3	11.3	73.4	13.3
y	67.6	14.9	71.1	14.3	69.3	10.4	72.3	13.3
s	64.1	11.3	69.4	10.3	71.3	13.4	71.8	13.2
sf	65.9	11.7	70.8	10.8	67.2	12.1	71.8	11.3
yf	77.3	9.6	79.1	11.2	75.3	9.3	76.2	10.7
sm	66.7	12.1	72.8	9.1	72.8	12.7	70.2	12.1
ym	63.9	10.9	70.6	11.2	71.1	8.7	67.8	10.7

In comparison to my own achieved SGI results, it can be observed that nearly all SGD results outperform the corresponding SGI result. The gender-differentiation benefits when RASTA-filtering is performed or SDC coefficients are included. These techniques show significant improvements on FS2 (f: $F = 7.9409$, $p = 0.0056$), FS3 (m: $F = 9.2174$, $p = 0.0030$, f: $F = 10.41050$, $p = 0.0016$) and FS4 (m: $F = 6.1911$, $p = 0.0143$). The age differentation shows a significant improvement when SDC coefficients are included (y: 4.0682, $p = 0.0437$, s: $F = 6.636$, $p = 0.0112$)

A distinguishing of both age and gender groups also leads to a remarkable improvement in comparison the the SGI classification. The best improvement, can be observed for FS2. In this case, the yf shows a significant improvement for all feature sets (FS1: $F = 16.0039$, $p = 0.0001$ FS2: $F = 11.6454$, $p = 0.0009$ FS3: $F = 11.1457$, $p = 0.0008$ FS4: $F = 9.0638$, $p = 0.0033$). But when SDC coefficients are utilised together with RASTA-filtering (FS4), the UAR of SGI is better than the results for the young male (ym) speaker group.

Finally, I combined the different results of speaker groupings, as only a combination of groupings allows a proper comparison to SGI. For instance the results for each male and female speaker are put together to get the overall result for the SGDg set. This result can then be directly compared with results gained on the SGI set. The outcome is shown in Figure 6.14 according to the different feature sets. Additionally, significant improvements against the corresponding FS of the SGI result are pointed out.

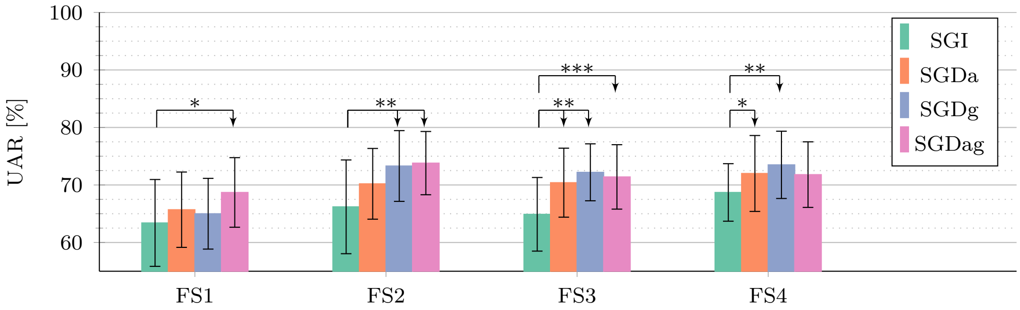


Figure 6.14: UARs for 2-class LMC for SGI and different SGD configurations utilising GMMs and LOSO on different feature sets. Stars denote the significance level: * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$).

The classification achieved with LMC shows that SGDag grouping could significantly outperform the SGI results for nearly all feature sets. The incorporation of either RASTA-filtering (FS2) or SDC-coefficients (FS3) contributes to a significant improvement also for SGDa or SGDg classifiers (cf. Figure 6.14). The best result of 73.3 % is achieved using FS2 and the SGDag approach ($F = 8.70644$, $p = 0.0032$).

When comparing the achieved UARs utilising either age or gender groups, it can be seen that for FS2, FS3, and FS4 the gender grouping outperforms the age grouping by 1.4 % to 3.2 %. But for FS1, where neither RASTA-filtering nor SDC-coefficients are incorporated, the gender grouping falls below the performance of the age grouping. Thus, no statement could be made if an age or a gender grouping should preferred. Hence, further experiments are needed.

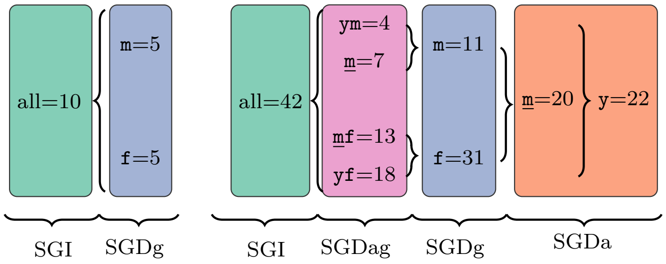
6.2.4 Experiments including additional Databases

In the previous section, the SGD approach has been successfully applied on LMC as a naturalistic emotional database. In this section, the method is extended on the databases emoDB (cf. Section 5.1.1) and VAM (cf. Section 5.2.2). EmoDB is a database of simulated emotions containing high quality emotionally neutral sentences for six emotions. VAM represents a naturalistic interaction corpus containing spontaneous and unscripted discussions between two to five persons from a German talk show.

To be able to compare the results on emoDB, VAM, and LMC, I use the two-class emotional set generated by [Schuller et al. 2009a]. For emoDB, they defined the combinations of boredom, disgust, neutral, and sadness as low arousal (A−) and anger, fear, surprise, and joy as high arousal (A+) (cf. Section 5.1.1). For my investigation om VAM, I also distinguish between A− and A+.

By using the simulated emotion database, I want to prove that my method is also applicable for very clear and expressive emotions, which may neglect the speaker variabilities. In this case it can be assumed that the acoustical differences between different emotions becomes very apparent. VAM is used, as it contains a different age-grouping than LMC and thus allows to investigate the age grouping influences as well. For a detailed introduction of these databases, I refer the reader to Chapter 5.

Unfortunately, for emoDB no distinction into age groups is possible (cf. Section 5.1.1). A special feature of VAM is its age distribution. In contrast to LMC, where by design two very opposed age groups (younger than 30 years and older than 60 years) are apparent, VAM utilises middle aged adults ($\underline{\mathtt{m}}$) ranging from 30 to 60 years in addition to young adults (y) (cf. Section 5.2.2). The used speaker groups and amount of speakers for emoDB and VAM are given in Figure 6.15. The age and gender grouping rely on a-priori information, given in the corpus description.



(a) emoDB (b) VAM

Figure 6.15: Distribution of subjects into speaker groups and their abbreviations.

According to the experiments on parameter tuning (cf. Section 6.2.1) and for comparing the results with LMC, I used the same set of features, namely 12 MFCCs, C0, F_0, and E. The Δ and $\Delta\Delta$ coefficients of all features are used to include contextual information. CMS is applied as channel normalisation technique (FS1) and I tested RASTA-filtering as alternative channel normalisation technique (FS2). Thus, the applied feature sets are the same as used for LMC (cf. Table 6.9 on page 145).

As classifier GMMs with 120 mixture components and 4 iteration steps are trained using a LOSO validation. As classification baseline, the SGI set is trained, disregarding the age-gender groupings. Furthermore, I define a two-class problem on emoDB by applying the clustering suggestion presented in [Schuller et al. 2009a] to get the two clusters: `low arousal` (`A`−) and `high arousal` (`A`+). My own results achieved with the SGI set and reported results from other research groups are given in Table 6.11.

Table 6.11: Achieved UARs of the SGI set on emoDB and VAM. The particular six-class problem of emoDB and the two-class problems are considered. For comparison the best reported results are given (cf. Table 3.2 on page 39). The FSs are explained in Table 6.6.

	emoDB				VAM	
	six-class		two-class			
	mean	std	mean	std	mean	std
FS1	74.6	6.4	92.6	5.6	70.1	15.8
FS2	75.2	8.3	92.7	5.6	71.8	17.1
	best reported result by other researchers					
	86.0[1] *Acc*		96.8[2] UAR		76.5[2] UAR	

[1] [Vogt & André 2006] [2] [Schuller et al. 2009a]

Regarding Table 6.11, it can be seen that the application of RASTA-filtering increases the classification performance of the SGI set of all used corpora. Although the improvement is between 0.6 % for the six-class problem of emoDB and 1.7 % for the two-class problem of VAM, none shows a significant increase.

My achieved results on emoDB and VAM are below the best results reported by other researchers. The authors of [Vogt & André 2006] applied a Naive Bayes classificator with a gender dependent set of features. Their gender-differentiation classification uses a-priori gender information as well. The authors of [Schuller et al. 2009a] achieved their results by using 6 552 features from 56 acoustic features and 39 functionals together with either an SVM (emoDB) or GMM (VAM) classifier.

Next, I performed the same experiments using the previously defined sets SGDa, SGDg, and SGDag on each database. To do so, the speakers are grouped according to their age and gender to train the corresponding classifiers in a LOSO manner. The number of available speakers and the speaker groups for each corpus are depicted in Figure 6.15. Afterwards, the mean and standard deviation of the achieved UARs are calculated. The individual results are presented in the following.

Speaker Group Dependent Classification Results on emoDB

Looking at the speaker group dependent results on emoDB, it should be noted that individual results are always substantially better than the gained baseline classification

Table 6.12: Achieved UARs in percent using SGD modelling for all available speaker groupings on emoDB. The FSs are explained in Table 6.6.

classes	Grouping	UAR [%]			
		FS1		FS2	
		mean	std	mean	std
6	m	75.9	7.3	75.7	8.3
	f	77.3	7.8	78.1	8.9
2	m	97.1	5.6	97.1	5.8
	f	98.2	2.6	98.2	3.0

(cf. Table 6.12). This is independent of the number of emotional clusters or the applied feature set. For the six-class problem, both speaker group dependent results are below the best reported result of 86.0 % [Vogt & André 2006].

Inspecting both gender groups independently, it is apparent that the improvement for females is about 2 % better than the recognition for males. This can be attributed to the smaller amount of training material available for the male group. On average, the emotional classifier could be trained with 6.6 utterances from each male speaker whereas from each female speaker 8.5 utterances can be used. The combined SGDg results just achieves approx. 77 % for both FSs (cf. Figure 6.16).

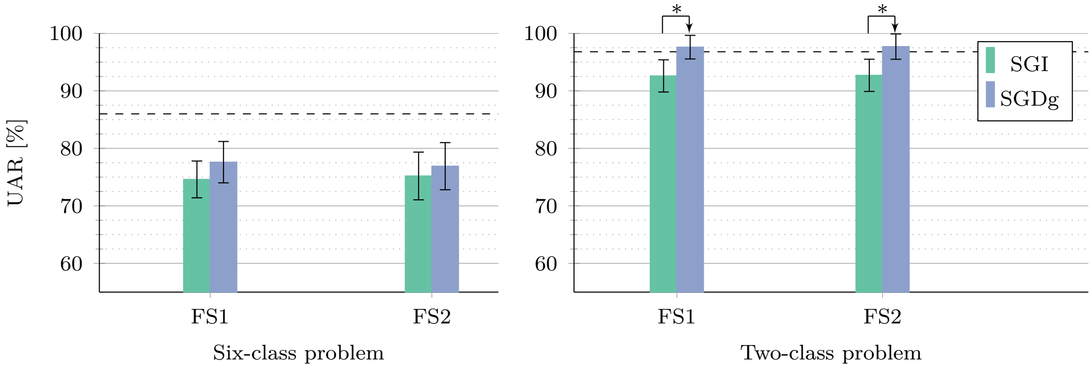


Figure 6.16: UARs in percent for emoDB's two-class and six-class problem comparing SGI and SGDg utilising LOSO validation on different feature sets. For comparison, the best reported results (cf. Table 3.2 on page 39) are marked with a dashed line. The stars denote the significance level: * ($p < 0.05$).

In contrast, the SGD results on two-class problem are outperforming the classification result of 96.8 % from [Schuller et al. 2009a], for both applied feature sets. In this case, a sufficient amount of material to train a robust classifier is available with more than 19 utterances per speaker. The combined results of both gender-specific classifiers are depicted in Figure 6.16. In comparison to the SGI results the SGD classifiers achieved an absolute improvement of 4 % to 5 % for both FSs. This improvement is significant for FS1 ($F = 4.8791$, $p = 0.0272$) and FS2 ($F = 4.48238$, $p = 0.0281$).

In summary, it can be stated that the low amount of training material for the six-class problem drives the presented approach to its limits. The tripling of the training material by combining certain emotions, in the second case shows a significant improvement of recognition performance for the investigated feature sets in comparison to SGI modelling. Furthermore, the SGD approach could slightly outperform the results of [Schuller et al. 2009a], whereby much fewer features (45 vs. 6 552), but a-priori knowledge about the speakers' group belonging are used.

Speaker Group Dependent Classification Results on VAM

Table 6.13: Achieved UAR in percent using SGD modelling for all available speaker groupings on VAM. The outlier, showing worse results than SGI is highlighted. The significance level is denoted as $p < 0.01$. The FSs are explained in Table 6.6.

Grouping	UAR [%]			
	FS1		FS2	
	mean	std	mean	std
m	72.2	16.3	73.4	15.3
f	80.1	12.9	80.2	14.3
y	76.8	13.1	77.1	15.4
$\underline{\mathtt{m}}$	76.1	13.1	75.1	14.2
$\underline{\mathtt{m}}\mathtt{f}$	75.9	13.9	75.2	13.2
yf	77.9	13.6	80.1	14.2
$\underline{\mathtt{m}}$	71.7	14.1	71.3	13.1
ym	75.3	12.7	71.8	14.8

The investigation of the speaker group dependent results for VAM, reveals that all individual results are better than the gained baseline classification (cf. Table 6.13). The utilised feature set only has a slight influence on the performance. A significant improvement of up to 10 % for the SGDg classifiers can be achieved. In this case, the female speaker group benefits from the high amount of training material resulting in significant improvements: FS1 ($F = 8.3160$, $p = 0.0052$) and FS2 ($F = 11.4370$, $p = 0.0010$), while for the m group, the small amount of training material is apparent.

In SGDa the improvement is between 3 % to 7 % for both speaker groups, which is only significant for the young speaker's group with FS2 ($F = 4.8509$, $p = 0.0277$).

Distinguishing both age and gender information (SGDag) demonstrates that both female speaker groups (yf and m̲f) have a high improvement of 3.4 % to 8.4 % while the improvement for the male speaker groups (ym and m̲) is only between −0.5 % to 5.2 %, whith one group (m̲), below the baseline. None of these combined groupings (m̲f, yf, m̲, and ym) show a significant improvement.

The best reported result of 76.5 % (cf. [Schuller et al. 2009a]) could only be outperformed by a few speaker groups (f, y, and yf). In the SGDag case, only the ym group outperformed this result. Especially, speaker groups containing middle aged speakers (m̲, m̲f, and m̲) show results clearly behind the reported result of 76.5 % (cf. Table 6.13). Also, the young male speakers stay behind the results of [Schuller et al. 2009a].

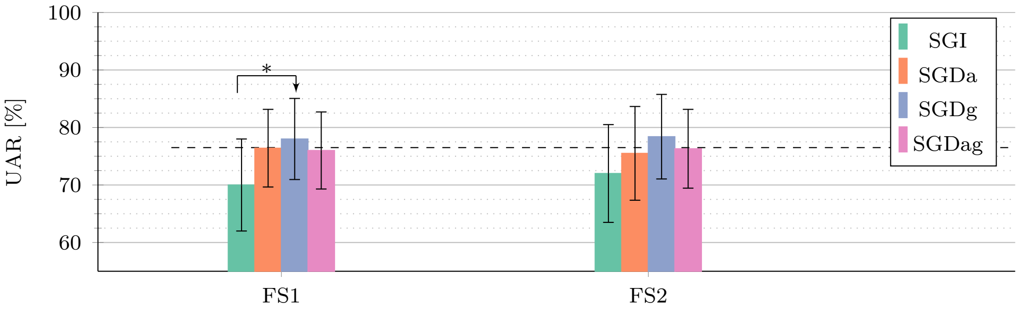


Figure 6.17: UARs for the two-class problem on VAM for SGI and SGDg and LOSO validation on different FSs. For comparison, the best reported result (cf. Table 3.2 on page 39) is marked with a dashed line. The star denotes significance level: * ($p < 0.05$).

Utilising VAM allows to examine a grouping of the speakers on different age ranges (y and m̲). The grouping comprises the speakers' age (SGDa), the speakers' gender (SGDg), and both information (SGDag), see Figure 6.17. For all three combinations, a substantial improvement was achieved in comparison to the baseline classification (cf. Table 6.11 on page 149). The improvement using FS1 for the SGDg approach ($F = 5.8451$, $p = 0.0178$) is significant.

Unfortunately, the SGDag achieves lower results than the classification using only one characteristic (age or gender). This is mostly caused by the declined performance for the m̲ group. It must be further investigated whether this can be attributed to the few amount of material available or to the fact that the present acoustical differences within the middle aged adults are larger than these in young adult's group.

In terms of the investigated features, the classification performance of the SGDag approach is mostly declining by using RASTA-filtering. Thus, a positive influence of

RASTA-filtering as seen in the SGI case cannot be obtained.

6.2.5 Intermediate Results

Before comparing the SGD results with VTLN, as acoustic normalisation technique, I present the SGI, SGDg, SGDa, and SGDag results in a summarised table (cf. Table 6.14) to directly compare the gained improvement across the corpora.

Table 6.14: Achieved UARs in percent for all corpora using SGD modelling. Additionally, absolute improvements against SGI results are given. The FSs are explained in Table 6.6.

Corpus	Problem	SGI	SGDg	SGDa	SGDag
LMC	two-class	68.7 (FS4)	73.4 (FS4)	72.0 (FS4)	73.8 (FS2)
improvement		–	4.7	3.3	5.1
emoDB	two-class	92.7 (FS2)	97.7 (FS2)	–	–
improvement		–	5.0		
emoDB	six-class	75.2 (FS2)	76.9 (FS2)	–	–
improvement		–	1.7		
VAM	two-class	71.8 (FS2)	78.4 (FS2)	76.4 (FS2)	76.3 (FS2)
improvement		–	6.6	4.5	4.5

Comparing the SGD results on the different corpora, it is evident that regarding speaker groups improves the recognition on all corpora. Here the smaller amount of data does not prevent this improvement. Even with emoDB, having very clear and expressive emotions, a quite high improvement is achieved. Utilising the complete quantity of emotions on the six-class problem shows the disadvantages of the SGD approach, as the small amount of training data drives the approach to its limits.

Comparing the utilization of age and gender characteristics on LMC and VAM reveals that the gender tends to be the dominating factor, as on both corpora, a higher improvement could be achieved, if the gender-information is used. Using the combination of both characteristics could slightly outperform the single characteristics' results on LMC. With VAM the SGDag result is behind both the SGDg and the SGDa result, but it is still better than the SGI result. It has to be further investigated whether this is caused by the small amount of training material for the male speakers or by an inadequate age grouping.

6.2.6 Comparison with Vocal Tract Length Normalisation

Another technique, dealing with speaker variabilities, is the utilization of VTLN. This method follows a conceptually different approach than speaker-group dependent modelling. Instead of separating the different speakers into certain groups, the acoustics of the different speakers are aligned (cf. Section 4.3.3). For this, a warping factor for each speaker, representing the degree of the necessary acoustical alignment, is estimated. It expresses the degree of frequency shift for the actual speaker's acoustics have to be changed to match a "general" speaker. This general speaker is modelled a-priori by all other speakers' acoustics of the data material.

To estimate the warping factor for each speaker, I used the maximum likelihood estimation (cf. [Kockmann et al. 2011]). Therein, a rather small GMM with only 40 mixtures and 4 iteration steps is trained on all unnormalised training utterances of the corpus using all speakers despite one. The left-out speaker is the target speaker, where the warped features are generated for each utterance in the range of 0.88 to 1.12 with a step size of 0.02.

The optimal warping factor is obtained afterwards by evaluating the likelihood of all warped instances against the unnormalised GMM and selecting the highest one (cf. [Cohen et al. 1995]). This procedure is repeated for all speakers. To estimate the warping factor, the same features as for the FS1 of the SGD approach are used, namely 12 MFCCs, C0, F_0, and E together with their Δ and $\Delta\Delta$ regression coefficients. Furthermore, CMS as channel normalisation technique is used. HTK is used for training, Vocal Tract Length Normalisation, and testing (cf. [Young et al. 2006]).

Estimated Warping Factors

Figure 6.18 depicts the estimated warping factors for emoDB and VAM. The acoustically high quality emoDB corpus has an equal number of male and female speakers covering an age range from 21 to 35 years. The estimated warping factors reflect the general description of warping VTLN: male speakers have a warping factor greater than 1 to stretch the vocal tract and female speaker's warping factor is smaller than 1 to compress the vocal tract. A performed k-means clustering reveals two clusters with centroids at 0.96 containing only female and 1.08 containing only male speakers. Thus, a good classification accuracy can be expected.

The estimated warping factors for VAM are quite different (cf. Figure 6.18). Although most of the male and female speakers have warping factors in the range of 0.96 to 1.04, which is close to 1, the general behaviour – compression for female and expansion for male speakers – is still apparent. The average warping factor for female

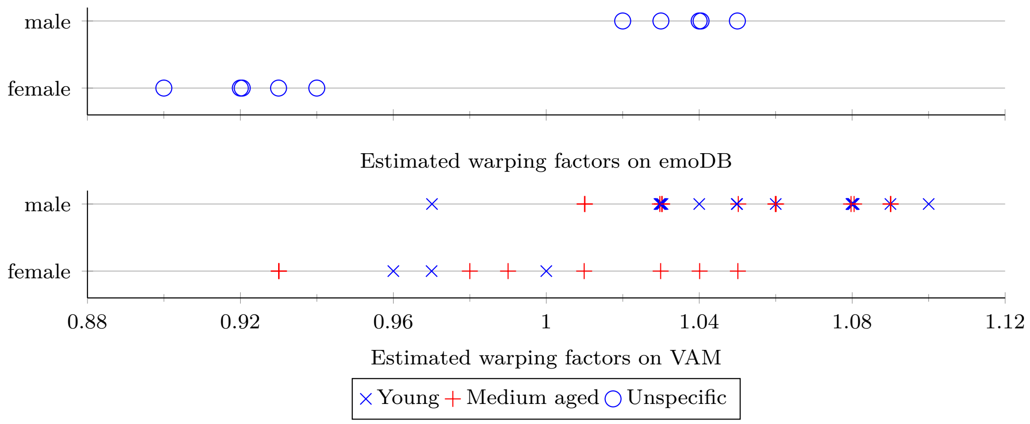


Figure 6.18: Estimated warping factors for every target speaker of the databases emoDB and VAM arranged for male and female speakers as proof of VTLN algorithm. The age groups are indicated where needed.

speakers is 0.99 with a standard deviation of 0.04 whereas for the male speakers an average warping factor of 1.05 with a standard deviation of 0.03 was estimated.

The age groups of the speakers were also known for this database. On the other hand, from Figure 6.18 it is apparent that age is not a separating factor. The mean for the young group is 1.05 with a standard deviation of 0.04. The old group has a mean warping factor of 1.03 and a standard deviation of 0.05. The anatomical differences are more prominent than the acoustic changes caused by the ageing.

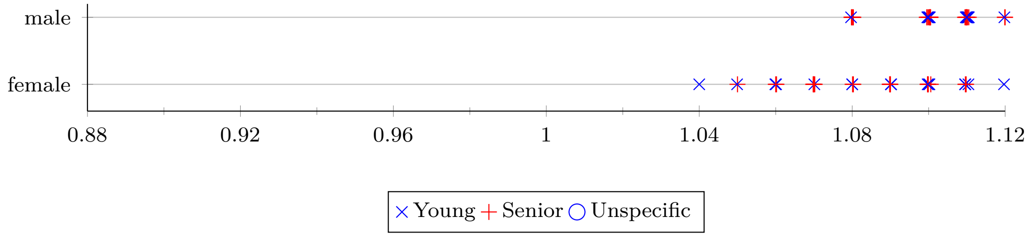


Figure 6.19: Estimated warping factors for every target speaker on LMC, as example for strange factor estimaton for male and female speakers. The age groups are indicated.

Estimating the warping factors on LMC reveals another picture (cf. Figure 6.19). Although the same procedure with identical features and modelling as for emoDB's and VAM's warping factor estimation are used, the achieved factors are quite different.

The general trend of the LMC's factors indicates a stretching. Female speakers have factors in the range of 1.04 to 1.12, whereas for the male speakers the factors are between 1.08 and 1.12. This can also be seen in the mean and standard deviation of both groups, the values of female speakers' factors have a mean of 1.08 and a standard deviation of 0.02. For male speakers the mean is 1.1 with a standard deviation of 0.01. Thus, by the estimated warping factors male and female speakers cannot longer be distinguished. Thus, although the warping factor estimation is able to separate male and female speakers on emoDB and VAM, this separation does not work for LMC. This may be due to the age groupings of LMC. The acoustic differences between the two groups, `y` and `s`, of LMC are more prominent, as for speakers over 60 years the fundamental frequency is dramatically changing. The female voice is declining and the male voice is increasing (cf. [Linville 2001; Hollien & Shipp 1972]. The influence of ageing on the fundamental frequency was discussed in Section 4.2.2. This can influence the warping factor estimation.

But, as for VAM the speakers' age is not a distinguishing factor, although both age groups in LMC are more apart than for VAM. Young adults have a warping factor with a mean of 1.1 and a standard deviation of 0.02, and seniors have a mean warping factor of 1.09 with a standard deviation of 0.02. Utilising other feature sets to infer RASTA-filtering and SDC-coefficients does not lead to different warping factors.

Classification Performance using VTLN

The obtained warping factors are then used to normalise the features of the different feature sets for each corpus accordingly. These features are then applied to train the classifiers using HTK and pursue a LOSO validation to preserve comparability to the previous presented SGD-experiments (cf. Section 6.2.3 and Section 6.2.4). In Figure 6.20 the achieved UAR of the VTLN approach is compared with results of the unnormalised features, which served as baseline classification results for the SGD approach as well. Furthermore, the best reported results by other researchers are marked with a dashed line for each corpus.

Investigating the achieved classification using VTLN, it can be stated that for all corpora the classification performance is improved in comparison to the baseline classification results but no significant improvement could be achieved (cf. Figure 6.20). Although the estimated warping factors on emoDB meet the expectations, the average classification performance is just slightly improved by about 0.5 % for the six-class problem and 0.7 % for the two-class problem. The highest improvement could be observed for FS1, where neither a RASTA-filtering is applied nor SDC-coefficients are inferred. All achieved improvements are not significant. Overall, the application of VTLN on emoDB could not outperform the results reported by other research groups.

The classification performance on both databases with naturalistic emotions benefits from the application of VTLN (cf. Figure 6.20), although the estimated warping factors

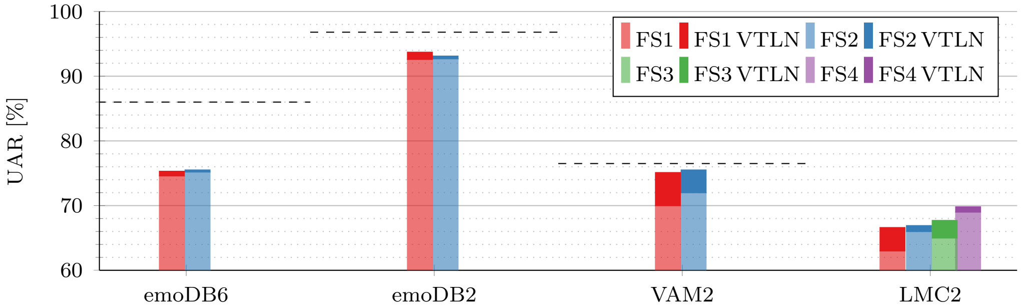


Figure 6.20: Mean UAR for VTLN-based classifiers in comparison to the baseline of different corpora utilising GMMs and LOSO on different feature sets. The number following the corpus abbreviation indicates the number of distinct classes. The best results reported for each database are marked with a dashed line (cf. Table 3.2 on page 39).

do not indicate that (cf. Figure 6.19 on page 155). The average improvement is 3.7 % for VAM and 2.0 % for LMC. An improvement of 5.1 % in total for VAM and of 3.6 % in total for LMC could be achieved using FS1. The highest improvement on VAM can be achieved with FS1 closely followed by FS2 and FS4.

With LMC the highest improvement is achieved using FS1 close followed by FS3. Unfortunately, the performance gains for FS1 and FS2 on VAM and for FS1 and FS3 on LMC are not significant as the standard deviation remains quite high. This suggests that VTLN cannot resolve speaker variabilities sufficiently. Furthermore, the gained improvements could not surpass the best reported results on both databases.

Comparison of SGD modelling and VTLN technique

To complete this study, I want to compare the results achieved by SGD modelling and VTLN. Therefore, I summarised the best results of all utilised corpora for both techniques in Table 6.15. I concentrated on feature sets with the particularly best result for each approach. As for most of the results, the standard deviation is already given, I do not note it in this overview table to preserve readability.

The comparison of SGD modelling and VTLN technique reveals that for all utilised corpora the improvement gained by distinguishing the different speaker groups achieve results that are in total 1.1 % to 4.3 % better than the VTLN results. Almost all corpora benefit from RASTA-filtering, which was not expectable especially for emoDB. For LMC, the feature sets achieving the best performance differ between SGI and VTLN on the one hand and SGD on the other hand. Furthermore, SGD improvements are significant, whereas VTLN does not produce significant improvements. It appears that the SDC-coefficients are able to adjust the acoustic variabilities.

Table 6.15: Achieved UARs in percent of SGD, VTLN, and SGI classification for all considered corpora. Furthermore, feature sets and speaker groupings are given. The FSs are explained in Table 6.6.

	emoDB		VAM	LMC
	six-class	two-class	two-class	two-class
SGI	75.2	92.8	71.8	68.7
FS	2	2	2	4
SGD	76.9	97.7	78.3	73.8
FS	2	2	2	2
Grouping	SGDg	SGDg	SGDg	SGDag
VTLN	75.5	93.4	75.5	69.8
FS	2	2	2	4

6.2.7 Discussion

In this section, I demonstrated that a speaker group dependent modelling leads to a significantly improved emotion classification. To do so, I first performed a parameter tuning of the two model parameters number of mixture components and iteration steps for GMMs. I was able to conclude that the best classification performance is obtained when choosing 120 mixture components. The type of emotion, simulated or naturalistic, does not influence the general trend of classification performance.

Simulated emotional databases are more sensitive for different iteration step numbers as naturalistic ones. As 4 appears as the best number of iteration steps in terms of classification and computation performance, I chose this for all further investigations. These findings confirm the results of [Böck et al. 2010], identifying a low number of iteration steps as suitable for both types of databases.

Furthermore, I investigated the influence of the contextual characteristics Δ, $\Delta\Delta$, and SDC-coefficients. Here I stated that the incorporation of Δ and $\Delta\Delta$ regression coefficients increases the recognition performance for both types of databases. In contrast to the utilization of SDC-coefficients for simulated databases where a decreased classification performance could be observed, the incorporating of SDC-coefficients increases the performance on LMC.

This is also partly true if different channel normalisation techniques are investigated. Here the application of CMS improves the classification performance on both types of databases. RASTA-filtering just slightly increases the performance on emoDB, but notably on LMC. My findings expand the investigation of [Kockmann et al. 2011], stating that RASTA-filtering together with ΔΔ regression coefficients has the most potency for acoustic emotion recognition.

Afterwards, I applied the speaker group modelling approach known from ASR to improve automatic emotion recognition from speech (Hypothesis 6.6 on page 136). As naturalistic recordings contain both more variability in the expression of emotions as well as less expressive emotions (cf. Section 3.3), I conclude that additional knowledge is required to successfully recognise emotions. Starting with the definition of speaker groups to reduce the acoustic variations, I assumed that separation of age and gender groups reduces the acoustic variability and thus, improve the emotion recognition.

Therefore, I performed experiments using the earlier defined model configuration and feature sets on LMC as naturalistic emotional corpus to support this hypothesis. This was successful: the incorporation of both age and gender for speaker grouping achieved significant improvements in comparison with SGI results and could outperform classifiers based only on age or gender as seperating characteristic.

The investigations presented on additional databases reveal that the emotion recognition can be improved through the separation of gender groups. The improvement is independent of the type of emotional content (simulated or naturalistic), the quality of recording, or the available speaker groups. I compared the SGD results with the previously gained baseline results (SGI) utilising the same features and classifiers as well as with results reported by other research groups. Hereby, I was able to demonstrate that this approach could be applied to several datasets. In comparison with the SGI classification, significant improvements for all databases using SGD-classifiers could be achieved, except for the six-class problem with emoDB.

Furthermore, for the first time this investigation allows to draw conclusions about the limitations of this approach. The lack of training material becomes most apparent. Regarding the classification performance in specific speaker groups, they outperform the SGI baseline in all cases. Especially, if the amount of available material for training is quite small, the performance is quite low, for instance the speaker groups of the six class problem on emoDB, the ym-group of LMC, or the m̳-group of VAM. Despite this, the trained SGD-model still outperforms a SGI-classifier trained on the same small amount of data (cf. [Siegert et al. 2013c]).

When comparing the classification performance of age-group dependent and gender-group dependent models a slightly better performance of the gender-group dependent models can be noticed. This indicates that the gender of a speaker has a higher influence on the variability of characteristics than his age, at least for the investigated age groups (cf. [Siegert et al. 2014d]). As this behaviour can be observed both in VAM and in LMC, this seems to be valid for different age groups, too. This result is supported by both the reported larger acoustic differences between male and female speakers as well as the gender effects on emotional responses as stated in Section 4.2. These findings are still not sufficient enough for general statements, but at least a tendency can be seen. On LMC a combination of both characteristics achieves the best performance. This may be supported by the optimal age grouping.

As further approach, I utilised VTLN, a method well known in ASR to compensate different vocal tract lengths. By applying this method, all SGI results could be improved. However, the improvement is neither significant nor does it achieve better results than the SGD-approach (cf. Table 6.15 on page 158). The comparison of the estimated warping factors reveals that a quite strange behaviour could be observed. Especially on LMC, the vocal tract of all speakers is stretched. The age-gender specific expression of emotions, as psychological research suggests, cannot be covered by VTLN (cf. Hypothesis 6.7 on page 136). One drawback this method has, is that the estimation of warping factors needs to be improved in order to cover the highly unbalanced age distribution. Thus, I advise using SGD modelling for emotion recognition.

6.3 Applying SGD-Modelling for Multimodal Fragmentary Data Fusion

As I have already stated in Section 3.2 and depicted in Section 3.3, naturalistic emotions require additional efforts in order to ensure a robust emotion recognition. Beside the improvement of emotion recognition methods on acoustic level I have presented earlier (cf. Section 6.2) that a multimodal emotion recognition approach can be utilised. This method is derived from the fact that humans express their emotions by using several channels, for instance facial expressions, gestures, and acoustics. Hence, the emotional response patterns are observable in different modalities which can be fused to robustly recognise the current emotion.

Although the main focus of this thesis is not classifier fusion, this topic is an important and arising issue for emotion recognition. In my case, I contributed in work which was done under the SFB/TRR 62 by colleagues at the Otto von Guericke University Magdeburg and the Ulm University (cf. [Böck et al. 2012a; Frommer et al. 2012b; Panning et al. 2012; Böck et al. 2013a; Krell et al. 2013; Siegert et al. 2013e]). A short introduction on common fusion techniques was given in Section 4.3.4.

In this section, I will concentrate on my contributions to improve the multimodal affect recognition for the naturalistic interaction on LMC. Therefore, contributions of other researchers are also presented and acknowledged accordingly. Afterwards, I discuss some of my contributions made under the constraint of fragmentary data.

6.3.1 Utilised Corpus

The conducted study, which will be presented afterwards, utilises the "79s" speaker set of the LMC (cf. Section 5.2.3) that has already been used in Section 6.2. Here, the focus is again on the two key events of the experiment, where the user should be set into a certain condition: `baseline` (`bsl`) and `challenge` (`cha`). The utilised material comprises the same automatically extracted and manually corrected 1 668 utterances as used in Section 6.2. The total length is 31 min, the average sample length per speaker group is nearly equal. But the distribution of samples is unbalanced, as a higher amount of samples is available for `bsl`.

The focus of this research is on the combination of different modalities. For visual classification only a subset of 13 speakers is available as the visual classifier was trained on manually FACS coded data, which is a time-consuming process. These 13 speakers are a subset of the 20s set, for which synchonised audio-visual data is available. The experimental codes of the 13 speakers, their amount of acoustic training material and their age grouping are given in Table 6.16.

Table 6.16: Detailed information for selected speakers of LMC.

ID	Speaker	`bsl`		`cha`		Group
		samples	length	samples	length	
1	20101013bkt	14	19.31	4	6.24	`yf`
2	20101115beh	26	28.30	15	13.36	`yf`
3	20101117auk	13	17.42	11	13.02	`yf`
4	20101117bmt	16	18.88	8	9.55	`ym`
5	20101213bsg	16	20.79	14	17.81	`sm`
6	20110110bhg	12	15.28	10	14.71	`ym`
7	20110112bkw	10	8.14	9	11.72	`ym`
8	20110117bsk	15	18.22	10	13.03	`ym`
9	20110119asr	15	18.73	2	3.45	`ym`
10	20110124bsa	10	10.88	4	4.04	`ym`
11	20110126bck	13	17.00	8	8.58	`ym`
12	20110131apz	16	12.21	3	2.42	`ym`
13	20110209bbh	15	12.61	5	5.63	`sm`

6.3.2 Fusion of Fragmentary Data without SGD Modelling

Using material of a naturalistic HCI, the classification applying different modalities, can be quite vague. One main reason is that the information in the different channels is not continuously available. This is mostly due to subjects not behaving ideally. They do not speak directly into the microphone, resulting in changing acoustic characteristics. They do not face the camera resulting in faces or the hands not always visible. Additionally, parts of the face can be hidden by hair or glasses. Furthermore, changes in illumination can make color-information unusable [Zeng et al. 2009; Gajšek et al. 2009; Navas et al. 2004]. We summarise these effects by denoting the data "fragmentary". Common reasons for fragmentary data are:

- prosodic features are only available if the user speaks,
- gestures are only detected if typical hand movements occur,
- facial expressions are usually temporary,
- user disappears from camera,
- face is hidden by hands,
- mouth speaking movement overlays facial expression.

This problem can, for instance, be observed in the LMC (cf. Section 5.2.3), too.

These fragmentary channel information can either be addressed by rejecting unfavourable data or by utilising a suitable fusion technique which is capable of handling such kind of data [Wagner et al. 2011]. Rejecting unusable data is not always feasible as it will reduce the over-all amount of data, leading in the end to nearly no remaining data. Furthermore, this approach is not applicable in real-time applications.

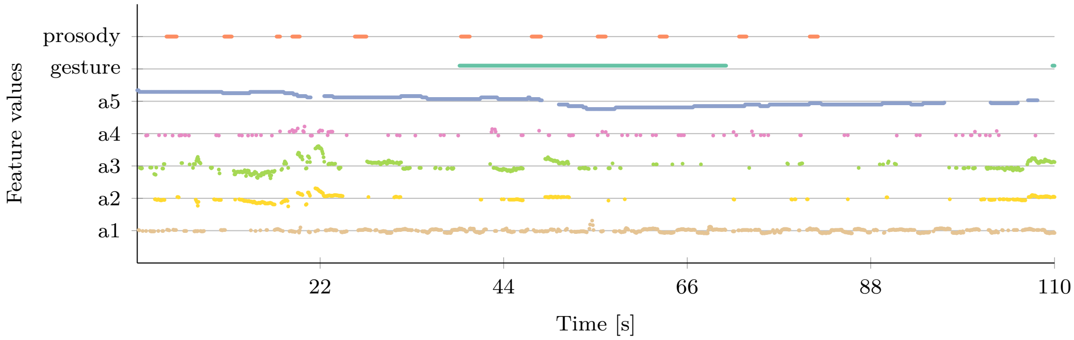


Figure 6.21: Observable features of the `challenge` event for subject 20101117auk of the LMC. Any dot in the figure represents a window with an extracted value for the specific feature. Facial measures: mean of left and right brow position (a1), mouth width (a2), mouth height (a3), head position (a4), eye blink frequency (a5). Gesture: line indicates self-touch. Prosody: line indicates utterance.

Figure 6.21 depicts the fragmentation of observable features for an excerpt of subject 20101117auk of the LMC. Within an excerpt of 110 s of the whole interaction, the subject speaks very rarely (∼13 s). Also, only two considered gestures can be observed, lasting for 32 s in total. The eye blink frequency (a5) and the brow position can be analysed nearly the entire time. Only longer closed eyes or fast head movements prevent a permanent observation. The three other facial features, mouth width (a2), mouth height (a3), and head position (a4), are also just partially observable. Detail on the extraction can be found in [Panning et al. 2012; Krell et al. 2013]. Thus, although acoustics, gestures and facial features are utilised, only for very few time points all characteristics can be evaluated. For instance, windowing all different features by using a 25 ms window, only in 880 out of 4400 frames all five video characteristics together with gesture information can be observed and for only 396 frames video and acoustic information are available. For the combination of all three characteristics, only 88 frames can be used.

Before discussing the applied decision fusion approach, I shortly present the unimodel classification results. The acoustic results are generated by myself, whereas the visual classifier is trained by Axel Panning. But, for the sake of completeness, I will briefly present his approach and results as well.

Acoustic Recognition Results For acoustic classification, I extracted a total of 39 features: 12 MFCCs as well as the C0 together with their Δ and $\Delta\Delta$ regression coefficients. These features are extracted frame-wise using a 25 ms Hamming window with a frame step of 15 ms. As classifier a GMM with 120 Gaussian mixture components and 4 iteration steps is trained (cf. Section 6.2.1). The whole “79s” speaker set of LMC is used for training. As the available material for facial analysis was limited to 13 speakers, only these speakers were analysed in a LOSO validation strategy. As confidence parameter, used later in the fusion approach, the log likelihoods for each test-utterance are stored as well. The unimodal classification results based on the UAR are given in Table 6.17. The overall mean UAR of the selected 13 speakers is 60.5 % with a standard deviation of 11.7 %. The achieved results are in line with similar investigations. In the previous section a mean UAR of 63.4 % with a standard deviation of 15.1 % could be achieved, also utilising F_0 (cf. Section 6.2.3). The results reported in [Prylipko et al. 2014a] achieved a UAR of 63 % using an SVM based on 81 turn level features comprising spectral acoustic features as well as voice quality features, and pitch related ones together with long-term statistical features (cf. Section 4.2.2). All these results show that a recognition of naturalistic emotions is quite difficult, as the acoustic variations are quite prominent and the emotional expressiveness is low (cf. [Batliner et al. 2000; Zeng et al. 2009]).

Facial Recognition Results The visual activities were analysed by Axel Panning, who considered mouth deformations, eyebrow movements, eye blink and global head movement as visual characteristics. Therefore, facial distances and head positions were measured (cf. [Panning et al. 2010; Panning et al. 2012]). As the emotional state is assumed to be reflected by the dynamics of observable features, and assumed to remain stable for a couple of frames (cf. [Panning et al. 2012]), a longer time window is used to analyse the visual activities. By a PCA the most important eigenvectors of the facial features are fed into an MLP to classify `bsl` and `cha`. The output of the MLP is a continuous value between 0 to 1, specifying the ratio that the actual feature context belongs either to the `bsl` (0) or the `cha` event (1). A general threshold of 0.5 is used to decide between `bsl` and `cha`. The overall mean of the facial classifications' UAR is 57.0 % with a standard deviation of 23.6 %. The individual unimodal classification results are given in Table 6.17. Although these results are close to the achieved acoustic performance, the rather high standard deviation depicts that an event decision can hardly be made on this modality alone. Emotion recognition from facial expressions on naturalistic interactions is hardly pursued, as the recording quality cannot be ensured and strict requirements onto illumination or gaze-direction to a camera device have to be fulfilled.

Table 6.17: Unimodal classification results (UAR) in percent for the 13 subjects.

ID	Speaker	Acoustic	Visual
1	20101013bkt	65.2	51.8
2	20101115beh	51.7	84.8
3	20101117auk	75.0	76.3
4	20101117bmt	57.1	65.0
5	20101213bsg	54.2	29.1
6	20110110bhg	60.0	55.3
7	20110112bkw	79.0	65.0
8	20110117bsk	66.7	91.4
9	20110119asr	68.0	25.6
10	20110124bsa	40.0	40.8
11	20110126bck	61.9	43.2
12	20110131apz	41.2	25.4
13	20110209bbh	66.7	88.3
	mean (std)	60.5 (11.7)	57.0(23.6)

Gesture Recognition Results The gesture detection, also performed by Axel Panning, is focussed on recognising self-touch actions. Self-touching was automatically detected in the video stream, by a skin colour detection algorithm. By a connected component analysis it is determined whether a self-touch in the face occurred (cf. [Saeed et al. 2011; Panning et al. 2012]). As, self-touch is a very rare event, the gestural analysis alone is not able to decide between `bsl` and `cha`. The absence of self-touch will give no evidences for one of these events. Thus, gestural analysis alone cannot be utilised for classification.

Decision Fusion For the fusion of the single modalities, an MFN is used to estimate the decision using all three modalities as input (cf. Section 4.3.4). Thus, the MFN mediates between the available decision of the unimodal classifications and additionally takes their temporal distances into account. In [Krell et al. 2013], the best performance could be achieved, when for each modality different input weights are used ($\mathbf{k_v} = 0.5$, $\mathbf{k_a} = 4$, $\mathbf{k_g} = 4$). The parameter $\mathbf{w}$, adjusting the lateral smoothness of the MFN, has only a slight influence on the performance. In the range of 50 to 1 000, the performance increases by just 4 % in total. The overall accuracy of this MFN is 79.8 % with a standard deviation of 21.2 %. The individual results are given in Table 6.18, together with improvements over the best unimodal channel.

Table 6.18: Multimodal classification results (ACC) in percent for the 13 subjects using an MFN. Using acoustic and visual classification results based on fragmentary data.

ID	Speaker	Fusion	Best single modality	
			Difference	Modality
1	20101013bkt	89.6	24.4	acoustic
2	20101115beh	95.7	10.9	visual
3	20101117auk	90.2	13.9	visual
4	20101117bmt	100.0	35.0	visual
5	20101213bsg	89.8	35.6	acoustic
6	20110110bhg	94.5	34.5	acoustic
7	20110112bkw	82.7	3.7	acoustic
8	20110117bsk	62.0	−29.4	visual
9	20110119asr	77.5	9.5	acoustic
10	20110124bsa	100.0	59.2	visual
11	20110126bck	71.2	9.3	acoustic
12	20110131apz	59.8	18.6	acoustic
13	20110209bbh	24.9	−63.4	visual
mean (std)		79.8 (21.2)	23.1 (16.5)	

The fusion results confirm the low expression level in the facial channel, which has already been presumed by FACS based unimodal classification. In comparison to the unimodal facial analysis, an absolute improvement of 22.8 % could be achieved. The prosodic analysis, which suffers from multiple missing decisions due to the silence of the speaker, shows a rather good framewise accuracy. An improvement of 19.3 % was achieved in comparison to the unimodal acoustic analysis.

The combination of both channels could lead to a quite high multimodal classification, as the poor results from facial expressions can be compensated by the good accuracy of the just sporadically appearing acoustic observations. Additionally, the mere occurrence of self-touch gestures can also be used, as the absence of a modality does not influence the MFN's decision. This fact can be supported by the structure of an MFN, providing the opportunity to utilise different weighting factors $\mathbf{k}$ for the various single modalities m.

6.3.3 Using SGD Modelling to Improve Fusion of Fragmentary Data

In this section, I present my investigations, combining the SGD approach and the MFN approach, to increase the affect recognition of fragmentary data. I already demonstrated that by inferring both gender and age information, the overall classification performance increased (cf. Section 6.2). By combining the improved acoustic classification with the decision fusion based on an MFN, I evaluated how an increased performance of an unimodal classifier influences the multimodal classification in total. For this, I raise the following hypothesis:

Hypothesis 6.8 *Although the acoustic channel is present quite rarely, an improved acoustic classification leads to an increased fused classification result.*

To do so, I utilised LMC (cf. Section 5.2.3) employing the same age-grouping as in Section 6.2 defining the SGDag classifiers, namely `yf`, `ym`, `sf`, `sm`. The same subset of LMC containing 79 speakers is used as well. The distribution of the training set is depicted for recapitulation in Table 6.19. The assignment of each subject to a speaker group is gathered from the corpus description.

Table 6.19: Distribution of utilised speaker groups in the "79s" set of LMC.

	male	female	$\sum$
young	16	21	37
old	18	24	42
$\sum$	34	45	79

Unimodal Classifiers The settings for the unimodal classifiers are the same as presented in Section 6.3.2. The visual classifier considers mouth deformations, eyeblink, eyebrow movement, and the general (global) head movement as well as the self-touch gesture information (cf. [Krell et al. 2013]). The visual analysis performed by Axel Panning is not modified. The individual results remain the same, as depicted in Table 6.17 on page 164. The overall UAR thus remains at 57.0 %.

The acoustic classifier now incorporates the age and gender information of each speaker. Apart from this, I utilised the same 39 features (12 MFCCs, C0 with Δ and ΔΔ) as well as the same classifier (GMM, 120 mixture components, and 4 iteration steps). Once again, a LOSO validation strategy is pursued. The individual results are presented in Figure 6.22.

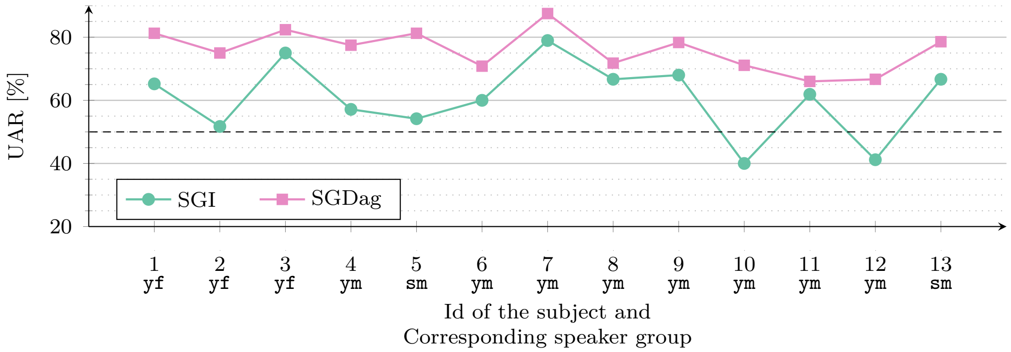


Figure 6.22: Comparison of the UARs achieved by the acoustic classification for each speaker with (SGDag) and without (SGI) incorporating the speaker group. The chance level is marked with a dashed horizontal line.

The overall UAR using the SGDag-classifiers for the incorporated 13 speakers increased to 76.0 % with a standard deviation of 6.5 %. Hence, an absolute improvement of 15.5 % could be achieved in comparison to the SGI classification, which is highly significant ($F = 17.4347$, $p = 0.0003$) (cf. [Siegert 2014]. This improvement outperforms the SGDag result presented in Section 6.2, where an improvement of 10.4 % could be achieved. The improvement is heavily influenced by the circumstance of an optimally selected subset. It mainly consist of young subjects. This age-group tends to express their emotions more clearly than senior adult speakers (cf. [Gross et al. 1997]) and thus, the separation of different age-groups results in an over-improvement. Additionally, the speakers selected in the “20s” set can be seen as quite expressive[30].

Fusion utilising the acoustic SGD Classifier The decision fusion system is based on the previously used MFN with the following values: the lateral smoothness

[30] The “20s” set was selected in such a way that suitable probands where asked to undergo the two other experiments. Thus, these subjects represent a best match selection of all subjects in terms of experimental expectations. As a result, the dialogue barriers caused a trouble in the communication and led to a recognizable event.

$\mathbf{w} = 1000$, the weighting factors $\mathbf{k_f} = 0.5$, $\mathbf{k_p} = 4$, and $\mathbf{k_g} = 4$. These values already demonstrated the best performance using the SGI acoustic classifier.

As pointed out above, each modality possesses an own distinct characteristic distribution of occurences over time. The recognition of the emotional state based on facial expressions requires the subject's face to be in the focus of the camera. However, in case the subject turns away and the feature extraction is hindered, decision making becomes infeasible. A similar problem can be observed regarding the prosodic analysis of the emotional state since it can be performed only in case the subject produces an utterance. In the given setting, the decisions derived from the gestural analysis are even more demanding because they only give evidence for the class `cha`. The classifier based on facial expression provides decision probabilities for all frames, while an acoustic analysis exists only for approx. 15.9 % of the frames and a gestural analysis only for about 9 % of the frames. Thus, it cannot be expected that the improvement for the final decision is as high as the improvement for the acoustic classification.

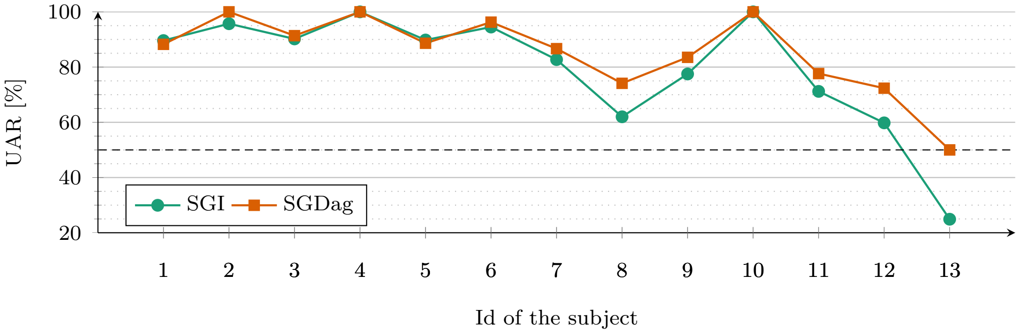


Figure 6.23: UARs after decision fusion over the continuous stream of the unimodal decisions comparing speaker group independent (SGI) and speaker group dependent (SGDag) acoustic classifier. The chance level is marked with a dashed horizontal line.

The individual results for each subject are presented in Figure 6.23. The over-all average accuracy is 85.29 % with a standard deviation of 14.22 %. In comparison to the SGI classifier fusion, the incorporation of speaker groups improved the performance by about 5.46 % in total. Although the acoustic classifier has a significant improvement, the current result is not significant ($F = 1.1186$, $p = 0.2902$) (cf. [Siegert 2014].

6.3.4 Discussion

As demonstrated in this section, the MFN is a powerful approach for fusing decisions from multiple modalities preserving temporal dependencies. The MFN reconstructs missing decisions and different temporal resolutions. Each channel can be individually weighted according to occurrences and reliability of the decisions. The unimodal recognition results provided by facial and acoustic analysis achieved moderate classification results that correspond with actual recognition results on naturalistic data material. The performed MFN fusion leads to a mean absolute improvement of 21.1 %.

Consequently, I combined my approach of SGD-modelling with the investigated MFN fusion to test my Hypothesis 6.8. An absolute improvement of 5.46 % could be achieved. The individual improvements are ranging from 0 % to 25.1 %. It can be further observed that an improved fusion could not be observed for all speakers where an acoustic improvement was gained. This could be attributed to the fact that acoustic observations occur very rarely (∼14.4 % `bsl`, ∼9.7 % `cha`). But in general, I was able to confirm my hypothesis. The marginal improvement of the fusion using the sophisticated acoustic classifier is related to the influence of the other two modalities competing with the acoustic modality as well as the sporadic utterances of the subject.

For the subjects 1 (20101013bkt) and 5 (20101213bsg) the acoustic classifier yields a notable performance gain, but the visual classification remains quite poor (cf. Figure 6.22 on page 167 and Table 6.16 on page 161). Furthermore, acoustic observations are quite rare for subject 1 and subject 5 with a total amount of approx. 15 s for (`bsl`) and 10 s for (`cha`) out of 110 s each. As the MFN generates a continuous prediction over the whole time, the recognition rates of all available modalities are incorporated. In cases when all modalities are available, the MFN can achieve quite good performances. Especially when the different inputs are weighted by ther reliableness, as in the present case, where the acoustic weight is 4 and the visual weight is just 0.5. But, if only one for channel is available, this channel determines the fusion result. In the investigated cases, the visual channel is nearly constantly available while the acoustic channel is just rarely available. Thus the MFN is tied by the low visual results.

Considering the young male subjects 8 (20110117bsk) and 12 (20110131apz), a remarkable improvement of the fusion can be observed. For these subjects the acoustic improvement yields to an improved fusion, as the acoustic channel is better represented. Especially within `cha`, a total amount of 18 s can be utilised. But for both subjects the acoustic classifications stay behind the mean of 76.0 % and thus, the fusion relying strongly on the acoustic's channel decision remains quite low.

The senior male subject 13 (20110209bbh) shows an overall low classification performance. Although, the individual classification results are quite high, the accuracy gained by the MFN is quite low (50.0 %), in comparison to the other subjects. In this case, a quite long self-touch event occurring during `bsl`, where a self-touch usually not occurs, prevents a better fusion result. This indicates that a unique thresholding for all speaker groups, for gestures and facial activities, could mislead the fusion.

6.4 Summary

In this chapter, I presented my own studies of the improvement of the automatic recognition of emotions from speech. Therein, I followed the established steps of pattern recognition. Starting with labelling issues, I presented a tool supporting the literal transcription and emotional annotation process. Furthermore, I presented studies using this tool on the emotional labelling of naturalistic emotions where I was able to support my hypothesis that emotional labelling methods derived from established methods in psychology results in samples with a proper emotional coverage. Following that, I investigated the inter-rater reliability measure to draw conclusions about the correctness of the labelling process. I was able to show to which extent a reliability is expectable for emotional labelling, and that the integration of visual cues as well as presenting the whole interaction in correct order helps to increase the IRR-value.

In the next step, I presented my improvements for the emotion classification itself. As the speech production literature shows that both the gender as well as the age have an influence on the speakers' acoustic characteristics, I hypothesised that separating the speakers' according to specific age groups and the speaker's gender could improve the emotion recognition. With this SGD approach I could significantly improve the emotion recognition on several databases with simulated and naturalistic emotional samples with various recording quality. I could show that my SGD-modelling approach has an effect for all kinds of different data. For highly expressive acted basic emotions and high quality data (emoDB) only a slight improvement of 1.2 % could be achieved. If the emotions are clustered dimensionally, the emotion-specific characteristics are increased and thus, the SGD approach earns a significant improvement for databases of both simulated and naturalistic emotions. I furthermore was able to show that this approach leads to better results than an acoustic normalisation through VTLN.

Finally, I combined my SGD approach with a multimodal fusion technique developed by colleagues at Ulm University to classify emotions within fragmentary multimodal data streams. In this, I showed that the improved acoustic classification by utilising the SGD approach leads to an improved fusion, although the acoustic information is just rarely present.

The basic requirement for all of these methods is, however, the presence of a measurable emotional reaction. It is known that HHI is controlled by further mechanisms, for instance feedback signals. Therefore, it is urging to regard also HCI under these premises. One of these feedback signals will be investigated in the following chapter.

Chapter 7

Discourse Particles as Interaction Patterns

Contents

THE previous chapters dealt with the automatic emotion recognition. I motivated emotion as an important information for a naturalistic HCI. In particular, this problem is addressed from an engineering perspective by formulating a pattern recognition problem, where the emotion is the pattern that has to be recognised.

In this chapter a new pattern is introduced that comprises information on the progress of HCI. For this, the so-called Discourse Particles (DPs) are used to evaluate the ongoing dialogue. These particles carry a very specific relation of their intonation with their function in the dialogie, which has not been used for automatic interaction evaluation so far. Therefore, I will present my studies on these patterns.

First, the importance of the investigated DPs as interaction patterns for HHI is depicted (cf. Section 7.1). In particular, their "form-function relation" is presented. This relation, introduced in [Schmidt 2001], is such that a specific meaning (function) is tied with different pitch-contours (form). This has been investigated by linguists for HHI, but has not yet been used for HCI.

To investigate these interaction patterns within an HCI, I verified that these particles are used within a naturalistic HCI and occur at specific situations of interest (cf. Section 7.2). In this context, I also incorporate age and gender dependencies. Afterwards, I investigate the assumed form-function relation by utilising two labelling tasks. This analysis is accompanied by an experiment to automatically distinguish the supported form-function relation (cf. Section 7.3).

Finally, I investigate the influence of further user characteristics such as personality traits on the use of DPs (cf. Section 7.4). It should be noted that the DP-usage is highly variable between different speakers, independent of the age- or gender-groups. This investigation is performed in conjunction with Matthias Haase, contributing psychological expertise in the evaluation of different personality traits.

7.1 Discourse Particles in Human Communication

During HHI several semantic and prosodic cues are exchanged between the interaction partners and used to signalise the progress of the dialogue [Allwood et al. 1992]. Especially the intonation of utterances transmits the communicative relation of the speakers and also their attitude towards the current dialogue. Furthermore, it is assumed that these short feedback signals are uttered in situations of a higher cognitive load [Corley & Stewart 2008] where a more articulated answer cannot be given.

As, for instance, stated in [Ladd 1996; Schmidt 2001], specific monosyllabic verbalisations, the Discourse Particles (DPs), have the same intonation as whole sentences and cover a similar functional concordance. These DPs like “hm” or “uhm” cannot be inflected but can be emphasised and are occurring at crucial communicative points. The DP “hm” is seen as a “neutral-consonant” whereas “uh” and “uhm” can be seen as “neutral-vocals” [Schmidt 2001][31]. The intonation of these particles is largely free of lexical and grammatical influences. Schmidt called that a “pure intonation”.

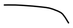
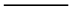
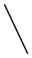
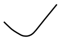
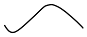
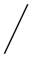
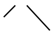
An empirical study of German presented in [Schmidt 2001] determined seven form-function relations of the DP “hm” due to auditory experiments (cf. Table 7.1)[32]. Several studies confirmed the form-function relation revealed. One investigation is presented in [Kehrein & Rabanus 2001]. The authors examined the data from four different conversational styles: talk-show, interview, theme-related talk, and informal discussion, with an overall length of 179 min taken from various German sources. They extracted 392 particles for the DP-type “hm” from the material and could confirm the form-function relation by a manual labelling. An investigation already carried out by Paschen in 1995 shows that the frequency of the different dialogical functions

[31] As the investigations are performed on a German corpus, I decided to rely on a perceptional translation: “ähm” is translated as “uhm” and “äh” as “uh” to be consistent with German sounds.

[32] In my thesis, I differentiate the term Discourse Particle from the two terms “filled pause” and “backchannel-signal”. The term “filled pauses” concerns only the particles which are used by the speaker to indicate uncertainty or to maintain control of the conversation. Thus, it does not comprise all functional meanings. The term “backchannel-signals” indicates all sorts of noises, gestures, expressions, or words used by a listener to indicate that he or she is paying attention to a speaker.

is depending on the conversation type [Paschen 1995]. By examining 2 913 kinds of “hm”s in eleven German conversations of different styles, the author concluded that confirmation signs dominate in conversations of narrative or cooperative character whereas in argumentative ones turn holding signals are more frequent.

Table 7.1: Form-function relation of the DP “hm” according to [Schmidt 2001]. Terms are translated into appropiate English ones.

Name	idealised pitch-contour	Description
DP-A		(negative) attention
DP-T		thinking
DP-F		finalisation signal
DP-C		confirmation
DP-P		positive assessment
DP-R		request to respond
DP-D		decline, can be seen as combination of DP-R and DP-F

A study using English conversations is presented in [Ward 2004]. The author investigated different acoustic features to discriminate different backchannel signals. As features the syllabification, duration, loudness, pitch slope, and pitch-contour of the acoustics is used. Ward could show that these features are appropriate to describe different feedback signals. Unfortunately, loudness cannot be reliably measured for realistic scenarios, since the distance between speaker and microphone varies (cf. Section 4.2). Additionally, syllabification can be split into single particles. Additionally, the pitch-contour is more exact than pitch slope. The features duration and pitch-contour are the same as in [Schmidt 2001]. In [Benus et al. 2007] the prosody of American English feedback cues is investigated and several DPs are annotated using eleven categories. Further information on the semantics of DPs within conversations can be found in [Allwood et al. 1992].

These particles are only very rarely investigated in the context of HCI (cf. Section 3.4.3). One of the very rare studies dealing with the occurrence of DPs during a HCI concluded that the number of partner-oriented signals decreases while the number of signals indicating a talk-organising, task-oriented, or expressive function is increasing [Fischer et al. 1996]. As the studies presented advise that the considered

DPs have a specific function within the conversation, this function could be helpful in assessing the interaction. But is is not analysed, whether the same mechanisms, such as backchanneling and cognitive load indication are expressed by humans within an HCI and, more importantly, can be detected. In the following, I will investigate the following hypotheses:

Hypothesis 7.1 *DPs are occurring more frequently at critical points within a naturalistic interaction.*

Hypothesis 7.2 *As the occurring DPs differ in their meaning, they can be automatically identified by their pitch-contour.*

7.2 The Occurrence of Discourse Particles in HCI

Now, I analyse whether DPs can be seen as interaction pattern occurring at interesting situations within an HCI (cf. Hypothesis 7.1). As representative corpus of a naturalistic HCI, I utilise the LMC. I start by using the whole session analysing global differences in the DP-usage. Afterwards, I analyse the local usage within significant situations and refer to the `challenge` barrier where caused by a suddenly arising luggage limit, the stress level of the user is rising as the luggage ahs to be re-packed. All investigations are performed on the "90s" set of LMC (cf. Section 5.2.3). The results are published in [Siegert et al. 2013a; Siegert et al. 2014a; Siegert et al. 2014c].

Based on the transcripts, all DPs are automatically aligned and extracted, utilising a manual correction phase. The preparation of the transcripts was conducted by Dmytro Prylipko, the manual correction by myself. I included the "hm"s as well as the "uh"s and "uhm"s as DPs. In this case, for each subject the whole session is used. In Figure 7.1, the total distribution of the three different DP-types is given.

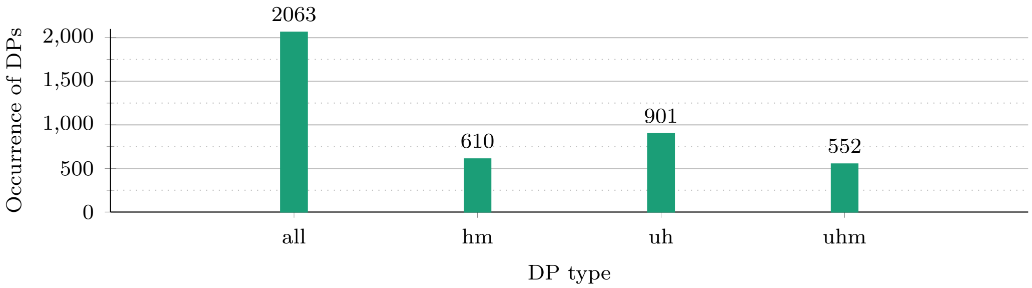


Figure 7.1: Number of extracted DPs distinguished into the three considered types.

The extraction results in a total number of 2 063 DPs, with a mean of 23.18 DPs per conversation and a standard deviation of 21.58. Only three subjects[33] do not utter any DP. One subject (20101206beg, `yf`) uses a maximum of 114 particles in the experiment. This result shows that DPs are used in HCI, although the conversational partner, the technical system, was not enabled to express them or react to them. The average DP length is 0.94 s, the standard deviation is 0.38 s. Only 32 min out of 40.4 h material represents DPs, illustrating the small amount of available data.

Before going more into details about the functional occurrence of DPs, I investigated the relation of the usage of DPs with the age and gender of the subjects. As I have shown in the previous chapter (cf. Section 6.2), the age and gender of a speaker influences the way emotions are uttered and therefore, these characteristics have to be included into the analysis. The group distribution of age and gender is as follows (cf. Table 7.2): 21 young male and 23 young female subjects and 19 senior male and 27 senior female subjects. For this investigation, I do not distinguish single DP-types.

Table 7.2: Distribution of utilised speaker groups in the "90s" set of LMC.

	male	female	$\sum$
young	21	23	44
senior	19	27	46
$\sum$	40	50	90

To provide valid statements on the DP-usage in a naturalistic HCI within the different SGD groups, two aspects have to be taken into account.

The first aspect that has to be taken into account is the verbosity of the speakers[34]. Verbosity denotes the number of verbalisations a speaker has made during the experiment. To model the verbosity and the DP-frequency, I assume that the underlying process is normally distribution. As the length of the experiment for each speaker was fixed and the time a speaker could spend for each category was pre-defined as well and furthermore, the speakers attending the experiment are all native Germans, the same general speaking rate can be assumed (cf. [Braun & Oba 2007]). Thus, on average, all speakers should have the same verbosity and the observed verbosity samples should vary around the unknown expected value. Unfortunately, this expected value is influenced by various factors, for instance age, gender, talking style or task difficulty. Thus several different populations have to be distinguished when grouping

[33] The subject codes are as follows, the age gender grouping is denoted as well: 20110208aib (`sm`), 20110315agw (`sm`), and 20110516bjs (`ym`).

[34] The verbosity analysis is based on the raw numbers provided by the Institute for Knowledge and Language Engineering at the Otto von Guericke Universität Magdeburg under the supervision of Prof. Rösner.

the observed samples. These factors have to be considered, as for the recruitment of participants, very opposing groups in terms of age, gender and educational level were considered (cf. [Prylipko et al. 2014a]). For this case and as the sample size is quite small, the observed samples do not perfectly reproduce a normal distribution. Nevertheless, I will make use of this model as it allows a comparison of the different influencing factors by utilising the ANOVA. To take into account that the samples do not form a normal distribution, the non-parametric Shapiro-Wilks version of the ANOVA (cf. Section 4.4.3) is used for the various statistical tests. The results of all calculations can be found in [Siegert 2014]. The test for normality distribution, taking into account several distinguishing factors, identifies for some cases a high significance that the data samples are normally distributed. The number of significant results for normality could be even increased, when the outliers below or above the quartiles are disregarded. The same considerations can be made for the frequency of DPs and the normalised DP-frequency (cf. Figure 7.3 on page 178 and Figure 7.4 on page 179).

The second aspect is the partition into two experimental phases with different dialogue styles. During the first phase, the personalisation, the subject gets familiar with communicating to a machine. The subjects are guided to talk freely. The second phase, the problem solving phase, has a more task focused command-like dialogue style. Of even more interest, is the combination of both aspects. As the dialogue styles are very different between the two phases, it could be assumed that also the verbosity differs. The verbosity for both experimental phases is depicted in Figure 7.2.

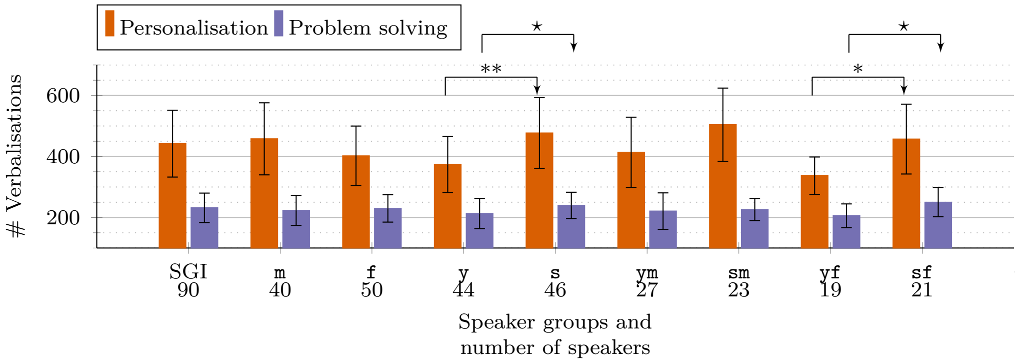


Figure 7.2: Mean and standard deviation for the verbosity regarding the two experimental phases and different speaker groups for LMC. For comparison the group independent frequency (SGI) is given, too. The stars denote the significance level: * ($p < 0.05$), ** ($p < 0.01$), ⋆ denotes close proximity to significance level.

Considering the verbosity of the two phases, the number of words between personalisation and problem solving phase differ significantly for each speaker group. The average number of words for the personalisation phase is 429. In contrast, for the problem solving phase the average number of words is just 226. Hereby, both phases are of nearly equal length. This can be attributed to the fact that the problem solving phase is more structured and thus less words are needed to fulfil the task of packing and unpacking clothes. Furthermore, with an average value of 93, the standard deviation for the problem solving phase is much lower than for the personalisation phase, where the averaged standard deviation is 210.

This also affects the previously mentioned age-related verbosity. Significant differences between the groups `y` and `s` ($F = 6.774$, $p = 0.009$) as well as between `yf` and `sf` ($F = 5.011$, $p = 0.025$) can be noticed for the personalisation phase. But these differences cannot be found for the problem solving phase, the p-values between `y` and `s` ($F = 3.566$, $p = 0.059$) as well as between `yf` and `sf` ($F = 3.457$, $p = 0.063$) are just close to significance level. A significant difference is hard to expect, as the number of observed samples is small, with just 19 `sm` and 27 `sf` speakers.

All speaker groupings have nearly similar verbosity values for the problem solving phase, ranging from 206 to 250. For the personalisation phase, the verbosity values differ between 337 and 504. The average verbosity increase factor between problem solving and personalisation phase is 1.89 for all speaker groupings, thus it can be stated that the subjects are more verbose during the personalisation phase. A DP should be more likely occur in the personalisation phase than in the problem solving phase, if it is just used for habituation.

7.2.1 Distribution of Discourse Particles for different Dialogue Styles

The aspects investigated above are now incorporated when the DP-usage is analysed. Thus, for each experiment the two phases, personalisation and problem solving, are distinguished. Furthermore, the user's verbosity is taken into account by using the relation of the users' DPs and their verbosity values. Additionally, I distinguish the different speaker groupings. The result is depicted in Figure 7.3.

First, it should be noted that for no speaker grouping significant differences between the DP-usage in both experimental phases can be observed. Furthermore, for the speaker groupings `f` and `sf`, the verbosity-normalised numbers of DPs between both phases are higher for the problem solving phase. This indicates that the DP-usage is not just an articulation-habituation occurring occasionally within the conversation with the system. Instead, the DPs can also be seen as an interaction pattern for HCI.

Considering the different speaker groupings, it can be stated that the difference in the personalisation phase is largely determined by the speaker's age. Young speakers

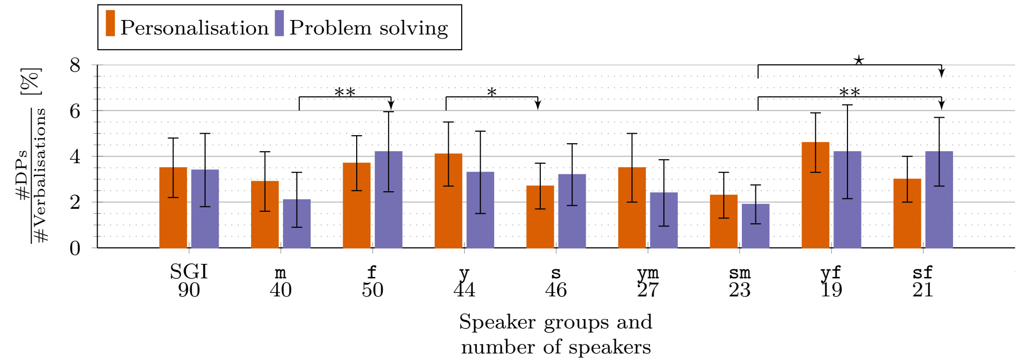


Figure 7.3: Mean and standard deviation for the DPs regarding different speaker groups for LMC. For comparison the group independent frequency (SGI) is given. The stars denote the significance level: * ($p < 0.05$), ** ($p < 0.01$), $\star$ denotes close proximity to significance level.

are significantly more verbose than senior speakers (($F = 5.195$, $p = 0.023$)) in the personalisation phase. As for the verbosity values, the difference between `yf` and `sf` ($F = 3.351$, $p = 0.067$) is close to significance level, although the verbosity is significantly less. This supports the assumption that younger users are used to talking to technical systems and therefore they tend to express themselves shorter and more concise, as stated in [Prylipko et al. 2014a], which I co-authored. But at the same time younger speakers are using DPs known from HHI more intuitively.

Regarding the problem solving phase, a significant difference can be observed along the gender dimension. Male speakers use significantly less DPs than female speakers ($F = 9115$, $p = 0.003$). A highly significant difference can also be observed between male and female senior adult speakers (`sm` and `sf`) with ($F = 8.111$, $p = 0.004$).

This results show that DPs are already used by humans (automatically) during the interaction with a technical system, although the present system is not able to react properly. Thus, the need to detect and interpret these signals is evident. For this, it is necessary to investigate the kind of dialogues causing an increased use of DPs.

7.2.2 Distribution of Discourse Particles for Dialogue Barriers

In the next step, I even go deeper into the problem solving phase and analyse whether the DPs show differences at the LMC's dialogue barriers. For this, I consider the dialogue barriers `baseline` and `challenge`, which already served as a basis for my SGD-dependend affect recognition in Section 6.2.

By way of reminder, I will shortly describe them. The `baseline` is the part of the experiment, where it is assumed that the first excitation is gone and the subject behaves naturally. The `challenge` barrier occurs, when the system refuses to pack further items, since the airline's weight limit is reached. Thus, the user has to unpack things. It is supposed that this barrier raises the subjects' stress level. I could support this statement, as already shown in Figure 6.7 on page 134. A main difference between `baseline` and `challenge` can be observed regarding the emotions `surprise`, `confusion`, and `relief`.

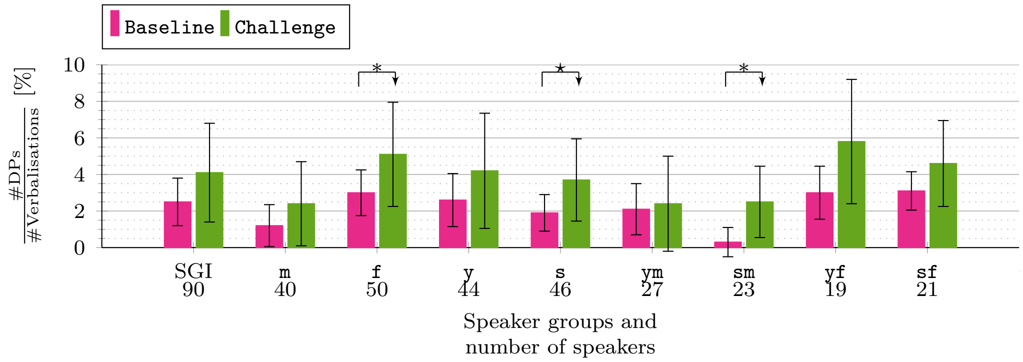


Figure 7.4: Mean and standard deviation for the DPs distinguishing the dialogue barriers `bsl` and `cha` regarding different speaker groups for LMC. For comparison the group independent frequency (SGI) is given. The stars denote the significance level: * ($p < 0.05$), ⋆ denotes close proximity to significance level.

From this description, I assume that due to a higher cognitive load due to the re-planning task in `challenge`, the DP-usage is also increasing, since DPs are known to indicate a high cognitive load (cf. [Corley & Stewart 2008]). For the analysis of this assumption, I calculated the relation of uttered DPs and verbosity within both dialogue barriers. Furthermore, I distinguished the previously used speaker grouping. The results are depicted in Figure 7.4 [35].

Regarding the DP-usage between the two dialogue barriers `baseline` and `challenge`, it is apparent that for all speaker groupings the average number of DPs for `challenge` is higher than for `baseline`. This is even significant for the speaker group `f` ($F = 4.622$, $p = 0.032$). The difference in the speaker group `s` is near the significance level ($F = 3.810$, $p = 0.051$). This observations support the statement from [Prylipko et al. 2014a] that male users and young users tend to have less problems to overcome the experimental barriers. Considering the combined age-gender grouping, only for

[35] The assumption of normality is heavily bent for this investigation (cf. [Siegert 2014]), also the assumption of independence can only be raised, if the influence of the situation is assumed to exceed the speakers' individual DP-uttering behaviour.

the sm grouping a significant difference between baseline and challenge can be observed ($F = 5.548$, $p = 0.018$). Thus, I can summarise that DPs are capable of serving as interaction pattern indicating situations where the user confronted with a critical situation in the dialogue (cf. Hypothesis 7.1 on page 174).

As one can see from Figure 7.4, particularly for the two groups yf and sf the standard deviation for challenge is quite high. This also indicates that other factors influence the individual DP-usage. I will analyse one kind of factors, the personality traits, which are in connection with stress-coping, in Section 7.4.

7.3 Experiments assessing the Form-Function-Relation

As presented in Section 7.2, DPs fulfil an important function within the conversation for both HHI and HCI. In the previous section, I could support my hypothesis (cf. Hypothesis 7.1 on page 174) that also in HCI the DPs are occurring at specific situations and thus it can be assumed that they fulfil the same functions for the dialogue, for instance indicating thinking, as in HHI.

In the following, I investigate my second hypothesis (cf. Hypothesis 7.2 on page 174) that the occurring DPs can be distinguished by their pitch-contour only. According to [Schmidt 2001], the pitch-contour allows to derive the function of the DP within the interaction. Schmidt called this the "form-function-relation". Thus, a reliable method for classifying the pitch-contour has to be developed. Then it is possible to evaluate the function of the occurring DPs. Furthermore, it has to be examined that the DPs are used as function indicators within the interaction and that this function is assessable by the acoustics as well as by the form-type.

For this investigation just a subset of 56 subjects of LMC with a total duration of 25 h could be used at this time. The manual correction of the transcripts, especially, the preparation of the DPs' alignment and manual labelling are quite time consuming. Furthermore, I considered the DP-type "hm", as only for this form type a well-founded theory exists (cf. [Schmidt 2001]). This results in a total number of 274 DPs.

To get an assessment for the functional use of the DPs, two manual labelling tasks were pursued. In the first task, the function of the DP within an HCI should be assessed based on the acoustic information. Afterwards, this label is cross-checked with the given graphical pitch-contour using the prototypes defined by Schmidt.

These labelling tasks followed the methodological improvements presented in Section 6.1. To this end, the labellers received the relevant parts of the interaction around each DP to be able to include contextual knowledge into their assessments. Furthermore, I utilised test-instructions and explanatory examples with replacement statements for the acoustic labelling (cf. Table 7.3).

7.3.1 Acoustical Labelling of the Dialogue Function

The acoustic labelling has been conducted with ten labellers. They had to assign the particles to one of the categories presented by Schmidt, see Table 7.1 on page 173. Additionally, I included the categories `other` (OTH) and `no hm` (DP-N). The category OTH should be assessed, when the categories by Schmidt were not suitable. For this, the labellers were instructed to give a suitable replacement statement as free-text. The category DP-N denoted cases where the automatic extraction failed, this means, the sample was not a "hm". To perform this labelling an adapted version of *ikannotate* (cf. [Böck et al. 2011b]) was used, where the emotional labelling module was exchanged by a module presenting a list of all DP function-label categories. The labellers also were given a manual describing the annotation and the several functional categories, which were paraphrased with suitable replacement statements (cf. Table 7.3).

Table 7.3: Replacement sentences for the acoustic form-type labelling, following the description given by [Schmidt 2001].

Label	Description	Replacement
DA-A	(negative) attention	shrug of the shoulders
DA-T	thinking	I need to get my head around; Wait a moment
DA-F	finalisation signal	Sighing; Oh!
DA-C	confirmation	Yes, Yes!
DA-D	decline	No, No!
DA-P	positive assessment	I see!
DA-AR	request to respond	What?

In the end, a majority voting was conducted to obtain the resulting function label. In this case, only those assessment were used, where five or more labellers agreed on the same label. The resulting labels are given in Table 7.4.

Table 7.4: Number and resulting label for all considered DPs. The categories are according to Table 7.1 on page 173, additionally used labels: DP-N and OTH. The label NONE indicates cases where no majority could be achieved.

Label	DP-A	DP-T	DP-F	DP-C	DP-P	DP-N	OTH	NONE
# Items	8	211	6	39	3	2	0	5

As a result, for 269 out of 274 DPs a majority vote could be achieved (cf. Table 7.4). Two samples were accordingly assessed not to represent a DP (DP-N). These samples had been wrongly included and contained short powerful hesitations. For five signals (NONE) no majority label could be found, the assessments for these particles varied between `attention` (DP-A) and `finalisation signal` (DP-F). Although, the labellers had the opportunity to assess meanings in addition to the ones given by Schmidt. Thus, it can be initially noticed that no other functional meaning than postulated by Schmidt are gained (OTH=0). Most of the DPs are used to indicate the task oriented function `thinking` (DP-T). Whereas partner-oriented signals as `attention` (DP-A), `finalisation signal` (DP-F), `confirmation` (DP-C), and `positive assessment` (DP-P) are just rarely used. Among these `confirmation` (DP-C) is used most frequently with approx. two-thirds of all partner-oriented signals. The DP-R (`request a respond`), commonly used in HHI, is not used. Presumably, the subjects do not expect the system to recognise this function properly.

Furthermore, I calculated Krippendorff's alpha to determine the reliability of the labelling process (cf. Section 4.1.3) and obtained a value of $\alpha_K = 0.55$. This indicates a moderate reliability according to Figure 4.7 on page 62 (cf. [Landis & Koch 1977]).

7.3.2 Form-type Extraction

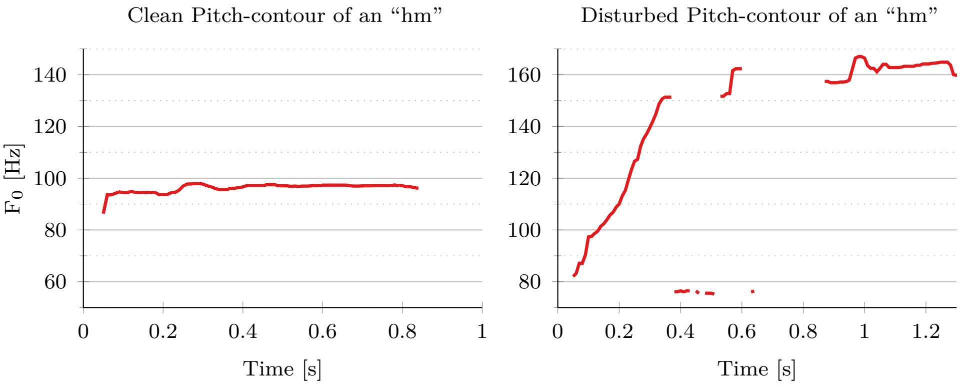


Figure 7.5: Samples of extracted pitch-contours. The left subfigure depicts a clean sample easily assignable to a form-prototype. The right subfigure depicts a disordered contour, where a form assignment is more difficult. The gaps indicate parts where no fundamental frequency could be extracted.

To extract the form of the DPs, namely the pitch-contour, I rely on commonly used methods, presented in detail in Section 4.2.2, the parameters of which are depicted in the following. The software PRAAT was used for this extraction (cf. [Boersma 2001]).

The DPs were windowed using a Hamming window of 30 ms width and a stepsize of 15 ms. A low-pass filter was applied with a band-pass frequency of 900 Hz and activated center-clipping. Autocorrelation was used to extract the pitch values for each frame. Having extracted the fundamental frequency for all windows of one utterance, a smoothing using a median filter utilising five values was applied to suppress outliers. In Figure 7.5 two extracted pitch-contours are depicted showing the difficulties of an automatic pitch extraction process for the DP “hm”. Due to the noise-like acoustic, the fundamental frequency cannot always be estimated for the whole expression.

7.3.3 Visual Labelling of the Form-type

The gathered labels of the acoustic labelling were compared to the extracted form-types. For this second labelling task, the extracted form-types were visually presented to the labellers together with the resulting label gained by the previous acoustic assessment. The prototypical form-types defined by Schmidt served as a reference to check whether the functional descriptions match the pitch-contours.

In this task, the labellers had to manually compare the extracted form-type with the defined prototypes and approve or reject the previous acoustically identified labels. I recruited five of the previous labellers. Thus, they were familiar with the task.

An inter-rater reliability of $\alpha_K = 0.7$ was achieved. Again, a majority voting was conducted, to obtain the resulting assessment. Since the samples labelled as `DP-N` are obviously no DPs, they were skipped for this task. The results are given in Figure 7.6.

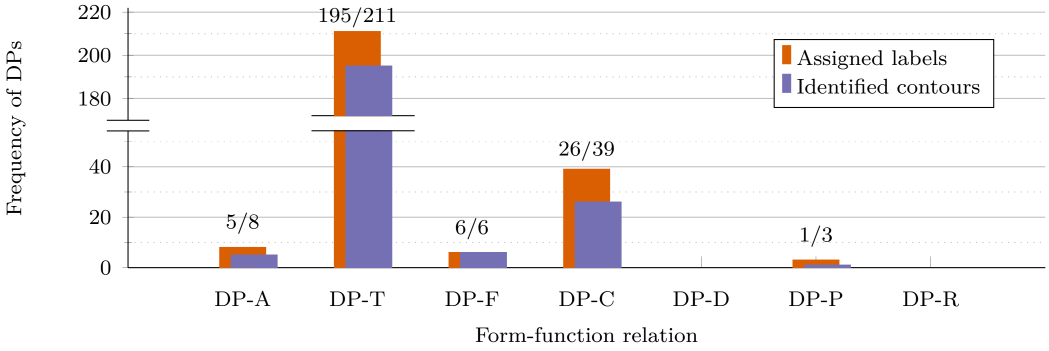


Figure 7.6: Comparison of the numbers of acoustically labelled functions with the visual presented form-types of the DP “hm”. The numbers above each bar indicate the number of matching functionals in relation to all samples for this functional.

It can be observed that for most (approx. 87 %) of the classified pitch-contours the form-type was labelled as matching, This also indicates the validity of the form-function relation defined by Schmidt for the considered naturalistic HCI. The functionals DP-D and DP-R could neither be found as pitch-contour nor were assessed. The non-occurrence of functional DP-D may be due to the experimental design itself, as no decline by the subjects is expected in this experiment. The lack of the functional DP-R indicates that the subjects do not fully assign human skills to the system. As a result from both labelling tasks, 195 particles could be successfully assigned to the functional DP-T, 26 to DP-C, 6 were identified as DP-F, 5 as DP-A, and 1 as DP-P.

The presented results are consistent with the findings of [Fischer et al. 1996], determining an increasing use of task-oriented signals. Furthermore, as no additional meanings were assessed, it can be assumed that the functionals determined by Schmidt are sufficient to distinguish the meaning of DPs within an HCI.

7.3.4 Automatic Classification

After the validity of the form-function relation has been approved, the next step is to investigate whether the form-function relation can be automatically utilised for a classification task. In order to allow a logically reproducible classification of the DPs, the set of labelled DPs is being divided into the two classes `thinking`, integrating all acoustically and visually correctly (DA-T) identified 195 samples, and `non-thinking`, representing the cumulated other 38 form functions of the DPs. From this, the distribution of DPs is quite unbalanced.

To perform a classification of the DPs of type “hm” into `thinking` and `non-thinking`, I performed several experiments utilising the pitch-contour as the only characteristic, consisting of the fundamental frequency (F_0) enhanced with different temporal context information (Δ, $\Delta\Delta$, $\Delta\Delta\Delta$, and SDC), as introduced in Section 4.2.2. Th utilised Feature Sets (FSs) are given in Table 7.5.

Table 7.5: Utilised FSs for the automatic form-function classification.

	Feature	Context Information	Size
FS1	F_0	Δ	2
FS2	F_0	Δ, $\Delta\Delta$	3
FS3	F_0	Δ, $\Delta\Delta$, $\Delta\Delta\Delta$	4
FS4	F_0	SDC	8
FS5	F_0	Δ, $\Delta\Delta$, $\Delta\Delta\Delta$, SDC	11

A three-state HMM is utilised as a classifier (cf. Section 4.3.1). The number of Gaussian mixture components is varied in the range of 12 to 39. The number of iterations was fixed to 5, according to the experiments presented in [Böck et al. 2010].

For validation, a ten-fold cross-validation is used. The results of these experiments can be found in Figure 7.7.

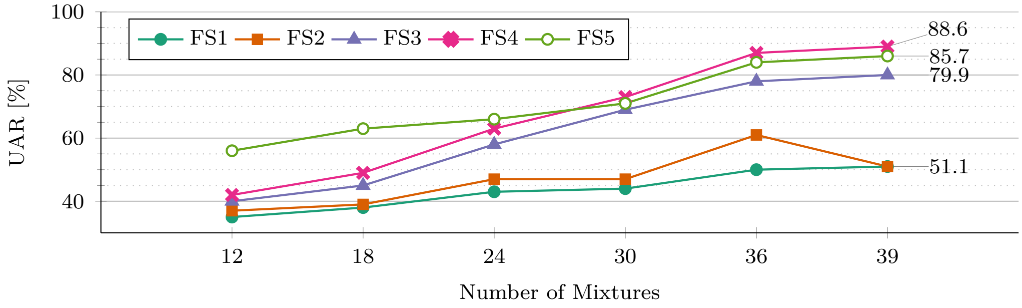


Figure 7.7: UARs of the implemented automatic DP form-function recognition based on the pitch-contour.

Considering the results presented, it can be observed that the classifier is able to learn the two classes by just utilising the pitch-contour. Using FS1 and FS2 (Δ and ΔΔ coefficients), the UAR is quite low with just 52.1 %. Using ΔΔΔ or SDC-coefficients, the accuracies increased remarkable. The best UAR is achieved utilising just the SDC-coefficients (FS4) with 89 % UAR (cf. Figure 7.7). Thus, it can be assumed that for an acceptable performance, additional context knowledge is necessary. According to the presented experiments, Δ and ΔΔ coefficients are not able to represent the necessary temporal resolution. This can only be ensured by the ΔΔΔ and SDC coefficients. The ΔΔΔ cover a width of seven windows and 120 ms time span. The SDC coefficients comprise ten windows with a time span of 165 ms.

Table 7.6: Example confusion matrix for one fold of the recognition experiment for FS4.

	Predicted Class	
True Class	`thinking`	`non-thinking`
`thinking`	14	6
`non-thinking`	2	2

As I have shown in Table 7.1 on page 173, prototypical pitch-contours characterise different functional meanings. So, a longer temporal context allows a better modelling of the pitch-contour and thus leads to an improved accuracy. Considering the misclassification rate for both classes, the class `thinking` is just slightly misclassified (up to 30 %) while the class `non-thinking` is misclassified more often (up to 50 %). This is shown exemplary in Table 7.6. This misclassification arises from both the highly unbalanced training set with just small sample sizes (38 `non-thinking` vs. 195

`thinking`) for the class `non-thinking` and the artificial combination of the different remaining particle functionals due to the small amount of present samples.

7.4 Discourse Particles and Personality Traits

In Section 7.2 I investigated the usage of DPs considering the subjects' age and gender. I showed that DPs are not equally distributed among the investigated age and gender groups (cf. Figure 7.3 on page 178). For some speaker groups, the usage of DPs is higher in the problem solving phase than in the personalisation phase. From these investigations, it can also be seen that the standard deviation is quite high. This indicates a high individuality of the users' DP-usage. Thus, it can be assumed that additional criteria, as specific psychological characteristics, are influencing the usage of DPs. As already mentioned earlier, the usage of DPs can be connected to the users' stress coping ability, therefore, I further analysed the DP-usage depending on these specific personality traits. This investigation is pursued together with Matthias Haase, from the Department of Psychosomatic Medicine and Psychotherapy at the Otto von Guericke University. For this analysis, I again set the DPs in relation to the total number of the user's acoustic utterances.

To analyse whether a specific personality trait influences the DP-usage, we differentiated between users with traits below the median (low trait) and those at or above the median (high trait). As a statistical test, an one-way non-parametric ANOVA is used to compare the means of our two median-split samples (cf. Section 4.4.3). We tested all personality traits available for the LMC, but I will report only those factors which allow for analysis close to the significance level. These factors are determined by the following personality questionnaires:

- NEO Five-Factor Inventory (NEO-FFI) [Costa & McCrae 1995]
- Inventory of interpersonal problems (IIP-C) [Horowitz et al. 2000]
- Stressverarbeitungsfragebogen (stress-coping quest., SVF) [Jahnke et al. 2002]

Considering each psychological trait, no significant differences are noticeable on the distinction between the two dialogue styles "personalisation" and "problem solving". But, the difference between personalisation and problem solving can be almost statistically analysed regarding the speakers above the median (cf. Table 7.7).

One could have expected that the discriminating factors (personality trait above median and trait below median) have an effect on the use of DPs in the two phases. However, as can be seen in Table 7.7 no significant differences in the usage of DPs can be seen between both groups whithin the personalisation phase. For the problem solving phase the difference is very close to significance level. Furthermore, the influence

Table 7.7: Archieved level of significance of DP-usage between personalisation and problem solving phase regarding personality traits.

	Personalisation		Problem solving	
trait	F	p	F	p
SVF positive distraction strategies (`SVF pos`)	2.015	0.156	3.546	0.058
SVF negative strategies (`SVF neg`)	1.271	0.260	3.515	0.061
IIP vindictive competing (`IIP vind`)	2.315	0.128	3.735	0.053
NEO-FFI Agreeableness (`NEO agree`)	1.777	0.183	3.479	0.062

of psychological characteristics heavily depends on the situation in which the user is located. The personalisation phase was not intended to cause mental stress, moreover it should make the user familiar with the system. Hence it can be assumed, that this was not a situation raising the users mental stress and thus this personality trait will have no influence. In the regulated problem solving phase very different situations are induced by the experimental design, which also produces partly contradictory user reactions, which can not be covered, when the whole pase is considered. Thus, as only few users were compared within a very heterogeneous sample, for statements with statistical significance the number of samples is not sufficient (cf. Figure 7.8).

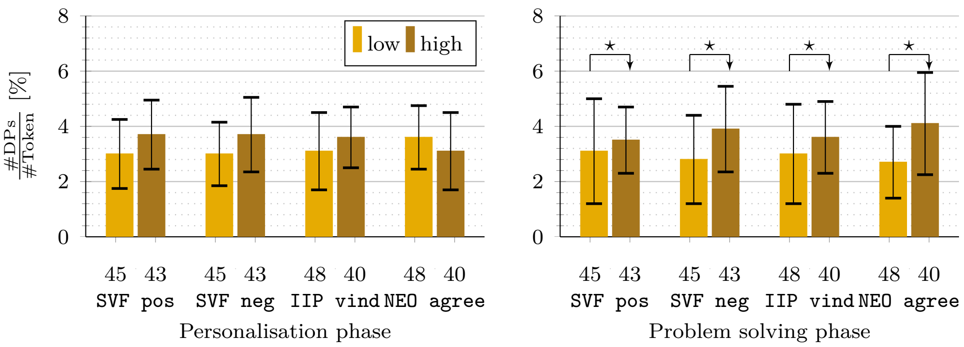


Figure 7.8: Mean and standard deviation for the DPs divided into the two dialogue styles regarding different groups of user characteristics. The symbol ⋆ denotes close proximity to significance level.

In addition to the analysis based on the two experimental phases, which are already published in [Siegert et al. 2014a], I also investigated the different usage of DPs between the dialogue barriers `baseline` and `challenge` regarding the personality trait factors `SVF pos`, `SVF neg`, `IIP vind`, and `NEO agree` (cf. Figure 7.9). In this case, the `SVF neg` factor shows significant results to distinct between the low and high group for both `baseline` ($F = 6.340$, $p = 0.012$) and `challenge` ($F = 4.617$, $p = 0.032$). Whereas for `SVF pos`, `IIP vind` and `NEO agree`, I can state that users belonging to the high group show an increased DP usage during the challenge barrier that is close to the significance level.

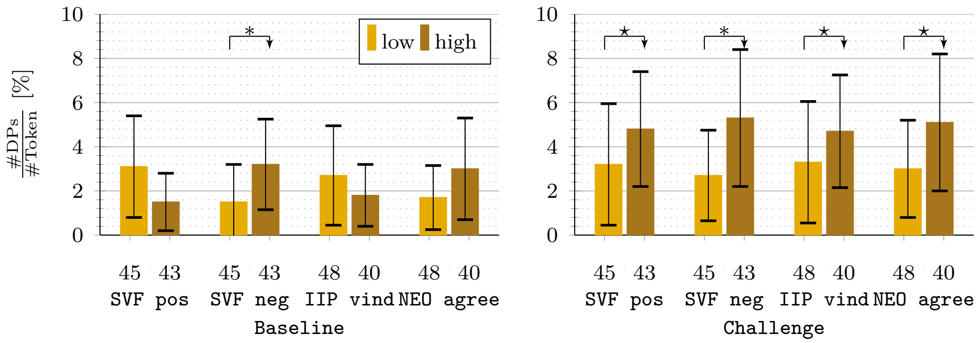


Figure 7.9: Mean and standard deviation for the DPs of the two barriers regarding different groups of user characteristics. The stars denote the significance level: * ($p < 0.05$), ⋆ denotes close proximity to significance level.

From both studies, the following conclusions can be drawn. Analysing the SVF positive distraction strategies (SVF pos), I can state that subjects having better skills in stress management with regard to positive distraction use substantially less DPs. The finding on SVF negative strategies (SVF neg) confirms this statement. Subjects not having a good stress management or even having negative stress management mechanisms use more DPs.

Evaluating the IIP vindictive competing (IIP vind) personality trait, it can be seen that subjects whi use DPs more frequently are also more likely to have problems in trusting others or are suspicious and rather quarrelsome against others. The interpretation of the NEO-FFI traits also confirms the IIP-findings. Subjects using fewer DPs show less confidence in dealing with other people with is determined by the agreeableness (NEO Agree).

Thus, it can be assumed that "negative" psychological characteristics stimulate the usage of DPs. A person having a bad stress regulation capability will be more likely to use DPs in a situation of higher cognitive load than a person having good stress regulation capabilities. This supports the assumption that DPs are an important pattern to detect situations of higher cognitive load (cf. [Corley & Stewart 2008]).

7.5 Summary

In this chapter, I introduced a new pattern, namely Discourse Particles, which I conjectured to be important for the evaluation of naturalistic HCI. Starting from the description of DPs and their role within HHI, I raised the hypothesis that they are also occurring in important situations within a HCI. To support this hypothesis, I analysed the LMC and incorporated the users' verbosity as well as two experimental modules "personalisation" and "problem solving" of the corpus.

My analyses reveal that DPs occur frequently within a HCI, although the system was not able to properly react to them. The occurrences are influenced by the user characteristics age and gender. In particular, the differences in the problem solving module are largely determined by the user's gender. These analyses support my Hypothesis 7.1 on page 174 that DPs occurr more frequently at critical points within the interaction. Furthermore, by comparing the occurrence of DPs between the interactions in undisturbed `baseline` and after the `challenge` dialogue barrier, I showed that DPs are good indicators for problematic interactions (cf. Section 7.2).

Afterwards, I performed several annotation experiments to support the form-function relation raised by Schmidt (cf. Section 7.3). Using the conducted acoustic function labelling and visual form-type labelling, I confirmed the form-function relation for the DP "hm". The consequently performed recognition experiments just used the pitch contour, characterised by the temporal contextual information of F_0, and achieved a UAR of 89 % distinguishing the class `thinking` from the combined class `non-thinking`. This exemplarily shows that DPs are indeed employable for the detection of situations of, for instance, higher cognitive load within an interaction and significantly contribute to the understanding of human behaviour in HCI systems. Furthermore, this supports my Hypothesis 7.2 on page 174 that the form-function relation can be recognised by using the pitch-contour.

I also investigated whether specific personality traits influence the DP-usage (cf. Section 7.4). This consideration was triggered by the fact that the high standard deviation could not be fully explained by the different age and gender groupings of the speakers. The investigations reveal that the usage of DPs is to a certain degree influenced by the users' stress coping ability.

The investigations presented in this chapter reveal that the occurrences of DPs can provide hints to specific situations of the interaction. My investigations show that not just the mere occurrence of the DPs is essential, but also their meaning. This meaning can be automatically recognised by their pitch-contour. Furthermore, I showed that DPs are occurring more frequently in situations of a higher cognitive load and thus, are an important interaction pattern. For the automatic usage of this phenomenon described in this chapter, obviously further steps, e.g. automatic DP allocation, are necessary. An account of this is given in Chapter 9.

Chapter 8

Modelling the Emotional Development

Contents

THE previous chapters dealt with finding correct and reliable labels and also presented a speaker group dependent modelling approach (cf. Chapter 6) and demonstrated the use of interaction patterns (cf. Chapter 7). This enabled us to build emotion-aware computers that recognise emotions and further interactive signs. But to develop "affective computers", more is necessary, as pointed out in [Picard 1997]. The observations of emotions and interaction signals alone are not sufficient to understand or predict human behaviour and intelligence. In addition to emotional observations and interaction patterns, a description of the progress of an interaction is also necessary.

As stated before, emotions are short-term affects usually bound to a specific event. Within HCI they are important affective states and should be recognised. But the recognised emotions should not lay the foundation for more in-depth decision on the dialogue strategy of the technical system. As emotions are only direct reactions to current occurring events, they are not related to the ongoing interaction and furthermore are also unable to give indications on the perceived interaction progress. Furthermore, a longer lasting affect, the mood, has to be used to deduce the interaction progress.

The mood, as discussed in Section 2.3, specifies the actual feeling of the user and is influenced by the user's emotional experiences. As an important fact for HCI, moods influence the user's cognitive functions, behaviour and judgements, and the individual (creative) problem solving ability (cf. [Morris 1989; Nolen-Hoeksema et al. 2009]). Thus, knowledge about the user's mood could support the technical system to decide whether additional assistance is necessary, for instance.

In the current chapter, I present a first approach, enabling technical systems to model the user's mood by using emotional observations as input values. I am starting with a motivation to introduce my considerations that lead to the mood model presented. Afterwards, the developed mood model itself and its implementation are described (Section 8.2). Here I also use personality traits to adjust emotional observations. The subsequent section presents the experimental evaluation (cf. Section 8.3). All results are published in [Siegert et al. 2012a; Siegert et al. 2013b; Kotzyba et al. 2012].

8.1 Motivation

Modelling the user's emotional development during an interaction with a system could be a first step towards a representation of the user's inner mental state. This provides the possibility to evolve a user interaction model and gives the opportunity to predict the continuous development of the interaction from the system's perspective. As stated in Section 2.3, moods reflect medium-term affects, generally not related to a concrete event (cf. [Morris 1989]). They last longer and are more stable than emotions and influence the user's cognitive functions directly. Furthermore, the mood of a user can be influenced by emotional experiences. These affective reactions can also be measured by a technical system. In this case, the mood can be technically seen as a long-time integration over the occurring emotional events to damp their strength.

In Section 2.3, I stated that according to [Mehrabian 1996], the mood can be located within the PAD-space (cf. Table 8.1). As further impact, moods are object to certain situational fluctuations caused by emotional experiences. Thus, the mood can be seen as a quite inert object within the PAD-space which can be influenced by emotional observations. Furthermore, certain personality traits such as `extraversion` could also influence the mood (cf. [Morris 1989]). As already stated in Section 2.3, for instance, [Tamir 2009] claims that extraverted persons regulate their affects more efficiently and show a slower decrease of positive affect.

Table 8.1: Mood terms for the PAD-space according to [Mehrabian 1996].

PAD-octant	mood	PAD-octant	mood
+ + +	Exuberant	− − −	Bored
+ + −	Dependent	− − +	Disdainful
+ − +	Relaxed	− + −	Anxious
+ − −	Docile	− + +	Hostile

Starting from the observation of the mood as a quite inert object within the PAD-space, being influenced by emotions, I define the following behaviour, which I intend to model:

- mood transitions in the PAD-space are caused by emotions
- single emotional observations do not directly change the mood's position
- repeated similar emotional observations facilitate a mood transition into the direction of the emotional observation
- repeated similar emotional observations hinder a mood transition in the opposite direction
- the personality trait `extraversion` can be seen as a reinforcement suppression factor on the emotional observation

From these observations I formulate the following hypothesis:

Hypothesis 8.1 *The mood can be modelled by a spring model, where emotions are considered as forces onto the mood object with a dimension specific self-adjustable damping term.*

Hypothesis 8.2 *The impact of an observed emotion on the mood is dependent of the personality trait* ***`extraversion`****, which can be modelled directly with one additional adjustment factor.*

8.2 Mood Model Implementation

The approach presented models the user's mood as an indicator of his inner mental state during an interaction with a technical system. However, it is not known how a user's mood can be deduced directly without utilising labelling methods based on questionnaires, for instance SAM or PANAS (cf. Section 4.1.2). Hence, for my approach the mood will be derived implicitly from observed emotional reactions of the user. Hence, the modelled mood can only be regarded as an approximation. Furthermore, the observation of single short term affects, the emotions, as well as the modelled mood will be located within the PAD-space in the range of -1 to 1 for each dimension.

This abstract definition of the mood's location by using the PAD-space allows the model to be independent of the chosen observed modality and to have the same representation for the emotional input values. For my approach, I use acoustically and visually labelled emotions as input values, which are also located in the PAD-space. This allows to get quite reliable emotional labels to validate the mood modelling without using a still imperfect automatic emotion recognition.

The implementation of the mood model can be described as follows: the emotional observations are influencing the actually felt mood by performing a force on the mood, which leads to an according mood translation within the PAD-space. My modelling is used to represent the users' mood during the interaction and is not limited to the `valence` dimension and refers to [Becker-Asano 2008]. An illustration of my mood modelling approach is given in Figure 8.1: An observed emotion causes a force onto the mood, as already stated, both, emotion and mood are placed within the PAD-space.

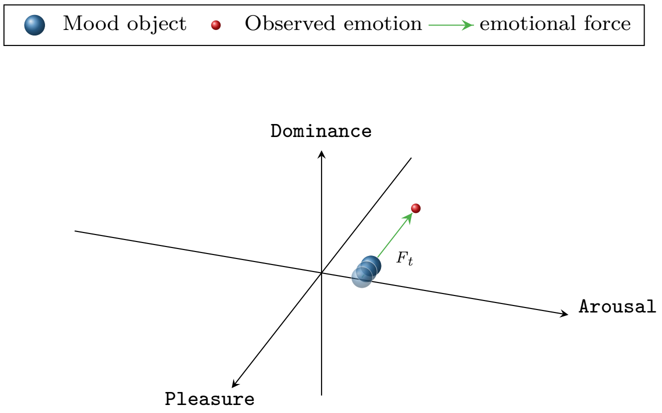


Figure 8.1: Illustration of the temporal evolution of the mood. The mood object is shifted by an observed emotion.

In my model, the mood is neither based on an emotional construction by a computational model (cf. [Gebhard 2005]) nor limited to the `valence` dimension (cf. [Becker-Asano 2008]). The approach presented by Gebhard implements the OCC model of emotions (cf. [Ortony et al. 1990]) outputting several co-existing emotions, where the computed emotions are afterwards mapped into the PAD-space. The mood is derived afterwards by using a mood change function in dependence of the computed emotion centre of all active emotions and their averaged intensity. The direction of the mood change is defined by the vector pointing from the PAD-space centre to the computed emotion centre. The strength of change is defined by the averaged intensity. Additionally, the authors utilise a time-span, defining the amount of time the mood change function needs to move a current mood from one mood octant centre to another. The mood simulation presented by Becker-Asano also relies on precomputed emotional objects located in the PAD-space. In contrast to [Gebhard 2005], the `valence` dimension is used to change the mood value. Thus, this model does not locate the mood within the PAD-space. Furthermore, in this model a computed emotional `valence` value results in a pulled mood adjusted by a factor indicating the "temperament" of an agent. A spring is then used to simulate the reset force to decrease steadily until neutrality is reached (cf. [Becker-Asano 2008]). Both mood models are used to equip virtual humans with realistic moods to produce a more human-like behaviour.

8.2.1 Modelling the Mood as three-dimensional Object with adjustable Damping

To illustrate the impact of recognised emotions on the mood, I modelled the observed emotion e_t at time t as the force F_t with the global weighting factor κ_0 (cf. Eq. 8.1). Furthermore, the emotions e_t are modelled for each dimension in the PAD-space separately. The calculation of the mood is conducted component-wise. The force F_t is used to update the mood M for that dimension by calculating a mood shift ΔL_M (cf. Eq. 8.2) utilising the damping D_t which is updated by using the current emotion force F_t and the previous damping D_{t-1}. This modelling technique is loosely based on a mechanical spring model: The emotional observation performs a force on the mood. This force is attenuated by a damping term, which is modified after each pulling.

$$\begin{aligned} F_t &= \kappa_0 \cdot e_t && (8.1)\\ \Delta L_M &= \frac{F_t}{D_t} && (8.2)\\ M_t &= M_{t-1} + \Delta L_M && (8.3)\\ D_t &= f(F_t, D_{t-1}, \mu_1, \mu_2) && (8.4) \end{aligned}$$

The main aspect of my model is the modifiable damping D_t. It is calculated according to Eq. 8.5 and Eq. 8.6. The damping is changed in each step by calculating ΔD_t, which is influenced by the observed emotion force F_t. The underlying function has the behaviour of a tanh-function, with the two parameters μ_1 and μ_2. The parameter μ_1 changes the oscillation behaviour of the function and the parameter μ_2 adjusts the range of values towards the maximum damping.

$$\begin{aligned} D_t &= D_{t-1} - \Delta D_t && (8.5)\\ \Delta D_t &= \mu_2 \cdot \tanh(F_t \cdot \mu_1) && (8.6) \end{aligned}$$

Al already stated emotions and moods are represented in PAD-space, the mood model should be considered within this space as well. For this, the mood calculation is carried out independently for each dimension and the result is formed from the combination of the single dimensional values of the mood model. The calculation over all three components is denoted as follows:

$$e_t^{pad} = \begin{pmatrix} e_t^p \\ e_t^a \\ e_t^d \end{pmatrix},\ F_t^{pad} = \begin{pmatrix} F_t^p \\ F_t^a \\ F_t^d \end{pmatrix},\ D_t^{pad} = \begin{pmatrix} D_t^p \\ D_t^a \\ D_t^d \end{pmatrix} \text{ and } M_t^{pad} = \begin{pmatrix} M_t^p \\ M_t^a \\ M_t^d \end{pmatrix} \qquad (8.7)$$

The block scheme for the mood modelling is illustrated in Figure 8.2.

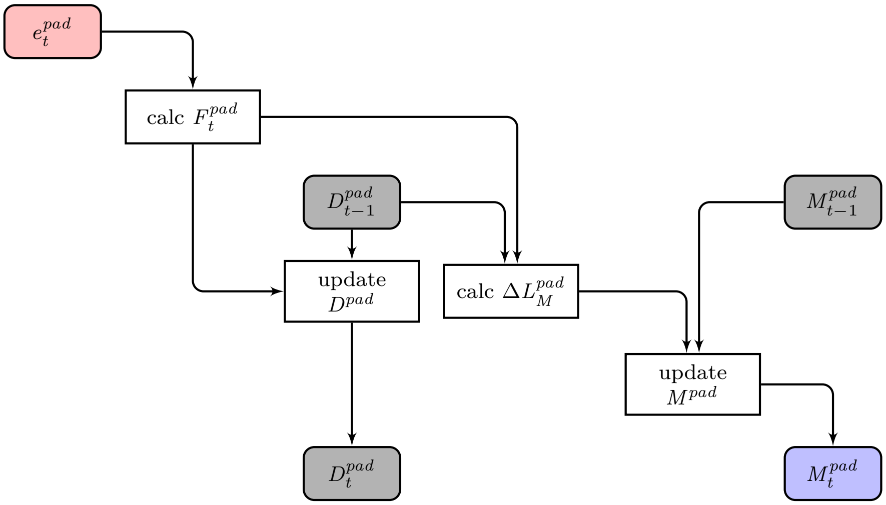


Figure 8.2: Block scheme of the presented mood model. The red box is an observed emotion, the blue box represents the modelled mood, grey boxes are inner model values, and white boxes are calculations. For simplification the combined components are used, internally the calculation is done for each dimension separately.

The mood model consist of two calculation paths. The diagonal one calculates the actual mood M_t^{pad}. The vertical one, updates the inner model parameter D_t^{pad}. As prerequisite the emotional force F_t^{pad} is compiled from the observed emotion e_t^{pad}.

8.2.2 Including Personality Traits

In Section 8.1, I explained that a user's mood is the result of the influence of the user's emotions over time with respect to the previous mood. The impact of the user's emotions on his mood depends also on the user's personality (cf. Section 2.3). The observed (external) affect needs not be felt (internally) with the same strength. For this, I considered to investigate how a mood model can translate the observed emotion into an emotional force with respect to the known differences in the external and internal representation.

It is known from literature that an external and an internal assessment lead to different interpretations (cf. [Truong et al. 2008]). Hence, these traits must be considered in the development of the mood model. Congruent with [Larsen & Fredrickson 1999; Carpenter et al. 2013], it can be noted that observed emotions, although similar in intensity and category, may be experienced differently by different users. Furthermore, depending on the user's personality, the way emotions are presented may vary. Hence, a translation of the observed emotion into the internal representation of the user is needed. For this, I focussed on the emotional intensity as an adjustment factor to determine the difference between external observation and internal feeling of the user's emotion. For this case, I use the personality trait `extraversion` that influences the adjustment factor (κ_η) (cf. Figure 8.3) where η indicates positive or negative.

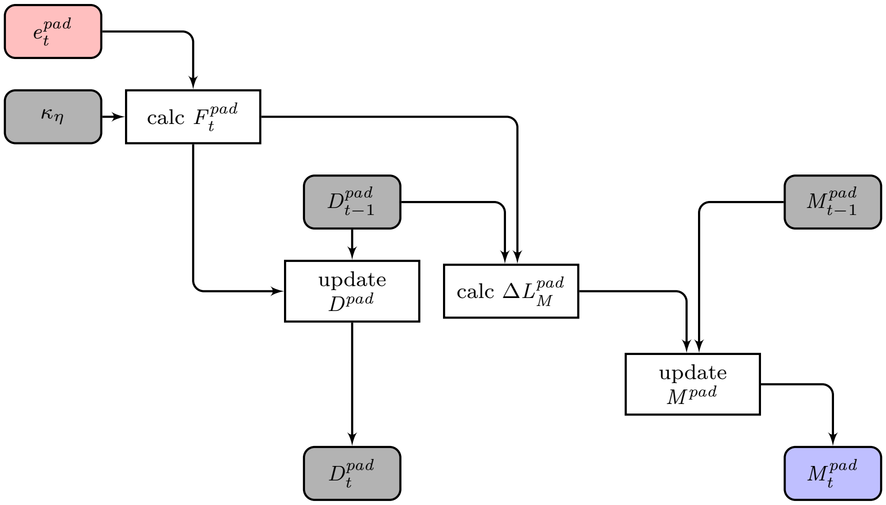


Figure 8.3: Block scheme of the mood model, including κ_η to include a personality trait dependent emotion force. For simplification the combined components are used, internally the calculation is done for each dimension separately.

The model scheme is still equivalent to Figure 8.2, except that an adjustment factor κ_η, determined from the user's `extraversion` value, is used for the calculation of F_t^{pad} (cf. Figure 8.3). The personality trait `extraversion` is particularly useful to divide subjects into the groups of users "showing" emotions and users "hiding" emotions (cf. [Larsen & Ketelaar 1991]). Additionally, users with `high extraversion` are more stable on positive affects. These considerations lead to a sign-dependent factor to distinguish between positive and negative values for emotional dimensions. This factor is used to weight positive and negative values according to the individual `extraversion` value of the actual user. For participants with `high extraversion` (≥ 0.6), the relation $\kappa_{\text{pos}} \geq \kappa_{\text{neg}}$ is used, for introverted participants (`low extraversion`, < 0.4) the relation $\kappa_{\text{pos}} < \kappa_{\text{neg}}$ is modelled. As for users with a `medium extraversion`, represented by values between 0.4 and 0.6, the effect of a higher stability for positive affect is not as salient (cf. [Larsen & Ketelaar 1991]), κ_{pos} and κ_{neg} are not distinguished.

$$\kappa_\eta \quad = \left(\begin{smallmatrix} \kappa_{\text{pos}} \\ \kappa_{\text{neg}} \end{smallmatrix} \right) \tag{8.8}$$

Since the κ_η depend on stable personality traits, they are (subject-) individually fixed and thus are constants of the model. Their value is determined from questionnaires, which will be discussed in Section 8.3.2.

8.3 Experimental Model Evaluation

The presented modelling technique needs sequences of emotion values to allow a mood prediction. Since this type of data is hardly obtainable, I chose two different databases already containing emotional sequences or having a quite strict experimental design for the emotional course. From these corpora, I used the emotional annotation as input data. The first experiment, utilising the SAL database, is an evaluation and plausibility test applying a quasi-continuous stream of emotional observations to model the user's mood within the PAD-space. The second experiment, based on the EmoRec corpus, tests if the modelled mood corresponds with the experimental presettings.

8.3.1 Plausibility Test

The database used for the mood model evaluation is the SAL database generated to build Sensitive Artificial Listener agents (cf. [McKeown et al. 2010]). This corpus consist of audio-visual recordings. Several communicatively competent agents with different emotional personalities are used to induce emotional user reactions. The annotation of this database was already discussed in Section 6.1. For my investigations, I concentrated on the transcriptions and annotations provided with the corpus. The data was labelled on five core dimensions, overall emotional intensity, `valence`, `activation`, `power`, and `expectation`, using GTrace (cf. Section 4.1.2) by three to six labellers each. To evaluate my proposed mood model, I chose the second session of speaker 1 (female), since annotations by five labellers for the both utilised traces, `pleasure` and `arousal`, are available. These labels are produced with a constant track having a step width of 0.02 s, which can be seen as quasi-continuous. The resulting labels fit the requirements of my mood modelling technique, as emotion labels with constant time-steps are needed. The reliability using Krippendorff's alpha (α_K) is 0.12 for `arousal` and 0.11 for `pleasure` (cf. Section 6.1.3). Both cases are interpreted as a *slight* reliability (cf. Figure 4.7 on page 62).

I have used the two dimensions `pleasure` and `arousal` for the mood modelling. To be able to use these labels, I calculated the mean of the five available annotation traces per dimension over all five involved annotators. Afterwards, each emotional label from both dimensions is used value-wise for the model. After performing some pre-tests to ensure that the mood remains within the limits of $[-1, 1]$ of the PAD-space, I defined the initial model parameters as given in Table 8.2.

Table 8.2: Initial values for mood model.

M_0	D_0	μ_1	μ_2
5	5	0.1	0.1

The results for the mood-evolvement over time are depicted in Figure 8.4, separately for the `pleasure` and `arousal` dimension. This figure depicts very clearly the delayed integration ability of the mood model. High emotion input amplitudes are delayed and dampened, the resulting mood is reacting with a quite high delay and just, if the amplitude remains within the same state for some time. The delay is dependent on the number and strength of emotional observations, and the damping term, as well as the inner-model parameters μ_1 and μ_2 (cf. Eq. 8.9):

$$\Delta L_M = \frac{F_t}{D_{t-1} - \mu_2 \tanh(F_t \cdot \mu_1)} \tag{8.9}$$

Furthermore, short changes of the amplitude do not change the mood course, as it can be seen for example around the time of 265 s in the `pleasure` dimension. The mood course calculated for both dimension separately is depicted in Figure 8.4.

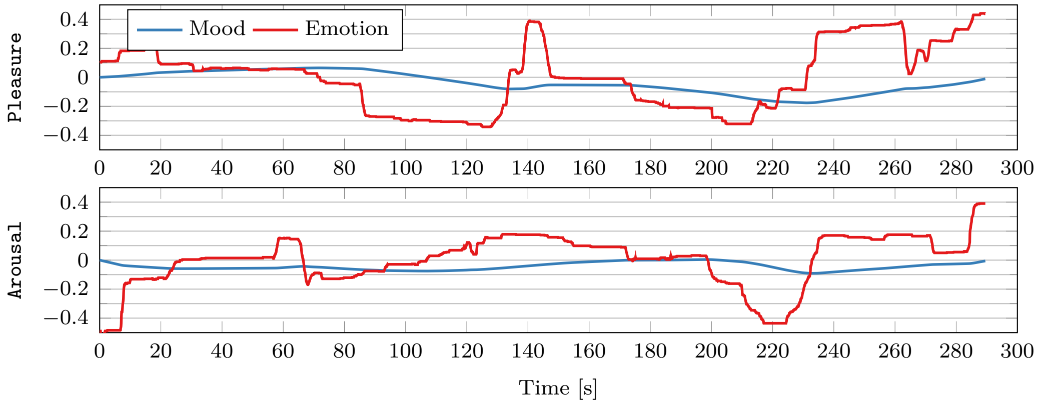


Figure 8.4: Mood development over time for separated dimensions using one sample speaker of the SAL corpus.

8.3.2 Test of Comparison with experimental Guidelines

For the second test, I rely on EmoRec-Woz I, a subset of the EmoRec corpus (cf. [Walter et al. 2011]). This database was generated within the SFB/TRR 62 during a Wizard-of-Oz experiment and contains audio, video, and bio-physiological data. The users had to play games of concentration (Memory) and each experiment was divided into two rounds with several Experimental sequences (ESs) (cf. Table 8.3).

Table 8.3: Sequence of ES and expected PAD-positions.

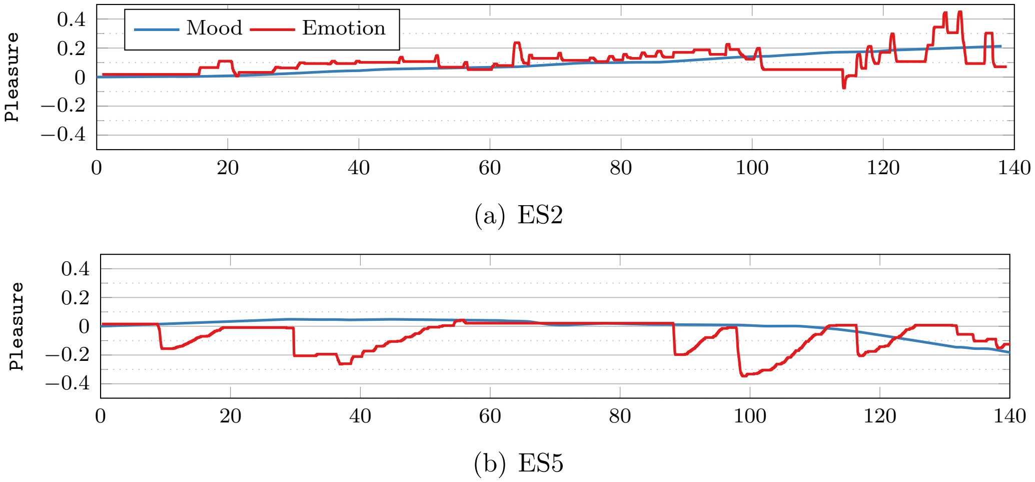
ES	Intro	1	2	3	4	5	6
user's PAD location	all	+ + +	+ − +	+ − +	− + −	− + −	+ − +
pleasure development	–	↗	↗	↗	→	↓	↑

The experiment was designed in such a way that different emotional states were induced through motivating feedback, wizard responses, and game difficulty level. The ESs with their expected PAD octants are shown in Table 8.3. The octands are indicated by their extreme-points. To model the mood development, I limited the investigation to `pleasure`, as this dimension is already an object of investigation (cf. [Walter et al. 2011; Böck et al. 2012b; Kotzyba et al. 2012; Böck et al. 2013a]), which simplifies the comparison of my mood modelling technique with other results.

The model calculation is based on the experiment of one participant (experimental code: 513). This session has a length of about 17.6 min. As initial model parameters, I used the same values as above (cf. Table 8.2). To gather the emotion data I again rely on labelled data, as it is still difficult to achieve a reliable emotion recognition over time for data of naturalistic interactions (cf. Section 3.3, [Böck et al. 2012b]). I used GTrace (cf. Section 4.1.2) and employed five labellers to label the two ESs on the `pleasure` dimension for the labelling of the EmoRec data.

As ES2 and ES5 were the most interesting sequences of the experiment, I first concentrated on these ESs. For this case, I relied on my results of Section 6.1 and utilised the complete ESs with both audio and video material. The average course is then achieved by calculating the mean over all single traces. The reliability of Krippendorff's alpha $\alpha_K = 0.10$ is comparable to reliabilities achieved on SAL. The modelled mood based on `pleasure` traces for both ES's are separately depicted in Figure 8.5.

Both ESs differ in the type of induced emotions. In ES2 the system tries to support and engage the user, while in ES5 negative feelings are induced. To do so, short positive or negative triggers are given by the wizard. These triggers can either be direct positive or negative feedback or given indirectly by memory card decks varied on difficult level. The user's reaction to these triggers are seen as changes of the



(a) ES2

(b) ES5

Figure 8.5: Gathered average labels on the dimension `pleasure` for ES2 and ES5 of participant 513.

annotated emotional trace (cf. Figure 8.5). Furthermore, the modelled mood within both ESs is in line with the experimental guidelines, as the modelled mood in the end of ES2 is 0.2128, which represents a positive mood, and at the end of ES5 the mood is negative with a value of −0.1987.

When including personality traits into the mood model, the following must be given: 1) The personality of the participants and 2) their subjective feelings. The first prerequisite is fulfilled by EmoRec-Woz I, as the Big Five personality traits for each participant were captured with the NEO-FFI questionnaire (cf. [Costa & McCrae 1985]). To integrate personality traits in the mood model I expand the adjustment factor κ_η (cf. Figure 8.6).

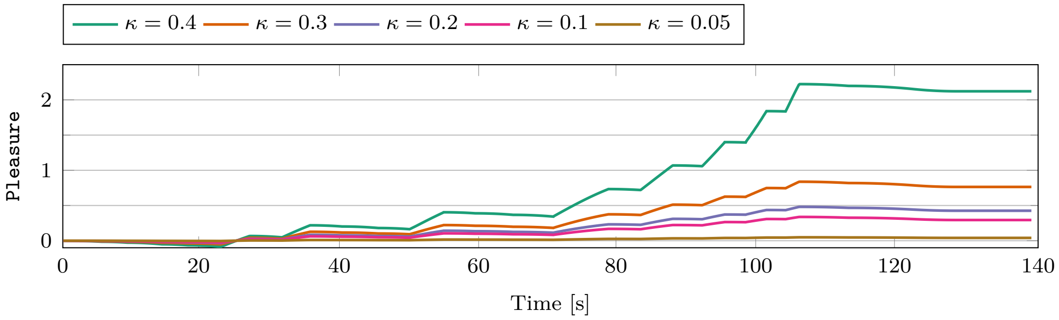


Figure 8.6: Mood development for different settings of κ_η, but not differing between κ_{pos} and κ_{neg}. As data the pleasure dimension for ES2, based on the labelled experimental data of participant 513, is used.

For this case, κ_η can be modified by choosing different values for κ. I will depict results for values in the range of 0.05 to 0.4. These values reproduce the strength of how an observed emotion is experienced by the observed person itself. An example of the different values for κ_η is given in Figure 8.6. For this experiment, emotional traces based on labelled observations for ES2 are used. It can be seen that values higher than 0.3 led to the mood rising too fast. This causes implausible moods, since the upper boundary of 1 for the PAD-space is violated. Hence, a $\kappa_\eta > 0.3$ should be avoided.

In contrast, for very small values ($\kappa_\eta < 0.1$) the mood becomes insensitive to emotional changes. Therefore, I suggest using values in the range from 0.1 to 0.3, as they seem to provide comprehensible mood courses. In Figure 8.7 it is depicted how the difference of κ_{pos} and κ_{neg} influences the mood development.

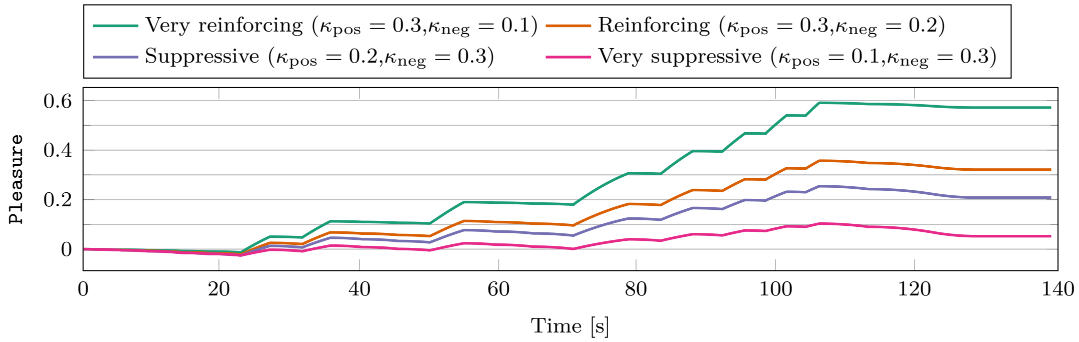


Figure 8.7: Mood development for different settings of κ_{pos} and κ_{neg}. As data the pleasure dimension for ES2, based on the labelled experimental data of participant 513, is used.

According to the phenomena described earlier, namely that persons with a `high extraversion` are more stable on positive affects, I tested different settings for the difference between κ_{pos} and κ_{neg} (cf. Figure 8.7). Here, I basically distinguish two different settings. First, a very reinforcing setting, where positive observations are emphasised and negative observations are suppressed. Secondly, a very suppressive setting where the mood development behaves the other way around, positive values are suppressed and negative ones emphasised. Again, the annotated emotional traces of ES2 are used and I included the previous considerations on the κ_η-values and chose only values between 0.1 to 0.3.

By varying these values, I could change the behaviour of the model to match the different settings of emotional stability. Although the input data remains the same, the emotional influence of positive observations on the mood can either be very suppressive or very reinforcing, depending only on the adjustment factor κ_η, as seen in Figure 8.7.

The `extraversion` for each subject can be obtained from the NEO-FFI questionnaire. The values for `extraversion` gathered from the questionaires are normalised in the range of $[0, 1]$. Thus, a high `extraversion` is denoted with values above 0.5 and a low `extraversion` by valies below 0.5. To obtain a mood model that reproduce the expected behaviour, the values for the parameter-pair κ_{pos} and κ_{neg} have to be chosen adequatly. In Table 8.4, I present my suggestions for plausible adjustment values based on the `extraversion` gathered from questionnaires.

Table 8.4: Suggested κ_{pos} and κ_{neg} values based on the `extraversion` personality trait.

`Extraversion`	κ_{pos}	κ_{neg}
>0.7	0.3	0.1
0.6-0.7	0.3	0.2
0.4-0.6	0.2	0.2
0.2-0.4	0.2	0.3
<0.2	0.1	0.3

Finally, I used the complete session of one experiment, labelled with GTrace by one annotator, again using the same initial model parameter (cf. Table 8.2 on page 199). This emotional annotation serves as input value for my mood modelling. By doing so, it is possible to form a mood development over a whole experiment and compare the calculated model with the experimental descriptions, as ground truth, of the complete experiment (cf. Table 8.3 on page 200). The whole mood development and the division into the single ESs are shown in Figure 8.8. I concentrated on the `pleasure`-dimension, as for this secured studies on the EmoRec corpus are available (cf. [Walter et al. 2011; Böck et al. 2012b]) showing that the experiment induces an "emotional feeling" that is measurable. Investigations with emotion recognisers using prosodic, facial, and biophysiological features and the comparison to the experimental design could support that the participants experienced ES2 as mostly positive and ES5 as mostly negative (cf. [Walter et al. 2011]). The underlying experimental design – the ground truth for my mood model – is described as follows: in ES1, ES2, ES3, and ES6 mostly positive emotions, in ES4 the emotional inducement goes back to a neutral degree. In ES5 mostly negative emotions were induced.

Using the emotional labels as input data for the modelling, I was able to demonstrate that the mood follows the prediction for the pleasure dimension of given ESs in the experiment (cf. Figure 8.8). The advantage of this modelling is that the entire sequence can be represented in one course. Furthermore, the influence of a preceding ES onto the actual ES is included in the mood modelling.

The resulting mood development is as follows: in the beginning of ES1 the mood rests in its neutral position and it takes some time, until the mood starts to shift towards the positive region. In ES2 and ES3 the mood continues to rise. In the beginning

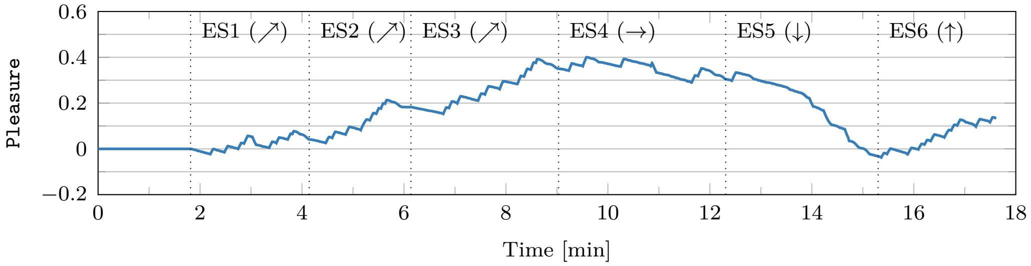


Figure 8.8: Course of the mood model using the whole experimental session of participant 513. The experimentals design ground truth, how the subject's `pleasure` value is changing over time is additionally denoted (cf. Table 8.3 on page 200).

of ES4 the mood reaches its highest value of 0.40 at 9.58 min. As in this ES, the inducement of negative emotions has started, the mood is decreasing afterwards. But as the previous induced positive emotions lead to a quite high damping in the direction of a negative mood, the mood falls just slowly. In ES5, when more negative emotions were induced due to negative feedback and time pressure, the mood is decreasing quite fast. At the end of ES5 the mood reaches its lowest value with −0.037 at 15.36 min. Here, it should be noted that negative emotional reactions are observed already in the very beginning of ES5, otherwise the strong decreasing of the mood could not have been observed. The mood remains quite low in the beginning of ES6. During the course of ES6, where many positive emotions were induced, the mood rises again and reaches 0.14 at the end of the experiment (17.60 min).

8.4 Summary

In this chapter I propose a mood modelling from a technical perspective that is able to incorporate several psychological observations. After describing the desired mood development, I presented my technical mood implementation. This implementation has the advantage of only three internal parameters (D_0, μ_1 and μ_2) and one user-specific parameter-pair, κ_{pos} and κ_{neg}. The mood development is oriented on a mechanical spring model.

Using two different experiments, I was able to evaluate the principal function of the proposed model on two different databases. Especially on the EmoRec Woz corpus, I was able to show that the generated mood course matched the experimental setting. By this, I could support my first hypothesis (cf. Hypothesis 8.1 on page 193) that the mood course can be modelled by a mechanical spring model.

By utilising the user-specific parameter-pair κ_{pos} and κ_{neg} the personality trait `extraversion` was integrated. This trait is supposed to regulate the individual emotional experiences. This supports my second hypothesis (cf. Hypothesis 8.2 on page 193) that the individual emotional experience can be modelled by one factor-pair.

As mentioned in Section 8.2, the presented results are based on labelled emotions, located in PAD-space according to their label. Until now, this model could only be tested with emotion values gathered due to a labelling process, as a continuous automatic recognition of emotional values in PAD-space is still under research. Another problem that has to be addressed is the need of equally distributed emotion values over time. To date, the emotional assessments are processed without regarding a gap between them, but this cannot always be guaranteed, especially when using automatically recognised emotion values. Here a further extension of my model is needed by, for example, a temporal averaging or weighting technique.

Chapter 9

Conclusion and Open Issues

Contents

THE preceding chapters of this thesis present my methods, improvements, and results achieved for speech-based emotion recognition, identified as open issues in Section 3.4. In this chapter an overall conclusion is given (cf. Section 9.1) and a further roadmap to integrate the issues addressed in this thesis is discussed in the outlook (cf. Section 9.2).

9.1 Conclusion

Emotion recognition is used to improve the HCI. The interaction should move beyond a pure command execution and reach a more human-like and more naturalistic way to interact. In this context the research field of “affective computing” was introduced (cf. [Picard 1997]). In this context, the term “companion” for a conversational system, having a human understanding of how something is said and meant, was introduced in [Wilks 2005]. The DFG-funded research project “A Companion-Technology for Cognitive Technical Systems”, under which this thesis has originated, also contributed to this research goal (cf. [Wendemuth & Biundo 2012]). Interdisciplinary research between psychology, neuroscience, engineering and computer science is needed to achieve the aim of developing a technical system that is able to understand the user’s abilities, preferences, and current needs. This thesis examined the speech-based emotion recognition from an engineer’s perspective and incorporates psychological as well as linguistic insights by transforming them into technically executable systems.

In addition to the motivation for the need of a speech-based emotion recognition for future HCI, in Chapter 1 I introduced the methodology of supervised pattern recognition, used as a foundation for speech-based emotion recognition, and I elaborated the three parts “annotation”, “modelling”, and “recognition”. I also emphasised that psychological insights are needed for a successful emotion recognition, especially for emotional annotation as well as for modelling.

The necessary psychological insights are presented and discussed in Chapter 2. This chapter deals with the question how emotions can be described, and how they become manifested in measurable acoustic characteristics. Therefore, several representations of emotions are depicted. In Section 2.1, I distinguish both categorial and dimensional representations. Categorial representations have the advantage that each category has a distinct label allowing to discriminate emotions. But the number of labels and their naming is still a matter of research. Dimensional representations in contrast have the advantage of indicating a relationship along dimensional axes. But the relevance of certain axes and the location of certain emotions within the emotional space is still being researched, although there is, up to a certain degree, an agreement on elementary emotions (cf. [Plutchik 1991; Ekman 1992]) and on main dimensions (cf. [Mehrabian & Russell 1977; Gehm & Scherer 1988]). The second part of this chapter (cf. Section 2.2) concentrates on the measurability of emotions. Here, the appraisal theory introduced by Scherer [2001] attempts to answer the question of how an observed event causes a bodily reaction. I discussed further studies to answer that question for both facial expressions and acoustic characteristics (cf. [Kaiser & Wehrle 2001; Johnstone et al. 2001]). Another impact that the appraisal theory implies is the problem of verbalisation of emotional experiences (cf. [Scherer 2005a]). This leads to the fact that not all emotional events can be correctly named, which increases the uncertainty of emotional annotation. Chapter 2 is completed by describing the further affective states "moods" and "personality" (cf. Section 2.3). Moods are defined as affective states lasting longer than emotions that are generally distinguished by their positive or negative value (cf. [Morris 1989]). Personality reflects the human's individual differences in mental characteristics and are nearly stable over the whole life (cf. [Becker 2001]). Both moods and personality are of importance for HCI as they influence the way different users judge the same situation, but they are often neglected in HCI research. This is mostly due to the difficulty of measurement as they are commonly only captured by questionnaires. And their impact on acoustic characteristics is hard to measure as it has been demonstrated, for instance, in the INTERSPEECH 2012 Speaker Trait Challenge (cf. [Schuller et al. 2012a]).

Chapter 3 reviews the current state-of-the-art in speech-based emotion recognition. In Section 3.1 I depict the evolution of datasets used for emotion recognition. In the last years the focus changed from datasets with simulated emotions to more naturalistic ones. Additionally, further modalities, like video or bio-physiological data are recorded in combination with audio. A non-exclusive list of emotional speech databases is presented in Table 3.1 on page 31. Afterwards several important methods for a speech-based emotion recognition are presented (cf. Section 3.2). This includes utilised features and pre-processing steps as well as applied classifiers. Efforts in classifier evaluation are also depicted by mentioning the various recognition challenges as well as benchmark corpora. Both reviews are summarised by discussing the development of recognition results on some example corpora (cf. Section 3.3). Chapter 3 is completed by emphasising certain open issues, which I identified as being not in the current focus of research and which are pursued in this thesis.

The Chapters 6 to 8 of this thesis will present my results in the aim of closing these open issues. The first two open issues are directly related to emotional pattern recognition and thus, are examined within one chapter. The last two issues are going beyond this approach and separate chapters are devoted to them.

Previously, however, necessary methods for this thesis are introduced in Chapter 4. Section 4.1 describes methods for the emotional annotation. I present the commonly used EWLs, GEWs and SAMs. This overview is completed by mentioning further labelling approaches used for special applications. This section also introduces the inter-rater reliability as a measure allowing a statement on the validity of the used labelling scheme. To date, considerations on reliability are just rarely taken into account while generating emotional datasets. Thus, the reliability is introduced with a focus on kappa-like coefficients and a comparison of interpretation schemes.

The next section describes the features utilised in this thesis in more detail (cf. Section 4.2), starting with the description of speech production and the variation of these characteristics due to emotional reactions. Afterwards common short-term segmental features such as MFCCs, PLP, and formants are introduced. The list of acoustic features is completed by describing longer-term features to include contextual information and prosodic cues such as pitch, jitter, or shimmer. Investigations of how these characteristics are influenced by ageing or the speaker's gender are also discussed, together with the description how these features can be extracted from a speech signal. Afterwards, the two classifiers utilised in this thesis, namely HMMs and GMMs are described and methods for defining optimal parameters are discussed (cf. Section 4.3). As an important part of this section also methods known from ASR to incorporate speaker characteristics by VTLN and SGD modelling are introduced. These methods have not been widely used in emotion recognition, yet. Lastly, common fusion techniques are depicted and the MFN, as it is later utilised for my own research, is explained in more detail.

The last section (cf. Section 4.4) describes common classifier evaluation methods. I present different data material arrangements to generate a validation set. Classifier performance measures and their differences are also discussed, concluding that UAR is the preferred measure in the emotion recognition community. For my later investigations, I also investigate the statistical significance of the achieved improvements. Thus, I introduce the needed principles of statistical analyses to perform an ANOVA.

As already stated, the classification performance heavily depends on the material used for training and validation. To reproduce the experiments, the applied corpora have to be described. This is done in Chapter 5, distinguishing the previously introduced two types of datasets, namely corpora of simulated and naturalistic emotions. The recordings for simulated emotions were conducted under controlled conditions and contained short, emotional statements, or alternatively, emotional stimuli performed by actors are presented to a subject, whose reactions were recorded. I have chosen one database, emoDB, which widely serves as a benchmark test to compare and evaluate new methods. Several publications reporting about recognition results exists, allowing the comparison of my own results to those from other researchers. This corpus contains prototypical emotional statements with a high expressiveness where a high classifier performance could be expected (cf. Section 5.1.1). In contrast, I have chosen three naturalistic datasets in order to meet the current development to migrate to this kind of corpora. The NIMITEK Corpus (cf. Section 5.2.1) was directly designed to investigate emotional speech during HCI. It was gathered during a WOZ setup and negative emotions should be elicited by increasing the stress level during the experiment. This corpus was mainly used in this thesis to develop emotional labelling techniques that were later transferred to other databases. The VAM corpus (cf. Section 5.2.2) contains spontaneous and unscripted discussions from a German talk show. It was labelled using SAM, allowing also to test the methods developed in this thesis on a dimensional representation of emotions. Furthermore, this dataset also provides information on the speakers' age and gender, allowing to include these speaker characteristics. The most important dataset for this thesis is LMC (cf. Section 5.2.3). It contains multimodal recordings of a WOZ experiment. This corpus was not intended to directly provoke emotional reactions, but to investigate how users interact with a technical system when significant dialogue barriers are arising. Thus, this corpus can be seen as the most naturalistic one, focussing just on the interaction. Of interest for my thesis are the dialogue barriers `baseline` and `challenge`. As this dataset contains quite long interactions of about 30 min, it also allows to investigate further interaction patterns. Additionally, this dataset provides information about the age and gender of the participants as well as analyses of their personality traits. This allows to incorporate these additional user traits into the investigations.

The following Chapter 6 presents my own research to improve the speech-based emotion recognition. As the community migrates from simulated emotions towards naturalistic interactions, the difficulty in the annotation of subjective emotional observations is arising. To support the process of manual emotional labelling, I present the tool *ikannotate*. This tool can be used to transcribe and annotate utterances. Then, and more importantly, the utterances can be labelled emotionally using various common methods namely EWLs, GEW and SAM, as presented in Section 4.1. This tool is used for my continuing investigations on emotional annotation.

Based on the reviews of emotional labelling efforts, I proposed the two hypotheses that the application of well-founded emotional labelling methods results in a proper emotion coverage with broader meaningful emotional labels (cf. Hypothesis 6.1 on page 119) and it results in the possibility to obtain a proper decision for emotional labels for all samples (cf. Hypothesis 6.2 on page 119). The experiments conducted on NIMITEK confirm Hypothesis 6.1 as they reveal that the GEW is able to cover a wide range of emotional observations and allow clustering of emotions, as the labels at a later time are related to each other. An EWL with basic emotions does not cover especially the weaker emotions occurring in HCI. SAM labellings are hard to compare as the interpretation of the graduation on each dimension is very subjective and translation into emotional labels is not possible. Hypothesis 6.2 is also confirmed, in this case labelling with an EWL consisting of basic emotions or a GEW results in no or just a few utterances remaining without a decision. In the case of SAM, this amount of undecided utterances is much larger, at around 30 %. Thus, regarding GEW as labelling method, both hypotheses could be confirmed with my experiments.

Afterwards, I investigated how the reliability of emotional annotation can be improved. The reliability is a measure for the quality of an annotation that is so far mainly neglected for emotional speech corpora. I raise two hypotheses: For emotional labelling the achieved IRR is generally low (cf. Hypothesis 6.3 on page 125) but the incorporation of visual and context information improves the IRR (cf. Hypothesis 6.4 on page 125). The preselection of emotional episodes circumvents the second kappa paradox (cf. Hypothesis 6.5 on page 125). In my investigations, I calculated the reliability for some popular emotional speech databases and for my own labelling pursued on NIMITEK. For all annotations the reliability is just *slight* to *fair* independent of the emotional content or the utilised labelling method. Thus, I deem Hypothesis 6.3 as confirmed. Afterwards, I conducted emotional labelling experiments on LMC. In these, I first defined a list of eleven emotional terms suited to label naturalistic emotions. By including visual information as well as the course of the interaction I could increase the reliability reaching a *moderate* reliability in the end. Through these experiments, I could confirm Hypothesis 6.4. In [Feinstein & Cicchetti 1990] the authors describe the paradox that kappa is decreasing when the distributions of agreement across different categories are not equal although the observed agreement remains high. In emotional labelling, especially the number of neutral labels is quite high and thus, the distribution of agreement is highly unbalanced. With an experiment that preselects any parts where an emotional reaction can be expected, I could rebalance the number of categories and thus also circumvent this second kappa paradox. This confirms Hypothesis 6.5. As a further result of my investigations, I could also prove that the occurring emotions in the dialogue barriers `baseline` and `challenge` are different. While in the `bsl` event `interest`, `relief` and `concentration` are dominating, the `cha` event is dominated by `surprise`, `confusion` and `concentration`.

Thus, the first open issue "a reliable ground truth for emotional pattern recognition" is considered as answered.

In Section 6.2, I investigated whether speaker characteristics improve the speech-based emotion recognition. For this, I raised the hypotheses that the consideration of the speaker's age and gender can improve the emotion recognition (cf. Hypothesis 6.6 on page 136) and that SGD modelling results in a higher improvement than performing an acoustic normalisation like VTLN (cf. Hypothesis 6.7 on page 136). For these investigations, I first performed a parameter tuning to adjust the number of mixture components and iterations at best. Afterwards, I defined speaker groups oriented at commonly used groupings for age and gender recognition as these groups tend to be a good starting point. The utilised datasets should represent a broad variety of different characteristics. To emphasise the general applicability of my presented method emoDB is chosen because of its high recording quality and very expressive acted basic emotions. VAM represents spontaneous emotions dimensionally labelled and LMC incorporates naturalistic interactions with broader emotional reactions after specific dialogue barriers. For all datasets, I define a gender grouping (SGDg) and for VAM and LMC age (SGDa) as well as age-gender grouping (SGDag) to train SGD emotion classifiers. The achieved results are compared to a corpus specific but speaker unspecific classifier (SGI) as well as to results from other research groups presented in Section 3.3. On each corpus, the SGDg classifier reaches an improvement between 1.7 % to 6.6 %, which is for some results even significant, against the corresponding SGI result. The SGDa classifiers on VAM and LMC also show an improved performance, but it falls behind the SGDg results. The combination of both groupings (SGDag) could further improve the performance just for LMC. These experiments confirm Hypothesis 6.6 on page 136 that emotion recognition has to consider the speaker's age and gender. In this context, it can be noticed that the achieved improvement is also influenced by the recording quality and expressiveness of the emotions. For emoDB, a dataset of very high recording quality and high expressiveness, the effects using SGD classifiers are weaker than on VAM and LMC. Afterwards, I conducted experiments using VTLN to adjust the acoustic differences and perform the emotional classification. The results are compared to my SGI results. I am able to state that the improvement achieved with VTLN falls with 0.5 % to 5.1 % behind the SGD modelling approach. Thus, I could show that an acoustic normalisation could not ensure an improvement to the same extent as my SGD modelling approach and could also confirm Hypothesis 6.7 on page 136.

In Section 6.3 the findings on SGD modelling are applied for multimodal fusion of fragmentary data. The difficulty is that for multimodal emotion recognition not all data streams are available all the time and thus the decisions have to be based on just partly available unimodal classification results, especially speech-based emotion recognition can only be pursued when the user is speaking. In this context, I raised the hypothesis that the by acoustic classification which was improved by SGD modelling also improves the fused classification result although the acoustic channel is present quite rarely (cf. Hypothesis 6.8 on page 166). The investigations are conducted on a sub-set, the "20s set", of LMC for validation. This time a continuous classification is pursued using an MFN which fuses the classification results from visual and acoustic information. Over the entire corpus of utilised material, the average amount of speech is just 12 %. By incorporating SGDag based classifiers, I could improve the fusion accuracy by about 5.5 % in total, which confirms Hypothesis 6.8.

Thus, the second open issue "incorporating speaker characteristics" has been answered.

Chapter 7 leaves the methodological improvement of speech-based emotion recognition and introduces a new pattern, Discourse Particles (DPs), that comprises information on the interaction progress of HCI. Thereby, I utilise a new pattern needed to evaluate longer interactions. Based on the investigations of Schmidt [2001], proposing a form-function relation of DPs, I raised the following two hypotheses: DPs occur more frequently at critical points within an interaction, which helps to identify potential dialogue abortions (cf. Hypothesis 7.1 on page 174). The differences in the pitch contour can be automatically recognised (cf. Hypothesis 7.2 on page 174). To conduct these investigations, I utilise another sub-set, the "90s set", of LMC. I furthermore incorporate the findings of Section 6.2, by also distinguishing the age and gender groupings to unfold the DPs analyses from these effects. My analyses reveal that DPs occur sufficiently within an HCI and are used significantly more often within critical situations, represented by the `challenge` barrier. Afterwards, I also showed that an automatic identification of DPs functions `thinking` and `not-thinking` purely based on the pitch-contour is possible, and thus, reached a UAR of 89 %. Thus my hypotheses are confirmed.

The third open issue "interactions and their footprints in speech" has therefore been addressed.

Chapter 8 is dedicated to the fact that observations of emotions and interaction signals alone are not sufficient to understand or predict human behaviour and intelligence. In addition to emotional observations and interaction patterns, a description of the interaction's progress is necessary as well. For this, a mood modelling is helpful and I raised the hypotheses that the mood can be modelled by a spring model considering emotional observations as forces and a changing damping term (cf. Hypothesis 8.1 on page 193). Furthermore, the observed emotion is dependent on the personality trait `extraversion` which can be easily included into the model (cf. Hypothesis 8.2 on page 193). My presented implementation has the advantage of having only three internal parameters. I used two different experiments to evaluate the general function of the proposed model for two different databases. Especially on the EmoRec-Woz I Corpus, I was able to show that the generated mood course matched the experimental setting. Hereby, I confirmed Hypothesis 8.1. By incorporating a user-specific parameter-pair, influenced by the trait `extraversion`, and by conducting experiments on EmoRec, I could support Hypothesis 8.2 on page 193 that the individual emotional experience can be modelled by just one factor.

Thus, the fourth open issue "modelling the temporal sequence of emotions in HCI" has been addressed.

9.2 Open Questions for Future Research

Summarizing, I can state that during my work presented in this thesis, the four open issues I identified in Chapter 3 are properly addressed. I showed that the extension of the pure acoustic emotion recognition by considering speaker characteristics, feedback signals and personality traits allows to examine longer-lasting natural interactions and to identify critical situations. Of course it is not possible to resolve the open issues identified in this thesis completely, as the work on this topic is still not finished, thus open questions are outstanding and have to be investigated in future research. In the current Section, I will address open questions, needed to develop technical systems that comprise the users individual interaction behaviour.

Reliable Ground Truth Although my work presents methodological improvements for the emotional labelling, there are still open questions. As I have shown, the commonly used interpretations for the reliability used in linguistic content analysis are not suited for emotional annotation. The reliability of emotional annotation does not exceed values above the 0.75 needed to be interpreted as *very good* or *excellent* in terms of content analysis. Thus, an open question for further research is:

- How must an adequate interpretation scheme for emotional labelling be defined?

Incorporation of Speaker Characteristics The speaker groupings, which I utilised for my SGD approach, were based on the speakers' characteristics age and gender. I reviewed physical and psychological evidences that these factors influence the acoustic characteristics. But it has to be investigated whether other factors are better suited to the improve recognition accuracy. To investigate this, corpora are necessary having further information about users. Unfortunately, such databases are quite rare.

Another aspect that has to be analysed is the question if all acoustic features are influenced by different speaker characteristics to the same extent. Additionally, a method is needed to adapt different SGD models. It has to be investigated if for this case the same technique as GMM-UBM with MAP and MLLR adaptation is useful. These considerations result in the following research questions:

- What are the best grouping factors for an improved emotional recognition?
- Which features are influenced by a speaker grouping and to what extent?
- Is it possible to utilise MAP and MLLR adaptation for different SGD models?

Discourse Particles as Interaction Patterns As I have demonstrated, DPs are useful patterns for the evaluation of HCI, especially to indicate stressful parts in the dialogue. But their detection is difficult. A first approach to detect “uh” and “uhm” is presented in [Prylipko et al. 2014b] which I co-authored, but a reliable automatic detection of “hm” could to date not be reached.

In addition to DPs, other interaction patterns are known, for instance crosstalk and offtalk. These patterns are mostly neglected in today's acoustic interaction analyses, but could reveal information about the user's turn-taking behaviour and the user's self-revelation. It has to be investigated whether these patterns are helpful for HCI and especially whether there is a general relationship between the occurrence of these patterns and specific states of the interaction. From this, the following questions arise:

- How can “hm” be automatically and reliably detected?
- Which other interaction patterns can be used to improve the analyses of HCI?

Modelling the Emotional Development My presented mood model, which indicates the emotional development during an interaction, shows the principle function of a user's mood prediction based on emotional observations. But to show its applicability, it has to be integrated into a realistic scenario, with the possibility to evaluate the model predictions against the user's actual feelings. This results in the following research questions:

- How can the model's parameters be automatically adjusted?

- In which way can the prediction of the model be utilised in the current dialogue?

If these open questions are satisfactorily answered, the research community is several steps closer to affective computing for naturalistic interactions. Thereby, it is possible to develop systems that interact in a cooperative and competent manner with their user so that he is supported in his daily life.

Glossary

n-fold cross validation

The samples of all speakers within a corpus are randomly partitioned into n subsamples (folds) of equal size. One subsample is retained for validation, the $n - 1$ folds are used for training. This procedure is repeated n times where each subsample is exactly used once for validation. The overall estimation is achieved by averaging the n results [Kohavi 1995].

Annotation

Annotation describes the step of adding the information how something has been said.

Appraisal

An appraisal is the theory in psychology that emotions are extracted from our evaluations of events that cause specific reactions in different people, measurable as emotional bodily reactions. Mainly the situational evaluation causes an emotional, or affective response [Scherer 2001].

Arousal

`Arousal` is an emotional dimension. It is also called `activation` or `excitement`. This dimension is mostly seen as second component in an emotion space. It determines the level of psychological arousal or neurological activation [Becker-Asano 2008].

Dominance

The emotional dimension `dominance` is also called `attention`, `control`, or `power`. Especially in the case of a high `arousal` the dimension of `dominance` is useful to distinguish certain emotion-describing adjectives [Scherer et al. 2006].

Emotion

Emotions reflect short-term affects, usually bound to a specific event, action, or object [Becker 2001].

Human-Computer Interaction

Human-Computer Interaction, also denoted as Human-Machine Interaction, describes the communication and interaction of one or multiple humans with a technical system.

Human-Human Interaction

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

Inter-Rater Reliability

Inter-Rater Reliability determines the extent to which two or more raters obtain the same result when measure a certain object [Kraemer 2008].

Intra-Rater Reliability

Intra-Rater Reliability compares the deviation of the assessment, which is completed by the same rater on two or more occasions [Gwet 2008a].

Labelling

Labelling describes the process of adding further levels of meaning. These levels are detached from the textual transcription and describe, for instance, affects, emotions, or interaction pattern.

Leave-One-Speaker-Group-Out

The corpus is partitioned into several parts, containing just material of a certain speaker group. All but one group is used for training, the remaining one for testing. This procedure is repeated for each speaker group. The overall estimation is achieved by averaging the speaker group's results [Kohavi 1995].

Leave-One-Speaker-Out

The material of all speakers within a corpus except one particular speaker is used for training. The remaining data of this speaker is applied for validation. This procedure is repeated for each speaker of the corpus. The overall estimation is achieved by averaging the speaker's individual results [Kohavi 1995].

Mood

Moods reflect medium-term affects, generally not related to a concrete event, but a subject of certain situational fluctuations that can be caused by emotional experiences. They last longer and are more stable affective states than emotions. Moods influence the user's cognitive functions directly [Morris 1989].

OCC model

Ortony, Clore and Collins's model of emotion is a widely used computational model for affective embodied agents. It states that the strength of a given emotion primarily depends on the events, agents, or objects in the environment of the agent exhibiting the emotion. OCC specifies about 22 emotion categories by using five processes to evaluate the events and getting the resulting emotional state (cf. [Ortony et al. 1990]).

Personality

Personality reflects a long-term affect and individual differences in mental characteristics. It comprises distinctive and characteristic patterns of thought, emotion, and behaviour that make up an individual's personal style of interacting with the physical and social environment [Nolen-Hoeksema et al. 2009].

Pleasure

`Pleasure` (`valence`) is agreed to be the first and most important dimensional emotion component as an emotion is either positive or negative [Becker-Asano 2008].

Transcription

Transcription denotes the process of translating the spoken content into a textual description.

Unweighted Average Recall

The Unweighted Average Recall is the extended version of the two-class recall definition, by calculation a class-wise recall and averaging over all classes.

Wizard-of-Oz scenario

In this scenario the application is controlled by an invisible human operator, while the subjects believe to talk to a machine.

Abbreviations

ABC	Airplane Behavior Corpus
ACC	Affective Callcenter Corpus
ANN	Artifical Neural Network
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
AR	Articulation Rate
ASF-E	Attributionsstilfragebogen für Erwachsene (Attributional style questionnaire for adults)
ASR	Automatic Speech Recognition
AU	Action Unit
AVEC	Audio/Visual Emotion Challenge and Workshop
AvR	Average Recall
BELMI	Berlin Everyday Language Mood Inventory
BIS/BAS	Questionnaire on the bipolar BIS/BAS scales
BNDB	Belfast Naturalistic Database
BW	Baum-Welch
c	children
CallfriendEmo	Emotional Enriched LDC CallFriend corpus
CC	Correlation Coefficient
CEICES	Combining Efforts for Improving automatic Classification of Emotional user States
CHAT	Codes for the Human Analysis of Transcripts
CMS	Cepstral Mean Subtraction
DCT	Discrete Cosine Transformation
DES	Danish Emotional Speech
DES-IV	Differential Emotions Scale (Version 4)
DFT	Discrete Fourier Transformation
DP	Discourse Particle
DTW	Dynamic Time Warping
EM	Expectation-Maximization
emoDB	Berlin Database of Emotional Speech
emoSDB	emotional Speech DataBase
eNTERFACE	eNTERFACE'05 Audio-Visual Emotion Database
ERQ	Emotion Regulation Questionnaire
ES	Experimental sequence
EWL	Emotion Word List
f	female speaker

FACS	Facial Action Coding System
FFT	Fast Fourier Transformation
FN	False Negative
FP	False Positive
FS	Feature Set
`ft`	female teens
GAT	Gesprächsanalytisches Transkriptionssystem (dialogue analytic transcription system)
GEMEP	GEneva Multimodal Emotion Portrayals
GEW	Geneva Wheel of Emotions
GMM	Gaussian Mixture Model
GSR	Global Speech Rate
GUI	Graphical User Interface
HCI	Human-Computer Interaction
HHI	Human-Human Interaction
HIAT	halb-interpretative Arbeits-Transkription (semi-interpretive working transcription)
HMM	Hidden Markov Model
HNR	Harmonics-to-Noise Ratio
HTK	Hidden Markov Toolkit
IIP	Inventory of Interpersonal Problems
ikannotate	*i*nterdisciplinary *k*nowledge-based *anno*tation *t*ool for *a*ided *t*ranscription of *e*motions
IRR	Inter-Rater Reliability
LLD	Low Level Descriptor
LMC	LAST MINUTE corpus
LOO	Leave-One-Out
LOSGO	Leave-One-Speaker-Group-Out
LOSO	Leave-One-Speaker-Out
LPC	Linear Predictive Coding
LSTMN	Long Short-Term Memory Network
`m`	male speaker
$\underline{\texttt{m}}$	middle aged
$\underline{\texttt{m}}\texttt{f}$	middle aged females
$\underline{\texttt{m}}$	middle aged males
MAP	Maximum A Posteriori
MFCC	Mel-Frequency Cepstral Coefficient
MFN	Markov Fusion Network
MLLR	Maximum Likelihood Linear Regression
MLP	Multi Layer Perceptron

mt	male teens
MV	Majority Vote
NEO-FFI	NEO Five-Factor Inventory
NES	Neuro-Endocrine System
NIAVE	New Italian Audio and Video Emotional Database
s	seniors
PAD	Pleasure-Arousal-Dominance
PANAS	Positive and Negative Affect Schedule
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
PPS	Phonemes Per Second
PrEmo	Product Emotion Measurement Tool
RASTA	RelAtive SpecTrAl
RBF	Radial Basis Function
RF	Random Forest
S-MV	Soft-Majority Vote
SAFE	Situation Analysis in a Fictional and Emotional Corpus
SAL	Belfast Sensitive Artificial Listener
SAM	Self Assessment Manikins
SDC	Shifted Delta Cepstra
SDMS	static and dynamic modulation spectrum
sf	senior female adults
SGD	Speaker Group Dependent
SGDa	age specific Speaker Group Dependent
SGDag	age and gender specific Speaker Group Dependent
SGDg	gender specific Speaker Group Dependent
SGI	Speaker Group Independent
SIFT	Simplified Inverse Filtering Technique
sm	senior male adults
SNS	Somatic Nervous System
SPM	Syllables Per Minute
SPS	Syllables Per Second
SPSS	Statistical Package for the Social Sciences
SR	Syllable Rate
SRN	Simple Recurrent Neural Network
SUI	Speech User Interface
SUSAS	Speech Under Simulated and Actual Stress Database
SVF	Stressverarbeitungsfragebogen (stress-coping questionnaire)
SVM	Support Vector Machine

t	teens
TA-EG	Questionnaire for the assessment of affinity to technology in electronic devices (Fragebogen zur Erfassung von Technikaffinität in elektronischen Geräten)
TEO	Teager Energy Operator
TN	True Negative
TP	True Positive
TUM AVIC	Audivisual Interest Corpus
UAH	UAH emotional speech corpus
UAR	Unweighted Average Recall
UBM	Universal Background Model
VAM	Vera am Mittag Audio-Visual Emotional Corpus
VTLN	Vocal Tract Length Normalisation
WAR	Weighted Average Recall
WIMP	Window, Icon, Menu, Pointing device
WOZ	Wizard-of-Oz
WOZdc	WOZ data corpus
WPM	Words Per Minute
y	young adults
yc	young children
yf	young female adults
ym	young male adults

List of Symbols

A_e	Expected agreement
A_o	Observed agreement
Acc	Accuracy rate
agr_i	Agreement value for sample i
α_κ	Artstein and Poesios's measure for inter-rater agreement
α_C	Cronbach's alpha
α_K	Krippendorffs' alpha
A_i	Peak-to-peak signal amplitude
$s_{XX}(\kappa)$	Autocorrelation function of the signal X
C0	Zeroth cepstral coefficient
c_t	Coefficient at frame t
T_{critic}	Critical value of significance niveau
$s_{XY}(\kappa)$	Cross-correlation function of the signals X and Y
D_e	Expected disagreement
D_o	Observed disagreement
D_t	Calculated damping of mood at time point t
Δ	Delta regression coefficient
$\Delta\Delta$	Double delta regression coefficient (Acceleration)
$\Delta\Delta\Delta$	Third order regression coefficient
disagr_i	Disagreement value for sample i
e_t	Emotional observation at time point t
E	Energy term
Err	Error rate
$u(n)$	Excitation
F_1	F_1-score
F_β	F-score
a_i	Filter coefficient for LPC
F_t	Derived force of emotional observation at time point t
F_i	ith Formant
F_0	Fundamental frequency
H_1	Alternative hypothesis
H_0	Null hypothesis
π_s	Initial parameter of a state s for an HMM
I	Intensity term
L_I	Sound intensity level

Jitter_{abs}	Absulute jitter
κ	Cohens' kappa
multi-κ	Cohens' multi-kappa
κ_w	Cohen's weighted kappa
K	Fleiss' kappa
π	Scott's pi
M_t	Mood object at time point t
κ_η	Adjustment factor to regulate the emotional force
$\mathbf{n}_c$	Number of items assigned by all raters to category c
$\mathbf{n}_{rc}$	Number of items assigned by raters r to category c
$\mathbf{n}_{ic}$	Number of raters who assigned item i to category c
NPV	Negative predictive value
V	Output alphabet of an HMM $V = \{v_1, \ldots, v_n\}$
T_i	Extracted F_0 period lengths
θ_0	Phenomena of H_0
θ	Postulated phenomena of H_1
PPV	Positive predictive value
P_{ac}	Acoustic sound power
Pre	Precision
$P(\mathbf{O}\|W)$	Conditional probability that the observation $\mathbf{O}$ is generated given the sequence of words W
$P(W\|\mathbf{O})$	Conditional probability that a sequence of words W is generated given the observation $\mathbf{O}$
$P(W)$	A priori probability of the sequence of words W
$P(c)$	Proportion of items assigned to category c despite the rater
$P(c\|r)$	Proportion of items assigned by rater r to category c
b_{jk}	Production probability of an HMM
Rec	Recall
$\mathit{Shimmer}_{abs}$	Absulute shimmer
α	Level of significance
Spe	Specificity
S	State sequence $S = \{s_1, \ldots, s_n\}$
a_{ij}	State transition probability of an HMM
T	Test statistic

Abbreviations of Emotions

A arousal
A+ high arousal
A− low arousal
adm admiration
agg aggressive
amu amusement
ang anger
anx anxiety

bor boredom
bsl baseline

cha challenge
che cheerful
coc concentration
cof confident
com contentment
con confusion
cot contempt

D dominance
D+ high dominance (dominating)
D− low dominance (submissive)
des despair
dis disgust
dou doubt

emp emphatic
exa exaltation

fea fear

han hot anger
hap happy
hel helpless
hes hesitation
hst high stress
hur hurt

int interest
inx intoxicated
iro irony
irr irritation

joy joy

LoI level of interest
lst listing

mot motherese
mst medium stress

neg negative emotions
ner nervousness
neu neutral
nne non-negative emotions

oth other

pai pain
pan panic
ple pleasure
pos positive emotions
pri pride
puz puzzled

rel relief
rep reprimanding

sad sadness
scr screaming
ser serenity
sur surprise

ten tenderness

V valence
V+ positive valence
V− negative valence

wai waiuku
wor worry

References

Abrilian, S.; Devillers, L.; Buisine, S. & Martin, J. C. (2005). "EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces". In: *Proc. of the 11th HCII.* Las Vegas, USA, s.p.

Ai, H.; Litman, D. J.; Forbes-Riley, K.; Rotaru, M.; Tetreault, J. & Pur, A. (2006). "Using system and user performance features to improve emotion detection in spoken tutoring dialogs". In: *Proc. of the INTERSPEECH-2006.* Pittsburgh, USA, pp. 797–800.

Albornoz, E. M.; Milone, D. H. & Rufiner, H. L. (2011). "Spoken emotion recognition using hierarchical classifiers". *Comput Speech Lang* 25 (3), pp. 556–570.

Allport, G. W. & Odbert, H. S. (1936). "Trait-names, a psycho-lexical study". *Psychological Monographs* 47 (1), pp. i–171.

Allwood, J.; Nivre, J. & Ahlsén, E. (1992). "On the Semantics and Pragmatics of Linguistic Feedback". *Journal of Semantics* 9 (1), pp. 1–26.

Altman, D. G. (1991). *Practical Statistics for Medical Research.* London, UK: Chapman & Hall.

Amir, N.; Ron, S. & Laor, N. (2000). "Analysis of an emotional speech corpus in Hebrew based on objective criteria". In: *Proc. of the SpeechEmotion-2000.* Newcastle, UK, pp. 29–33.

Anagnostopoulos, T. & Skourlas, C. (Feb. 2014). "Ensemble Majority Voting Classifier for Speech Emotion Recognition and Prediction". *Journal of Systems and Information Technology* 16 (3), s.p.

Anusuya, M. A. & Katti, S. K. (2009). "Speech Recognition by Machine: A Review". *International Journal of Computer Science and Information Security* 6 (3), pp. 181–205.

Arlot, S. & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". *Statistics Surveys* 4, pp. 40–79.

Artstein, R. & Poesio, M. (Dec. 2008). "Inter-Coder Agreement for Computational Linguistics". *Comput Linguist* 34 (4), pp. 555–596.

Atal, B. S. (June 1974). "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification". *J Acoust Soc Am* 55 (6), pp. 1304–1312.

Atal, B. S. & Hanauer, S. L. (Aug. 1971). "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave". *J Acoust Soc Am* 50 (2), pp. 637–655.

Atal, B. S. & Stover, V. (1975). "Voice-excited predictive coding system for low-bit-rate transmission of speech". In: *Proc. of the IEEE ICC-1975.* San Francisco, USA, pp. 30–37.

Bachorowski, J. A. (1999). "Vocal expression and perception of emotion". *Curr Dir Psychol Sci* 8 (2), pp. 53–57.

Bahari, M. H. & Hamme van, H. (2012). “Speaker age estimation using Hidden Markov Model weight supervectors”. In: *Proc. of the 11th IEEE ISSPA*. Montréal, Canada, pp. 517–521.

Balomenos, T.; Raouzaiou, A.; Ioannou, S.; Drosopoulos, S.; Karpouzis, A. & Kollias, S. (2005). “Emotion Analysis in Man-Machine Interaction Systems”. In: *Machine Learning for Multimodal Interaction*. Ed. by Bengio, S. & Bourlard, H. Vol. 3361. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 318–328.

Banse, R. & Scherer, K. R. (1996). “Acoustic profiles in vocal emotion expression”. *J Pers Soc Psychol* 70, pp. 614–636.

Batliner, A.; Fischer, K.; Huber, R.; Spilker, J. & Nöth, E. (2000). “Desperately seeking emotions or: actors, wizards, and human beings”. In: *Proc. of the SpeechEmotion-2000*. Newcastle, UK, pp. 195–200.

Batliner, A.; Hacker, C.; Steidl, S.; Nöth, E.; Russell, M. & Wong, M. (2004). ““You stupid tin box”- children interacting with the AIBO robot: A Cross-Linguistic Emotional Speech Corpus”. In: *Proc. of the 4th LREC*. Lisbon, Portugal, pp. 865–868.

Batliner, A.; Steidl, S.; Hacker, C. & Nöth, E. (2008). “Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech”. *User Model User-Adap* 18 (1), pp. 175–206.

Batliner, A.; Fischer, K.; Huber, R.; Spilker, J. & Nöth, E. (2003). “How to find trouble in communication”. *Speech Commun* 40 (1-2), pp. 117–143.

Batliner, A.; Steidl, S.; Schuller, B.; Seppi, D.; Laskowski, K.; Vogt, T.; Devillers, L.; Vidrascu, L.; Amir, N.; Kossous, L. & Aharonson, V. (2006). “Combining Efforts for Improving Automatic Classification of Emotional User States”. In: *Proc. of the IS-LTC 2006*. Ljubljana, Slovenia, pp. 240–245.

Batliner, A.; Seppi, D.; Steidl, S.; & Schuller, B. (2010). “Segmenting into Adequate Units for Automatic Recognition of Emotion-Related Episodes: A Speech-Based Approach”. *Advances in Human-Computer Interaction* 2010, s.p.

Batliner, A.; Steidl, S.; Schuller, B.; Seppi, D.; Vogt, T.; Wagner, J.; Devillers, L.; Vidrascu, L.; Aharonson, V.; Kessous, L. & Amir, N. (Jan. 2011). “Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-related User States in Speech”. *Comput Speech Lang* 25 (1), pp. 4–28.

Baum, L. E. & Petrie, T. (Dec. 1966). “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”. *Ann. Math. Stat.* 37 (6), pp. 1054–1063.

Becker, P. (2001). “Structural and Relational Analyses of Emotions and Personality Traits”. *Zeitschrift für Differentielle und Diagnostische Psychologie* 22 (3), pp. 155–172.

Becker-Asano, C. (2008). “WASABI: Affect Simulation for Agents with Believable Interactivity”. PhD thesis. University of Bielefeld.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour.* Princeton, USA: Princeton University Press.

Benesty, J.; Sondhi, M. M. & Huang, Y. (eds.). *Springer Handbook of Speech Processing.* Berlin, Heidelberg, Germany: Springer.

Benus, S.; Gravana, A. & Hirschberg, J. (2007). "The Prosody of Backchannels in American Englisch". In: *Proc. of the 16th ICPhS.* Saarbrücken, Germany, pp. 1065–1068.

Berry, C. C. (Nov. 1992). "The kappa statistic". *JAMA-J Am Med Assoc* 268 (18), pp. 2513–2514.

Bishop, C. M. (2011). *Pattern Recognition and Machine Learning.* 2nd ed. Berlin, Heidelberg, Germany: Springer.

Bitouk, D.; Verma, R. & Nenkova, A. (2010). "Class-level spectral features for emotion recognition". *Speech Commun* 52 (7-8), pp. 613–625.

Bocklet, T.; Maier, A.; Bauer, J. G.; Burkhardt, F. & Noth, E. (2008). "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines". In: *Proc. of the IEEE ICASSP-2008.* Las Vegas, USA, pp. 1605–1608.

Boersma, P. (2001). "Praat, a system for doing phonetics by computer". *Glot International* 5 (9-10), pp. 341–345.

Bogert, B. P.; Healy, M. J. R. & Tukey, J. W. (1963). "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking". In: *Proc. of the Symp. on Time Series Analysis.* New York, USA. Chap. 15, pp. 209–243.

Bone, D.; Black, M. P.; Li, M.; Metallinou, A.; Lee, S. & Narayanan, S. S. (2011). "Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors". In: *Proc. of the INTERSPEECH-2011.* Florence, Italy. Chap. 15, pp. 3217–3220.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler.* 7. vollständig überarbeite Auflage. Berlin, Heidelberg, Germany: Springer.

Bozkurt, E.; Erzin, E.; Erdem, Ç. E. & Erdem, A. T. (2011). "Formant position based weighted spectral features for emotion recognition". *Speech Commun* 53 (9-10), pp. 1186–1197.

Bradley, M. M. & Lang, P. J. (1994). "Measuring emotion: The self-assessment manikin and the semantic differential". *J Behav Ther Exp Psy* 25 (1), pp. 49–59.

Braun, A. & Oba, R. (2007). "Speaking tempo in emotional speech – A cross-cultural study using dubbed speech". In: *Proc. of the ParaLing'07.* Saarbrücken, Germany, pp. 77–82.

Broekens, J. & Brinkman, W.-P. (2009). "AffectButton: Towards a standard for dynamic affective user feedback". In: *Proc. of the 3rd IEEE ACII.* Amsterdam, The Netherlands, s.p.

Broekens, J. & Brinkman, W.-P. (June 2013). "AffectButton: A Method for Reliable and Valid Affective Self-report". *Int J Hum-Comput St* 71 (6), pp. 641–667.

Brown, M. B. & Forsythe, A. B. (1974). "Robust tests for equality of variances". *J Am Stat Assoc* 69 (346), pp. 364–467.

Brown, P. F.; deSouza, P. V.; Mercer, R. L.; Pietra, V. J. D. & Lai, J. C. (Dec. 1992). "Class-based N-gram Models of Natural Language". *Comput Linguist* 18 (4), pp. 467–479.

Bruder, C.; Clemens, C.; Glaser, C. & Karrer-Gauß, K. (2009). *TA-EG – Fragebogen zur Erfassung von Technikaffinität*. Tech. rep. FG Mensch-Maschine Systeme TU Berlin.

Brückl, M. & Sendlmeier, W. (2005). "Junge und alte Stimmen". In: *Stimmlicher Ausdruck in der Alltagskommunikation*. Ed. by Sendlmeier, W. & Bartels, A. Vol. 4. Mündliche Kommunikation. Berlin, Germany: Logos Verlag, pp. 135–163.

Burg, J. P. (1975). "Maximum entropy spectral analysis". PhD thesis. Department of Geophysics, Stanford University.

Burger, S.; MacLaren, V. & Yu, H. (2002). "The ISL meeting corpus: The impact of meeting type on speech style". In: *Proc. of the INTERSPEECH-2002*. Denver, USA, pp. 301–304.

Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W. & Weiss, B. (2005). "A database of German emotional speech". In: *Proc. of the INTERSPEECH-2005*. Lisbon, Portugal, pp. 1517–1520.

Burkhardt, F.; Eckert, M.; Johannsen, W. & Stegmann, J. (2010). "A Database of Age and Gender Annotated Telephone Speech". In: *Proc. of the 7th LREC*. Valletta, Malta, s.p.

Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C. M.; Kazemzadeh, A.; Lee, S.; Neumann, U. & Narayanan, S. (2004). "Analysis of emotion recognition using facial expressions, speech and multimodal information". In: *Proc. of the 6th ACM ICMI*. State College, USA, pp. 205–211.

Butler, L. D. & Nolen-Hoeksema, S. (1994). "Gender differences in responses to depressed mood in a college sample". *Sex Roles* 30 (5-6), pp. 331–346.

Bänziger, T.; Mortillaro, M. & Scherer, K. R. (2012). "Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception". *Emotion* 12 (5), pp. 1161–1179.

Böck, R. (2013). "Multimodal Automatic User Disposition Recognition in Human-Machine Interaction". PhD thesis. Otto von Guericke University Magdeburg.

Böck, R.; Hübner, D. & Wendemuth, A. (2010). "Determining optimal signal features and parameters for HMM-based emotion classification". In: *Proc. of the 15th IEEE MELECON*. Valetta, Malta, pp. 1586–1590.

Böck, R.; Siegert, I.; Vlasenko, B.; Wendemuth, A.; Haase, M. & Lange, J. (2011a). "A Processing Tool for Emotionally Coloured Speech". In: *Proc. of the 2011 IEEE ICME*. Barcelona, Spain, s.p.

Böck, R.; Siegert, I.; Haase, M.; Lange, J. & Wendemuth, A. (2011b). "ikannotate – A Tool for Labelling, Transcription, and Annotation of Emotionally Coloured Speech". In:

Affective Computing and Intelligent Interaction. Ed. by D'Mello, S.; Graesser, A.; Schuller, B. & Martin, J.-C. Vol. 6974. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 25–34.

Böck, R.; Limbrecht, K.; Siegert, I.; Glüge, S.; Walter, S. & Wendemuth, A. (2012a). "Combining Mimic and Prosodic Analyses for User Disposition Classification". In: *Proc. of the 23th ESSV.* Cottbus, Germany, pp. 220–227.

Böck, R.; Limbrecht, K.; Walter, S.; Hrabal, D.; Traue, H. C.; Glüge, S. & Wendemuth, A. (2012b). "Intraindividual and Interindividual Multimodal Emotion Analyses in Human-Machine-Interaction". In: *Proc. of the IEEE CogSIMA.* New Orleans, USA, pp. 59–64.

Böck, R.; Limbrecht-Ecklundt, K.; Siegert, I.; Walter, S. & Wendemuth, A. (2013a). "Audio-Based Pre-classification for Semi-automatic Facial Expression Coding". In: *Human-Computer Interaction. Towards Intelligent and Implicit Interaction.* Ed. by Kurosu, M. Vol. 8008. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 301–309.

Böck, R.; Glüge, S. & Wendemuth, A. (2013b). "Dempster-Shafer Theory with Smoothness". In: *Integrated Uncertainty in Knowledge Modelling and Decision Making.* Ed. by Qin, Z. & Huynh, V.-N. Vol. 8032. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 13–22.

Callejas, Z. & López-Cózar, R. (2005). "Implementing Modular Dialogue Systems: A Case Study". In: *Proc. of the ASIDE 2005.* Aalborg, Denmark, s.p.

– (May 2008). "Influence of contextual information in emotion annotation for spoken dialogue systems". *Speech Commun* 50 (5), pp. 416–433.

Carletta, J. (June 1996). "Assessing agreement on classification tasks: the kappa statistic". *Comput Linguist* 22 (2), pp. 249–254.

Carpenter, S. M.; Peters, E.; Västfjäll, D. & Isen, A. M. (2013). "Positive feelings facilitate working memory and complex decision making among older adults". *Cognition Emotion* 27 (1), pp. 184–192.

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

Carver, C. S. & White., T. L. (1994). "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales". *J Pers Soc Psychol* 67 (2), pp. 319–333.

Cauldwell, R. T. (2000). "Where did the anger go? The role of context in interpreting emotion in speech". In: *Proc. of the SpeechEmotion-2000.* Newcastle, UK, pp. 127–131.

Chasaide, A. N. & Gobl, C. (1993). "Contextual variation of the vowel voice source as a function of adjacent consonants". *Lang Speech* 36, pp. 303–330.

Chateau, N.; Maffiolo, V. & Blouin, C. (2004). "Analysis of emotional speech in voice mail messages: The influence of speakers' gender". In: *Proc. of the INTERSPEECH-2004.* Jeju, Korea, pp. 39–44.

Cicchetti, D. V. & Feinstein, A. R. (June 1990). “High agreement but low kappa: II. Resolving the paradoxes”. *J Clin Epidemiol* 43 (6), pp. 551–558.

Clavel, C.; Vasilescu, I.; Devillers, L.; Ehrette, T. & Richard, G. (2006). “Fear-type emotions of the SAFE corpus: Annotation issues”. In: *Proc. of the 5th LREC.* Genova, Italy, pp. 1099–1104.

Cohen, J. (Apr. 1960). “A coefficient of agreement for nominal scales”. *Educ Psychol Meas* 24 (1), pp. 37–46.

Cohen, J. (Oct. 1968). “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.” *Psychol Bull* 70 (4), pp. 213–220.

Cohen, J.; Kamm, T. & Andreou, A. G. (1995). “Vocal tract normalization in speech recognition: Compensating for systematic speaker variability”. *J Acoust Soc Am* 97 (5), pp. 3246–3247.

Colombetti, G. (2009). “From affect programs to dynamical discrete emotions”. *Philo Psychol* 22 (4), pp. 407–425.

Corley, M. & Stewart, O. W. (2008). “Hesitation Disfluencies in Spontaneous Speech: The Meaning of *um*”. *Language and Linguistics Compass* 2 (4), pp. 589–602.

Costa, P. T. & McCrae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, USA: Psychological Assessment Resources.

– (1992). *NEO-PI-R Professional manual. Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI).* Odessa, USA: Psychological Assessment Resources.

– (1995). “Domains and Facets: Hierarchical Personality Assessment Using the Revised NEO Personality Inventory”. *J Pers Assess* 64 (1), pp. 21–50.

Cotton, J. C. (1936). “Syllabic rate: A new concept in the study of speech rate variation”. *Commun Monogr* 3, pp. 112–117.

Cowie, R. & Cornelius, R. R. (2003). “Describing the emotional states that are expressed in speech”. *Speech Commun* 40 (1-2), pp. 5–32.

Cowie, R.; Douglas-Cowie, E.; Savvidou, S.; McMahon, E.; Sawey, M. & Schröder, M. (2000). “FEELTRACE: An Instrument for Recording Perceived Emotion in Real Time”. In: *Proc. of the SpeechEmotion-2000.* Newcastle, UK, pp. 19–24.

Cowie, R. & McKeown, G. (2010). *Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme.* Tech. rep. SEMAINE deliverable D6b.

Crawford, J. R. & Henry, J. D. (Sept. 2004). “The positive and negative affect schedule (PANAS): construct validity, measurement properties and normative data in a large non-clinical sample”. *Brit J Clin Psychol* 43 (3), pp. 245–265.

Cronbach, L. J. (1951). “Coefficient alpha and the internal structure of tests”. *Psychometrika* 16 (3), pp. 297–334.

Cullen, A. & Harte, N. (2012). “Feature sets for automatic classification of dimensional affect”. In: *Proc. of the 23nd IET Irish Signals and Systems Conference.* Maynooth, Ireland, pp. 1–6.

Cuperman, R. & Ickes, W. (2009). “Big Five Predictors of Behavior and Perceptions in Initial Dyadic Interactions: Personality Similarity Helps Extraverts and Introverts, but Hurts ’Disagreeables’”. *J Pers Soc Psychol* 97 (4), pp. 667–684.

Cutler, A. & Clifton, C. E. (1985). “The use of prosodic information in word recognition”. In: *Attention and performance.* Ed. by Bouma, H. & Bowhuis, D. G. Vol. 10. Hillsdale, USA: Erlbaum, pp. 183–196.

Cutler, A.; Ladd, D. R. & Brown, G. (1983). *Prosody, models and measurements.* Heidelberg, Berlin, Germany: Springer.

Daily, J. A. (2002). “Personality and Interpersonal Communication”. In: *Handbook of Interpersonal Communication.* Ed. by Knapp, M. L. & Daily, J. A. Thousand Oaks, USA: Sage, pp. 133–180.

Darwin, C. (1874). *The Descent of Man, and Selection in Relation to Sex.* 2nd ed. London,UK: John Murray.

Davies, M. & Fleiss, J. L. (Dec. 1982). “Measuring Agreement for Multinomial Data”. *Biometrics* 38 (4), pp. 1047–1051.

Davis, K. H.; Biddulph, R. & Balashek, S. (Nov. 1952). “Automatic Recognition of Spoken Digits”. *J Acoust Soc Am* 24 (6), pp. 637–642.

Davis, S. & Mermelstein, P. (1980). “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. *IEEE Trans. Acoust., Speech, Signal Process.* 28 (4), pp. 357–366.

de Boer, B. (2000). “Self-organization in vowel systems”. *J Phonetics* 28 (4), pp. 441–465.

Dellwo, V.; Leemann, A. & Kolly, M.-J. (2012). “Speaker idiosyncratic rhythmic features in the speech signal”. In: *Proc. of the INTERSPEECH-2012.* Portland, USA, s.p.

Dempster, A. P.; Laird, N. M. & Rubin, D. B. (1977). “Maximum likelihood from incomplete data via the EM algorithm”. *J Roy Stat Soc B* 39 (1), pp. 1–38.

Denes, P. (Apr. 1959). “The design and operation of the mechanical speech recognizer at University College London”. *J Brit I R E* 19 (4), pp. 219–229.

Desmet, P. M. A.; Porcelijn, R. & Dijk, M. B. (2007). “Emotional Design. Application of a Research-Based Design Approach”. *Knowledge, Technology & Policy* 20 (3), pp. 141–155.

Devillers, L. & Vasilescu, I. (2004). “Reliability of lexical and prosodic cues in two real-life spoken dialog corpora”. In: *Proc. of the 4th LREC.* Lisbon, Portugal, pp. 865–868.

Devillers, L. & Vidrascu, I. (2006). “Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs”. In: *Proc. of the INTERSPEECH-2006.* Pittsburgh, USA, pp. 801–804.

Devillers, L.; Cowie, R.; Martin, J.; Douglas-Cowie, E.; Abrilian, S. & McRorie, M. (2006). “Real life emotions in French and English TV video clips: An integrated annotation protocol combining continous and discrete approaches”. In: *Proc. of the 5th LREC.* Genova, Italy, pp. 1105–1110.

Diamantidis, N. A.; Karlis, D. & Giakoumakis, E. A. (Jan. 2000). “Unsupervised Stratification of Cross-validation for Accuracy Estimation”. *Artif Intell* 116 (1-2), pp. 1–16.

Dobrišek, S.; Gajšek, R.; Mihelič, F.; Pavešić, N. & Štruc, V. (2013). “Towards Efficient Multi-Modal Emotion Recognition”. *Int J Adv Robot Syst* 10 (53), s.p.

Doost, H. V.; Akbari, M.; Charsted, P. & Akbari, J. A. (2013). “The Role of Psychological Traits in Market Mavensim Using Big Five Model”. *J Basic Appl Sci Res* 3 (2), pp. 744–751.

Douglas-Cowie, E.; Cowie, R. & Schröder, M. (2000). “A New Emotion Database: Considerations, Sources and Scope”. In: *Proc. of the SpeechEmotion-2000.* Newcastle, UK, pp. 39–44.

Douglas-Cowie, E.; Devillers, L.; Martin, J.-C.; Cowie, R.; Savvidou, S.; Abrilian, S. & Cox, C. (2005). “Multimodal databases of everyday emotion: facing up to complexity”. In: *Proc. of the INTERSPEECH-2005.* Lisbon, Portugal, pp. 813–816.

Dumouchel, P.; Dehak, N.; Attabi, Y.; Dehak, R. & Boufaden, N. (2009). “Cepstral and long-term features for emotion recognition”. In: *Proc. of the INTERSPEECH-2009.* Brighton, UK, pp. 344–347.

Ekman, P. (1992). “Are there basic emotions?” *Psychol Rev* 99, pp. 550–553.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement.* Palo Alto, USA: Consulting Psychologists Press.

Ekman, P. (2005). “Basic Emotions”. In: *Handbook of Cognition and Emotion.* Hoboken, USA: John Wiley & Sons, pp. 45–60.

Elsholz, J.-P.; Melo, G. de; Hermann, M. & Weber, M. (2009). “Designing an extensible architecture for Personalized Ambient Information”. *Pervasive and Mobile Computing* 5 (5), pp. 592–605.

Emori, T. & Shinoda, K. (2001). “Rapid vocal tract length normalization using maximum likelihood estimation”. In: *Proc. of the INTERSPEECH-2001.* Aalborg, Denmark, pp. 1649–1652.

Engberg, I. S. & Hansen, A. V. (1996). *Documentation of the danish emotional speech database (DES).* Tech. rep. Internal aau report. Denmark: Center for Person, Kommunikation, Aalborg University.

Eppinger, B. & Herter, E. (1993). *Sprachverarbeitung.* Munich, Germany: Carl-Hanser-Verlag.

Esposito, A. & Riviello, M. T. (2010). “The New Italian Audio and Video Emotional Database”. In: *Development of Multimodal Interfaces: Active Listening and Synchrony.*

Ed. by Esposito, A.; Campbell, N.; Vogel, C.; Hussain, A. & Nijholt, A. Vol. 5967. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 406–422.

Eyben, F.; Wöllmer, M. & Schuller, B. (2010). “openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor”. In: *Proc. of the ACM MM-2010.* Firenze, Italy, s.p.

Fahy, F. (2002). *Sound Intensity.* 2nd ed. New York, USA: Taylor & Francis.

Fant, G. (1960). *The Acoustic Theory of Speech Production.* Description and analysis of contemporary standard Russian. Hague, The Netherlands: Mouton & Co.

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

Feinstein, A. R. & Cicchetti, D. V. (June 1990). “High agreement but low kappa: I. The problems of two paradoxes”. *J Clin Epidemiol* 43 (6), pp. 543–549.

Felzenszwalb, P. F. & Huttenlocher, D. P. (Jan. 2005). “Pictorial Structures for Object Recognition”. *Int J Comput Vision* 61 (1), pp. 55–79.

Fernandez, R. & Picard, R. W. (2003). “Modeling drivers’ speech under stress”. *Speech Commun* 40 (1-2), pp. 145–159.

Fischer, K.; Wrede, B.; Brindöpke, C. & Johanntokrax, M. (1996). “Quantitative und funktionale Analysen von Diskurspartikeln im Computer Talk (Quantitative and functional analyzes of discourse particles in Computer Talk)”. *International Journal for Language Data Processing* 20 (1-2), pp. 85–100.

Fleiss, J. L. (Nov. 1971). “Measuring nominal scale agreement among many raters”. *Psychol Bull* 76 (5), pp. 378–382.

Fleiss, J. L.; Levin, B. & Paik, M. C. (2003). *Statistical Methods for Rates & Proportions.* 3rd ed. Hoboken, USA: John Wiley & Sons.

Flew, A. (ed.). *A Dictionary of Philosophy.* London, UK: Pan Books.

Forgas, J. P (2002). “Feeling and doing: Affective influences on interpersonal behavior”. *Psychol Inq* 13 (1), pp. 1–28.

Fox, A. (2000). *Prosodic Features and Prosodic Structure : The Phonology of ’Suprasegmentals’.* Oxford,UK: Oxford University Press.

Fragopanagos, N. F. & Taylor, J. G. (Nov. 2005). “Emotion recognition in human-computer interaction”. *Neural Networks* 18 (4), pp. 389–405.

Frijda, N. H. (1969). “Recognition of emotion”. *Adv Exp Soc Psychol* 4, pp. 167–223.

– (1986). *The emotions.* Cambridge, UK: Cambridge University Press.

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

Frommer, J.; Michaelis, B.; Rösner, D.; Wendemuth, A.; Friesen, R.; Haase, M.; Kunze, M.; Andrich, R.; Lange, J.; Panning, A. & Siegert, I. (2012b). "Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus". In: *Proc. of the 8th LREC*. Istanbul, Turkey, pp. 3064–3069.

Fu, K. S. (1982). *Syntactic pattern recognition and applications*. Englewood Cliffs, USA: Prentice-Hall.

Funder, D. C. & Sneed, C. D. (1993). "Behavioral manifestations of personality: An ecological approach to judgmental accuracy". *J Pers Soc Psychol* 64 (3), pp. 479–490.

Gajšek, R.; Štruc, V.; Vesnicer, B.; Podlesek, A.; Komidar, L. & Mihelič, F. (2009). "Analysis and Assessment of AvID: Multi-Modal Emotional Database". In: *Text, Speech and Dialogue*. Ed. by Matoušek, V. & Mautner, P. Vol. 5729. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 266–273.

Galton, F. (1892). *Finger Prints*. London, UK: Macmillan. Facsimile from: http://galton.org/books/finger-prints.

Gebhard, P. (2005). "ALMA A Layered Model of Affect". In: *Proc. of the 4th ACM AAMAS*. Utrecht, The Netherlands, pp. 29–36.

Gehm, T. & Scherer, K. R. (1988). "Factors determining the dimensions of subjective emotional space". In: *Facets of emotion: Recent research*. Ed. by Scherer, K. R. Hillsdale, USA: Lawrence Erlbaum, pp. 99–114.

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

Gharavian, D.; Sheikhan, M. & Ashoftedel, F. (2013). "Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model". *Neural Comput Appl* 22 (6), pp. 1181–1191.

Giuliani, D. & Gerosa, M. (2003). "Investigating recognition of children's speech". In: *Proc. of the IEEE ICASSP-2003*. Vol. 2. Hong Kong, pp. 137–140.

Glodek, M.; Schels, M.; Palm, G. & Schwenker, F. (2012). "Multi-Modal Fusion Based on Classification Using Rejection Option and Markov Fusion Network". In: *Proc. of the 21st IEEE ICPR*. Tsukuba, Japan, pp. 1084–1087.

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

Glüge, S.; Böck, R. & Wendemuth, A. (2011). "Segmented-Memory Recurrent Neural Networks versus Hidden Markov Models in Emotion Recognition from Speech". In: *Proc. of the 3rd IJCCI*. Paris, France, pp. 308–315.

Gnjatović, M. & Rösner, D. (2008). "On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System". In: *Proc. of the 7th LREC*. Marrakech, Morocco, s.p.

Gold, B. & Morgan, M. (2000). *Speech and Audio Signal Processing. Processing and Perception of Speech and Music.* Hoboken, USA: John Wiley & Sons.

Goldberg, L. R. (1981). “Language and individual differences: The search for universals in personality lexicons”. In: *Review of Personality and Social Psychology.* Ed. by Wheeler, L. Vol. 2. Beverly Hills, USA: Sage, pp. 141–165.

Gosztolya, G.; Busa-Fekete, R. & Tóth, L. (2013). “Detecting autism, emotions and social signals using adaboost”. In: *Proc. of the INTERSPEECH-2013.* Lyon, France, pp. 220–224.

Grandjean, D.; Sander, D. & Scherer, K. R. (Apr. 2008). “Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization”. *Conscious Cogn* 17 (2), pp. 484–495.

Grimm, M. & Kroschel, K. (2005). “Evaluation of natural emotions using self assessment manikins”. In: *Proc. of the IEEE ASRU.* Cancún, Mexico, pp. 381–385.

Grimm, M.; Kroschel, K. & Narayanan, S. (2008). “The Vera am Mittag German Audio-Visual Emotional Speech Database”. In: *Proc. of the 2008 IEEE ICME.* Hannover, Germany, pp. 865–868.

Grimm, M.; Kroschel, K.; Mower, E. & Narayanan, S. (2007). “Primitives-based evaluation and estimation of emotions in speech”. *Speech Commun* 49 (10-11), pp. 787–800.

Gross, J. J. & John, O. P. (2003). “Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being”. *J Pers Soc Psychol* 85 (2), pp. 348–362.

Gross, J. J.; Carstensen, L. L.; Pasupathi, M.; Tsai, J.; Skorpen, C. G. & Hsu, A. Y. (1997). “Emotion and aging: experience, expression, and control.” *Psychol Aging* 12 (4), pp. 590–599.

Gwet, K. L. (2008a). “Intrarater Reliability”. In: *Wiley Encyclopedia of Clinical Trials.* Ed. by D’Agostino, R. B.; Sullivan, L. & Massaro, J. Hoboken, USA: John Wiley & Sons, pp. 473–485.

Gwet, K. L. (2008b). “Computing inter-rater reliability and its variance in the presence of high agreement”. *Brit J Math Stat Psy* 61 (1), pp. 29–48.

Haji, T.; Horiguchi, S.; Baer, T. & Gould, W. J. (1986). “Frequency and amplitude perturbation analysis of electroglottograph during sustained phonation”. *J Acoust Soc Am* 80 (1), pp. 58–62.

Harrington, J.; Palethorpe, S. & Watson, C. I. (2007). “Age-related changes in fundamental frequency and formants : a longitudinal study of four speakers”. In: *Proc. of the INTERSPEECH-2007.* Vol. 2. Antwerp, Belgium, pp. 1081–1084.

Hassan, A; Damper, R. & Niranjan, M. (July 2013). “On Acoustic Emotion Recognition: Compensating for Covariate Shift”. *IEEE Trans. Audio, Speech, Language Process.* 21 (7), pp. 1458–1468.

Hassenzahl, M.; Burmester, M. & Koller, F. (2003). “AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität”. In: *Mensch & Computer 2003*. Ed. by Szwillus, G. & Ziegler, J. Vol. 57. Berichte des German Chapter of the ACM. Wiesbaden, Germany: Vieweg+Teubner, pp. 187–196.

Hawk, S. T.; Kleef van, G. A.; Fischer, A. H. & Schalk van der, J. (June 2009). “’Worth a thousand words’: absolute and relative decoding of nonlinguistic affect vocalizations”. *Emotion* 9 (3), pp. 293–305.

Hayes, A. F. & Krippendorff, K. (Dec. 2007). “Answering the Call for a Standard Reliability Measure for Coding Data”. *Communication Methods and Measures* 24 (1), pp. 77–89.

Hermansky, H. (2011). “Speech recognition from spectral dynamics”. *Sadhana* 36 (5), pp. 729–744.

Hermansky, H.; Morgan, N.; Bayya, A. & Kohn, P. (1992). “RASTA-PLP speech analysis technique”. In: *Proc. of the IEEE ICASSP-1992*. Vol. 1. San Francisco, USA, pp. 121–124.

Hermansky, H. (1990). “Perceptual linear predictive (PLP) Analysis of speech”. *J Acoust Soc Am* 87 (4), pp. 1738–1752.

Hermansky, H. & Morgan, N. (1994). “RASTA processing of speech”. *IEEE Speech Audio Process.* 2 (4), pp. 578–589.

Ho, C.-H. (2001). “Speaker Modelling for Voice Conversion”. PhD thesis. London: Brunel University.

Hollien, H. & Shipp, T. (1972). “Speaking Fundamental Frequency and Chronologic Age in Males”. *Journal Speech Hear Res* 15, pp. 155–159.

Holsti, O. R. (1969). *Content analysis for the social sciences and humanitities.* Reading, USA: Addison-Wesley.

Honda, K. (2008). “Physiological Processes of Speech Production”. In: *Springer Handbook of Speech Processing.* Ed. by Benesty, J.; Sondhi, M. M. & Huang, Y. Berlin, Heidelberg, Germany: Springer.

Horowitz, L. M.; Strauß, B. & Kordy, H. (2000). *Inventar zur Erfassung interpersonaler Probleme (IIPD).* 2nd ed. Weinheim, Germany: Beltz.

Hrabal, D.; Kohrs, C.; Brechmann, A.; Tan, J.-W.; Rukavina, S. & Traue, H. C. (2013). “Physiological effects of delayed system response time on skin conductance”. In: *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction.* Ed. by Schwenker, F.; Scherer, S. & Morency, L.-P. Vol. 7742. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 52–62.

Huang, D.-Y.; Ge, S. S. & Zhang, Z. (2011). “Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines”. In: *Proc. of the INTERSPEECH-2011.* Florence, Italy. Chap. 15, pp. 3301–3304.

Hubeika, V. (2006). “Estimation of Gender and Age from Recorded Speech”. In: *Proc. of the ACM Student Research competition.* Prague, Czech Republic, pp. 25–32.

Hussain, M. S.; Calvo, R. A. & Aghaei Pour, P. (2011). “Hybrid Fusion Approach for Detecting Affects from Multichannel Physiology”. In: *Affective Computing and Intelligent Interaction.* Ed. by D’Mello, S.; Graesser, A.; Schuller, B. & Martin, J.-C. Vol. 6974. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 568–577.

Ibáñez, J. (Jan. 2011). “Showing emotions through movement and symmetry”. *Comput Hum Behav* 27 (1), pp. 561–567.

Ivanov, A. & Chen, X. (2012). “Modulation spectrum analysis for speaker personality trait recognition”. In: *Proc. of the INTERSPEECH-2012.* Portland, USA, pp. 278–281.

Iwarsson, J. & Sundberg, J. (1998). “Effects of lung volume on vertical larynx position during phonation”. *J Voice* 12 (2), pp. 159–165.

Izard, C. E.; Libero, D. Z.; Putnam, P. & Haynes, O. M. (May 1993). “Stability of emotion experiences and their relations to traits of personality”. *J Pers Soc Psychol* 64 (5), pp. 847–860.

Jahnke, W.; Erdmann, G. & Kallus, K. (2002). *Stressverarbeitungsfragebogen mit SVF 120 und SVF 78.* 3rd ed. Göttingen, Germany: Hogrefe.

Jain, A. K.; Duin, R. P. W. & Mao, J. (2000). “Statistical pattern recognition: a review”. *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (1), pp. 4–37.

Jelinek, F.; Bahl, L. R.; & Mercer, R. L. (May 1975). “Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech”. *IEEE Trans. Inf. Theory* 21 (3), pp. 250–256.

Jeon, J. H.; Xia, R. & Liu, Y. (2010). “Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence”. In: *Proc. of the INTERSPEECH-2010*, pp. 2802–2805.

John, O. P.; Hampson, S. E. & Goldberg, L. R. (1991). “Is there a basic level of personality description?” *J Pers Soc Psychol* 60 (3), pp. 348–361.

Johnstone, T.; Reekum van, C. M. & Scherer, K. R. (2001). “Vocal Expression Correlates of Appraisal Processes”. In: *Appraisal Processes in Emotion: Theory, Methods, Research.* Ed. by Scherer, K. R.; Schorr, A. & Johnstone, T. Oxford, UK: Oxford University Press, pp. 271–284.

Juang, B.-H. & Rabiner, L. R. (2006). “Speech Recognition, Automatic: History”. In: *Encyclopedia of Language & Linguistics.* Ed. by Brown, K. 2nd ed. Oxford, UK: Elsevier, pp. 806–819.

Juslin, P. N & Scherer, K. R. (2005). “Vocal expression of affect”. In: *The new handbook of methods in nonverbal behavior research.* Ed. by Harrigan J. A. Rosenthal R., S. K. R. New York, USA: Oxford University Press, pp. 66–135.

Jähne, B. (1995). *Digital Image Processing.* 6th ed. Berlin, Heidelberg, Germany: Springer.

Kaiser, S. & Wehrle, T. (2001). “Facial Expressions as Indicator of Appraisal Processes”. In: *Appraisal Processes in Emotion: Theory, Methods, Research.* Ed. by Scherer, K. R.; Schorr, A. & Johnstone, T. Oxford, UK: Oxford University Press, pp. 285–305.

Kameas, A. D.; Goumopoulos, C.; Hagras, H.; Callaghan, V.; Heinroth, T. & Weber, M. (2009). “An Architecture That Supports Task-Centered Adaptation In Intelligent Environments”. In: *Advanced Intelligent Environments.* Ed. by Kameas, A. D.; Callagan, V.; Hagras, H.; Weber, M. & Minker, W. Berlin Heidelberg, Germany: Springer, pp. 41–66.

Kane, J.; Scherer, S.; Aylett, M. P.; Morency, L.-P. & Gobl, C. (2013). “Speaker and language independent voice quality classification applied to unlabelled corpora of expressive speech”. In: *Proc. of the IEEE ICASSP-2013.* Vancouver, Canada, pp. 7982–7986.

Kehrein, R. & Rabanus, S. (2001). “Ein Modell zur funktionalen Beschreibung von Diskurspartikeln (A Model for the functional description of discourse particles)”. In: *Neue Wege der Intonationsforschung.* Vol. 157-158. Germanistische Linguistik. Hildesheim, Germany: Georg Olms Verlag, pp. 33–50.

Kelly, F. & Harte, N. (2011). “Effects of Long-Term Ageing on Speaker Verification”. In: *Biometrics and ID Management.* Ed. by Vielhauer, C.; Dittmann, J.; Drygajlo, A.; Juul, N. & Fairhurst, M. Vol. 6583. LNCS. Berlin Heidelberg, Germany: Springer, pp. 113–124.

Khan, A. & Rayner, G. D. (2003). “Robustness to non-normality of common tests for the many-sample location problem”. *J Appl Math Decis Sci* 7 (4), pp. 187–206.

Kim, J. (2007). “Bimodal emotion recognition using speech and physiological changes”. In: *Robust Speech Rcognition and Understanding.* Ed. by Grimm, M. & Kroschel, K. Vienna, Austria: I-Tech Education and Publishing, pp. 265–280.

Kim, J.-B.; Park, J.-S. & Oh, Y.-H. (2012a). “Speaker-Characterized Emotion Recognition using Online and Iterative Speaker Adaptation”. *Cognitive Computation* 4 (4), pp. 398–408.

Kim, J.; Kumar, N.; Tsiartas, A.; Li, M. & Narayanan, S. S. (2012b). “Intelligibility Classification of Pathological Speech Using Fusion of Multiple Subsystems”. In: *Proc. of the INTERSPEECH-2012.* Portland, USA, pp. 534–537.

Kinnunen, T. & Li, H. (2010). “An overview of text-independent speaker recognition: From features to supervectors”. *Speech Commun* 52 (1), pp. 12–40.

Kipp, M. (2001). “Anvil - A Generic Annotation Tool for Multimodal Dialogue”. In: *Proc. of the INTERSPEECH-2001.* Aalborg, Denmark, pp. 1367–1370.

Knox, M. & Mirghafori, M. (2007). “Automatic laughter detection using neural networks”. In: *Proc. of the INTERSPEECH-2007.* Antwerp, Belgium, pp. 2973–2976.

Kockmann, M.; Burget, L. & Černocký, J. (2009). “Brno University of Technology System for Interspeech 2009 Emotion Challenge”. In: *Proc. of the INTERSPEECH-2009.* Brighton, GB, pp. 348–351.

Kockmann, M.; Burget, L. & Černocký, J. H. (2011). "Application of speaker- and language identification state-of-the-art techniques for emotion recognition". *Speech Commun* 53 (9-10), pp. 1172–1185.

Koelstra, S.; Mühl, C. & Patras, I. (2009). "EEG analysis for implicit tagging of video data". In: *Proc. of the 3rd IEEE ACII.* Amsterdam, The Netherlands, pp. 27–32.

Kohavi, R. (1995). "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection". In: *Proc. of the 14th IJCAI.* Vol. 2. Montréal, Canada, pp. 1137–1143.

Kopp, S.; Allwood, J.; Grammer, K.; Ahlsen, E. & Stocksmeier, T. (2008). "Modeling Embodied Feedback with Virtual Humans". In: *Modeling Communication with Robots and Virtual Humans.* Ed. by Wachsmuth, I. & Knoblich, G. Vol. 4930. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 18–37.

Kotzyba, M.; Deml, B.; Neumann, H.; Glüge, S.; Hartmann, K.; Siegert, I.; Wendemuth, A.; Traue, H. C. & Walter, S. (2012). "Emotion Detection by Event Evaluation using Fuzzy Sets as Appraisal Variables". In: *Proc. of the 11th ICCM.* Berlin, Germany, pp. 123–124.

Kraemer, H. C. (Dec. 1979). "Ramifications of a population model for κ as a coefficient of reliability". *Psychometrika* 44 (4), pp. 461–472.

– (June 1980). "Extension of the Kappa Coefficient". *Biometrics* 36 (2), pp. 207–216.

– (2008). "Interrater Reliability". In: *Wiley Encyclopedia of Clinical Trials.* Ed. by D'Agostino, R. B.; Sullivan, L. & Massaro, J. Hoboken, USA: John Wiley & Sons.

Krell, G.; Glodek, M.; Panning, A.; Siegert, I.; Michaelis, B.; Wendemuth, A. & Schwenker, F. (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.; Scherer, S. & Morency, L.-P. Vol. 7742. LNAI. Berlin, Heidelberg, Germany: Springer, pp. 116–130.

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

Kruskal, W. & Wallis, W. A. (1952). "Use of Ranks in One-Criterion Variance Analysis". *J Am Stat Assoc* 47 (260), pp. 583–621.

Kulkarni, D. & Simon, H. (1988). "The processes of scientific discovery: The strategy of experimentation". *Cognitive Sci* 12, pp. 139–175.

Ladd, R. D. (1996). "Intonational Phonology". In: *Studies in Linguistics.* Vol. 79. Cambridge, UK: Cambridge University Press.

Landis, J. R. & Koch, G. G. (Mar. 1977). "The measurement of observer agreement for categorical data". *Biometrics* 33 (1), pp. 159–174.

Lang, P. J. (1980). "Behavioral treatment and bio-behavioral assessment: Computer applications". In: *Technology in Mental Health Care Delivery Systems.* Ed. by Sidowski, J. B.; Johnson, J. H. & Williams, T. A. New York, USA: Ablex Publishing, pp. 119–137.

Lange, J. & Frommer, J. (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.

Larsen, R. J. & Fredrickson, B. L. (1999). “Measurement Issues in Emotion Research”. In: *Well-being: Foundations of hedonic psychology.* Ed. by Kahneman, D.; Diener, E. & Schwarz, N. New York, USA: Russell Sage Foundation, pp. 40–60.

Larsen, R. J. & Ketelaar, T. (July 1991). “Personality and susceptibility to positive and negative emotional states”. *J Pers Soc Psychol* 61 (1), pp. 132–140.

Lee, C.-C.; Mower, E.; Busso, C.; Lee, S. & Narayanan, S. (2009). “Emotion Recognition Using a Hierarchical Binary Decision Tree Approach”. In: *Proc. of the INTERSPEECH-2009.* Brighton, GB, pp. 320–323.

Lee, C. M. & Narayanan, S. S. (Mar. 2005). “Toward detecting emotions in spoken dialogs”. *IEEE Trans. Speech Audio Process.* 13 (2), pp. 293–303.

Lee, L. & Rose, R. (1998). “A frequency warping approach to speaker normalization”. *IEEE Speech Audio Process.* 6 (1), pp. 49–60.

Lee, L. & Rose, R. C. (1996). “Speaker normalization using efficient frequency warping procedures”. In: *Proc. of the IEEE ICASSP-1996.* Vol. 1. Atlanta, USA, pp. 353–356.

Lee, M.-W. & Kwak, K.-C. (Dec. 2012). “Performance Comparison of Gender and Age Group Recognition for Human-Robot Interaction”. *International Journal of Advanced Computer Science and Application* 3 (12), pp. 207–211.

Lee, S.; Potamianos, A. & Narayanan, S. (1997). “Analysis of children’s speech: Duration, pitch and formants”. In: *Proc. of the EUROSPEECH-1997.* Vol. 1. Rhodes, Greece, pp. 473–476.

Lefter, I.; Rothkrantz, L. J. M. & Burghouts, G. (2012). “Aggression detection in speech using sensor and semantic information”. In: *Text, Speech and Dialogue.* Ed. by Sojka, P.; Horák, A.; Kopeček, I. & Pala, K. Vol. 7499. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 665–672.

Levinson, S. E.; Rabiner, L. R. & Sondhi, M. M. (Apr. 1983). “An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition”. *Bell Syst. Tech. J.* 62 (4), pp. 1035–1074.

Levy, D.; Catizone, R.; Battacharia, B.; Krotov, A. & Wilks, Y. (1997). “CONVERSE: a conversational companion”. In: *Proc. 1st. Int. Workshop on Human-Computer Conversation.* Bellagio, Italy, s.p.

Li, M.; Jung, C.-S. & Han, K. J. (2010). “Combining Five Acoustic Level Modeling Methods for Automatic Speaker Age and Gender Recognition”. In: *Proc. of the INTERSPEECH-2010.* Makuhari, Japan, pp. 2826–2829.

Li, X.; Tao, J.; Johnson, M. T.; Soltis, J.; Savage, A.; Leong, K. M. & Newman, J. D. (2007). "Stress and Emotion Classification using Jitter and Shimmer Features". In: *Proc. of the IEEE ICASSP-2007.* Vol. 4. Honolulu, USA, pp. 1081–1084.

Linville, S. E. (2001). *Vocal Aging.* San Diego, USA: Singular Publishing Group.

Lipovčan, L. K.; Prizmić, Z. & Franc, R. (2009). "Age and Gender Differences in Affect Regulation Strategies". *Drustvena istrazivanja: Journal for General Social Issues* 18 (6), pp. 1075–1088.

Liu, Y.; Shriberg, E.; Stolcke, A. & Harper, M. (2005). "Comparing HMM, maximum entropy, and conditional random fields for disfluency detection". In: *Proc. of the INTERSPEECH-2005.* Lisbon, Portugal, pp. 3033–3036.

MacWhinney, B. (2000). *The CHILDES project: tools for analyzing talk.* 2nd ed. Mahwah, USA: Lawrence Erlbaum.

Markel, J. D. (1972). "The SIFT algorithm for Fundamental Frequency estimation". *IEEE Trans. Audio Electroacoust.* 20 (5), pp. 367–377.

Marsella, S. C. & Gratch, J. (2009). "EMA: A process model of appraisal dynamics". *Cognitive Systems Research* 10 (1), pp. 70–90.

Marti, R.; Heute, U. & Antweiler, C. (2008). *Advances in Digital Speech Transmission.* Hoboken, USA: John Wiley & Sons.

Martin, O.; Kotsia, I.; Macq, B. & Pitas, I. (2006). "The eNTERFACE'05 Audio-Visual Emotion Database". In: *Proc. of the 22nd IEEE ICDE Workshops.* Atlanta,USA, s.p.

Mauss, I. B. & Robinson, M. D. (2009). "Measures of emotion: A review". *Cognition Emotion* 23 (2), pp. 209–237.

McDougall, W. (1908). *An introduction to Social Psychology.* 2nd ed. London, UK: Methuen & Co.

McKeown, G.; Valstar, M. F.; Cowie, R. & Pantic, M. (2010). "The SEMAINE corpus of emotionally coloured character interactions". In: *Proc. of the 2010 IEEE ICME.* Singapore, pp. 1079–1084.

McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M. & Schroder, M. (2012). "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent". *IEEE Trans. Affect. Comput.* 3 (1), pp. 5–17.

McLennan, C. T.; Luce, P. A. & Charles-Luce, J. (2003). "Representation of lexical form". *J Exp Psychol Learn* 29 (4), pp. 539–553.

McRae, K.; Ochsner, K. N.; Mauss, I. B.; Gabrieli, J. J. D. & Gross, J. J. (2008). "Gender Differences in Emotion Regulation: An fMRI Study of Cognitive Reappraisal". *Group Processes & Intergroup Relations* 11 (2), pp. 143–162.

Mehrabian, A. (Oct. 1970). "A semantic space for nonverbal behavior". *J Consult Clin Psych* 35 (2), pp. 248–257.

Mehrabian, A. (1996). “Analysis of the Big-five Personality Factors in Terms of the PAD Temperament Model”. *Aust J Psychol* 48 (2), pp. 86–92.

Mehrabian, A. & Russell, J. A. (Sept. 1977). “Evidence for a three-factor theory of emotions”. *J Res Pers* 11 (3), pp. 273–294.

Mehrabian, A. & Russell, J. A. (1974). *An Approach to Environmental Psychology.* Cambridge, USA: MIT Press.

Meinedo, H. & Trancoso, I. (Aug. 2011). “Age and gender detection in the I-DASH project”. *ACM Trans. Speech Lang. Process.* 7 (4), pp. 1–16.

Meng, H. & Bianchi-Berthouze, N. (2011). “Naturalistic Affective Expression Classification by a Multi-stage Approach Based on Hidden Markov Models”. In: *Affective Computing and Intelligent Interaction.* Ed. by D’Mello, S.; Graesser, A.; Schuller, B. & Martin, J.-C. Vol. 6975. LNCS. Berlin Heidelberg, Germany: Springer, pp. 378–387.

Meng, H.; Huang, D.; Wang, H.; Yang, H.; AI-Shuraifi, M. & Wang, Y. (2013). “Depression Recognition Based on Dynamic Facial and Vocal Expression Features Using Partial Least Square Regression”. In: *Proc. of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge.* Barcelona, Spain, pp. 21–30.

Mengistu, K. T. (2009). “Robust Acoustic and Semantic Modeling in a Telephone-based Spoken Dialog System”. PhD thesis. Otto von Guericke University Magdeburg.

Meudt, S.; Bigalke, L. & Schwenker, F. (2012). “ATLAS – an annotation tool for HCI data utilizing machine learning methods”. In: *Proc. of the 1st APD.* San Fransisco, USA, pp. 5347–5352.

Michaelis, D.; Fröhlich, M.; Strube, H. W.; Kruse, E.; Story, B. & Titze, I. R. (1998). “Some simulations concerning jitter and shimmer measurement”. In: *Proc. of the 3rd International Workshop on Advances in Quantitative Laryngoscopy.* Aachen, Germany, pp. 71–80.

Montacié, C. & Caraty, M.-J. (2012). “Pitch and Intonation Contribution to Speakers’ Traits Classification”. In: *Proc. of the INTERSPEECH-2012.* Portland, USA, pp. 526–529.

Morris, J. D. (1995). “SAM: the self-assessment manikin an efficient cross-cultural measurement of emotional response”. *J Advertising Res* 35 (6), pp. 63–68.

Morris, J. D. & McMullen, J. S. (1994). “Measuring Multiple Emotional Responses to a Single Television Commercial”. *Adv Consum Res* 21, pp. 175–180.

Morris, W. N. (1989). *Mood: the frame of mind.* New York, USA: Springer.

Mower, E.; Metallinou, A.; Lee, C.; Kazemzadeh, A.; Busso, C.; Lee, S. & Narayanan, S. (2009). “Interpreting ambiguous emotional expressions”. In: *Proc. of the 3rd IEEE ACII.* Amsterdam, The Netherlands, s.p.

Mozziconacci, S. J. L. & Hermes, D. J. (2000). “Expression Of Emotion And Attitude Through Temporal Speech Variations”. In: *Proc. of the INTERSPEECH-2000.* Vol. 2. Beijing, China, pp. 373–378.

Mporas, I.; Ganchev, T.; Kotinas, I. & Fakotakis, N. (2007). “Examining the Influence of Speech Frame Size and Issue of Cepstral Coefficients on the Speech Recognition Performance”. In: *Proc. of the 12th SpeCom-2007.* Moscow, Russia, pp. 134–139.

Murray, I. R. & Arnott, J. L. (1993). “Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion”. *J Acoust Soc Am* 93 (2), pp. 1097–1108.

Nadler, R. T.; Rabi, R. & Minda, J. P. (Dec. 2010). “Better mood and better performance. Learning rule-described categories is enhanced by positive mood”. *Psychol Sci* 21 (12), pp. 1770–1776.

Navas, E.; Castelruiz, A.; Luengo, I.; Sánchez, J. & Hernáez, I. (2004). “Designing and Recording an Audiovisual Database of Emotional Speech in Basque”. In: *Proc. of the 4th LREC.* Lisbon, Portugal, s.p.

Niedenthal, P. M.; Halberstadt, J. B. & Setterlund, M. B. (1997). “Being Happy and Seeing ”Happy”: Emotional State Mediates Visual Word Recognition”. *Cognition Emotion* 11 (4), pp. 403–432.

NIST/SEMATECH (2014). *e-Handbook of Statistical Methods.* URL: http://www.itl.nist.gov/div898/handbook/.

Nolen-Hoeksema, S.; Fredrickson, B. L.; Loftus, G. R. & Wagenaar, W. A. (2009). *Atkinson & Hilgard’s Introduction to Psychology.* 15th ed. Hampshire, UK: Cengage Learning EMEA.

Noll, A. M. (Feb. 1967). “Cepstrum Pitch determination”. *J Acoust Soc Am* 41, pp. 293–309.

Nwe, T. L.; Foo, S. W. & Silva, L. C. D. (2003). “Speech emotion recognition using hidden Markov models”. *Speech Commun* 41 (4), pp. 603–623.

Olson, D. L. & Delen, D. (2008). *Advanced Data Mining Techniques.* Berlin, Heidelberg, Germany: Springer.

Ortony, A.; Clore, G. L. & Collins, A. (1990). *The Cognitive Structure of Emotions.* Cambridge, UK: Cambridge University Press.

Ortony, A. & Turner, T. J. (1990). “What’s basic about basic emotions?” *Psychol Rev* 97 (3), pp. 315–331.

Ozer, D. J. & Benet-Martinez, V. (2006). “Personality and the prediction of consequential outcomes”. *Annu Rev Psychol* 57 (3), pp. 401–421.

Paleari, M.; Huet, B. & Chellali, R. (2010). “Towards multimodal emotion recognition: a new approach”. In: *Proc. of the ACM CIVR-2010.* Xi’an, China, pp. 174–181.

Paliwal, K. K. & Rao, P. V. S. (1982). “On the performance of Burg’s method of maximum entropy spectral analysis when applied to voiced speech”. *Signal Process* 4 (1), pp. 59–63.

Panning, A.; Al-Hamadi, A. & Michaelis, B. (2010). “Active Shape Models on adaptively refined mouth emphasizing color images”. In: *Proc. of the 18th WSCG (Communication Papers).* Plzen, Czech Republic, pp. 221–228.

Panning, A.; Siegert, I.; Al-Hamadi, A.; Wendemuth, A.; Rösner, D.; Frommer, J.; Krell, G. & Michaelis, B. (2012). "Multimodal Affect Recognition in Spontaneous HCI Environment". In: *Proc. of 2012 IEEE ICSPCC.* Hong Kong, China, pp. 430–435.

Paschen, H. (1995). "Die Funktion der Diskurspartikel HM (The function of discourse particles HM)". MA thesis. University Mainz.

Patel, S. (2009). "An Acoustic Model of the Emotions Perceivable from the Suprasegmental Cues in Speech". PhD thesis. University of Florida.

Pavot, W.; Diener, E. & Fujita, F. (1990). "Extraversion and happiness". *Pers Indiv Differ* 11 (11), pp. 1299–1306.

Pearson, A. V. & Hartley, H. O. (1972). *Biometrica Tables for Statisticians.* Vol. 2. Cambridge, UK: Cambridge University Press.

Pedersen, W. C.; Bushman, B. J.; Vasquez, E. A. & Miller, N. (2008). "Kicking the (barking) dog effect: The moderating role of target attributes on triggered displaced aggression". *Pers Soc Psychol B* 34, pp. 1382–1395.

Philippou-Hübner, D.; Vlasenko, B.; Böck, R. & Wendemuth, A. (2012). "The Performance of the Speaking Rate Parameter in Emotion Recognition from Speech". In: *Proc. of the 2012 IEEE ICME.* Melbourne, Australia, pp. 296–301.

Picard, R. W. (1997). *Affective Computing.* Cambridge, USA: MIT Press.

Picard, R. R. & Cook, R. D. (Sept. 1984). "Cross-Validation of Regression Models". *J Am Stat Assoc* 79 (387), pp. 575–583.

Pieraccini, R. (2012). *The Voice in the Machine. Building Computers That Understand Speech.* Cambridge, USA: MIT Press.

Pittermann, J.; Pittermann, A. & Minker, W. (2010). *Handling Emotions in Human-Computer Dialogues.* Amsterdam, The Netherlands: Springer.

Plutchik, R. (1980). *Emotion, a psychoevolutionary synthesis.* New York, USA: Harper & Row.

– (1991). *The emotions.* revised. Lanham, USA: University Press of America.

Pols, L. C. W.; Kamp van der, L. J. T. & Plomp, R. (1969). "Perceptual and physical space of vowel sounds". *J Acoust Soc Am* 46, pp. 458–467.

Poppe, P.; Stiensmeier-Pelster, J. & Pelster, A. (2005). *Attributionsstilfragebogen für Erwachsene (ASF-E).* Göttingen, Germany: Hogrefe.

Potamianos, A. & Narayanan, S. (2007). "A review of the acoustic and linguistic properties of children's speech". In: *Proc. of the 9th IEEE MMSP.* Crete, Greece, pp. 22–25.

Powers, D. M. W. (2011). "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation". *J of Mach Lear Tech* 2 (1), pp. 37–63.

– (2012). "The problem with kappa". In: *Proc. of the 13th ACM EACL.* Avignon, France, pp. 345–355.

Prylipko, D.; Rösner, D.; Siegert, I.; Günther, S.; Friesen, R.; Haase, M.; Vlasenko, B. & Wendemuth, A. (2014a). "Analysis of significant dialog events in realistic human–computer interaction". *Journal on Multimodal User Interfaces* 8 (1), pp. 75–86.

Prylipko, D.; Egorow, O.; Siegert, I. & Wendemuth, A. (2014b). "Application of Image Processing Methods to Filled Pauses Detection from Spontaneous Speech". In: *Proc. of the INTERSPEECH-2014*. Singapore, s.p.

Ptacek, P.; Sander, E.; Maloney, W. & Jackson, C. (1966). "Phonatory and Related Changes with Advanced Age". *Journal Speech Hear Res* 9, pp. 353–360.

Rabiner, L. R.; Cheng, M.; Rosenberg, A. E. & McGonegal, C. (1976). "A comparative performance study of several pitch detection algorithms". *IEEE Trans. Acoust., Speech, Signal Process.* 24 (5), pp. 399–418.

Rabiner, L. R. & Juang, B.-H. (1993). *Fundamentals of Speech Recognition.* Upper Saddle River, USA: Prentice Hall.

Ramig, L. & Ringel, R. (1983). "Effect of Psychological Aging on Selected Acoustic characteristics of Voice". *Journal Speech Hear Res* 26, pp. 22–30.

Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004). *Handbuch für das computergestützte Transkribieren nach HIAT.* Tech. rep. SFB 538 Mehrsprachigkeit.

Reynolds, D. A.; Quatieri, T. F. & Dunn, R. B. (2000). "Speaker Verification Using Adapted Gaussian Mixture Models". *Digit Signal Process* 10 (1-3), pp. 19–41.

Rochester, S. R. (1973). "The significance of pauses in spontaneous speech". *J Psycholinguist Res* 2 (1), pp. 51–81.

Rogers, Y.; Sharp, H. & Preece, J. (2011). *Interaction Design - Beyond Human-Computer Interaction.* 3rd ed. Hoboken, USA: John Wiley & Sons.

Rosenberg, A. (2012). "Classifying Skewed Data: Importance Weighting to Optimize Average Recall". In: *Proc. of the INTERSPEECH-2012.* Portland, USA, s.p.

Rowe, G.; Hirsh, J. B. & Anderson, A. K. (2007). "Positive affect increases the breadth of attentional selection". *P Natl Acad Sci USA* 104 (1), pp. 383–388.

Russel, J. A. (Dec. 1980). "Three dimensions of emotion". *J Pers Soc Psychol* 39 (9), pp. 1161–1178.

Russel, J. A. & Mehrabian, A. (1974). "Distinguishing anger and anxiety in terms of emotional response factors". *J Consult Clin Psych* 42, pp. 79–83.

Ruvolo, P.; Fasel, I. & Movellan, J. R. (Sept. 2010). "A Learning Approach to Hierarchical Feature Selection and Aggregation for Audio Classification". *Pattern Recogn Lett* 31 (12), pp. 1535–1542.

Rösner, D.; Frommer, J.; Friesen, R.; Haase, M.; Lange, J. & Otto, M. (2012). "LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions". In: *Proc. of the 8th LREC.* Istanbul, Turkey, pp. 96–103.

Sacharin, V.; Schlegel, K.; & Scherer, K. R. (2012). *Geneva Emotion Wheel rating study.* Tech. rep. NCCR Affective Sciences: Center for Person, Kommunikation, Aalborg University.

Saeed, A.; Niese, R.; Al-Hamadi, A. & Panning, A. (2011). “Hand-face-touch Measure: a Cue for Human Behavior Analysis”. In: *Proc. of the IEEE ICIS 2011.* Vol. 3. Guangzhou, China, pp. 605–609.

Sahidullah, M. & Saha, G. (2012). “Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition”. *Speech Commun* 54 (4), pp. 543–565.

Sakai, T. & Doshita, S. (1962). “The Phonetic Typewriter”. In: *Proc. of the IFIP Congress 62.* Munich, Germany, pp. 445–450.

Sakoe, H. & Chiba, S. (Feb. 1978). “Dynamic Programming Algorithm Quantization for Spoken Word Recognition”. *IEEE Trans. Acoust., Speech, Signal Process.* 26 (1), pp. 43–49.

Salmon, W. C. (1983). *Logic.* 3rd ed. Englewood Cliffs, USA: Prentice-Hall.

Savran, A.; Cao, H.; Shah, M.; Nenkova, A. & Verma, R. (2012). “Combining Video, Audio and Lexical Indicators of Affect in Spontaneous Conversation via Particle Filtering”. In: *Proc. of the 14th ACM ICMI’12.* Santa Monica, USA, pp. 485–492.

Schaffer, C. (1993). “Selecting a classification method by cross-validation”. *Machine Learning* 13 (1), pp. 135–143.

Scherer, K. R. (1984). “On the nature and function of emotion: A component process approach”. In: *Approaches to emotion.* Ed. by Scherer, K. R. & Ekman, P. Hillsdale, USA: Lawrence Erlbaum, pp. 293–317.

– (1994). “Affect Bursts”. In: *Emotions.* Ed. by Goozen van, S. H. M.; Poll van de, N. E. & Sergeant, J. A. Hillsdale, USA: Lawrence Erlbaum, pp. 161–193.

– (2001). “Appraisal Considered as a Process of Multilevel Sequential Checking”. In: *Appraisal Processes in Emotion: Theory, Methods, Research.* Ed. by Scherer, K. R.; Schorr, A. & Johnstone, T. Oxford, UK: Oxford University Press, pp. 92–120.

– (2005a). “Unconscious Processes in Emotion: The Bulk of the Iceberg”. In: *Emotion and Consciousness.* Ed. by Niedenthal, P.; Feldman-Barrett, L. & Winkielman, P. New York, USA: Guilford Press, pp. 312–334.

– (2005b). “What are emotions? And how can they be measured?” *Soc Sci Inform* 44 (4), pp. 695–729.

Scherer, K. R.; Banse, R.; Wallbott, H. G. & Goldbeck, T. (1991). “Vocal cues in emotion encoding and decoding”. *Motivation and Emotion* 15.2, pp. 123–148.

Scherer, K. R.; Dan, E. & Flykt, A. (2006). “What determines a feeling’s position in affective space? A case for appraisal”. *Cognition Emotion* 20 (1), pp. 92–113.

Scherer, K. R.; Shuman, V.; Fontaine, J. R. J. & Soriano, C. (2013). “The GRID meets the Wheel: Assessing emotional feeling via self-report”. In: *Components of emotional meaning: A sourcebook*. Ed. by Fontaine, J. R. J.; Scherer, K. R. & Soriano, C. Oxford, UK: Oxford University Press, s.p.

Scherer, S. (2011). “Analyzing the user’s state in HCI: from crisp emotions to conversational dispositions”. PhD thesis. Ulm University.

Scherer, S.; Siegert, I.; Bigalke, L. & Meudt, S. (2010). “Developing an Expressive Speech Labeling Tool Incorporating the Temporal Characteristics of Emotion”. In: *Proc. of the 7th LREC*. Valletta, Malta, pp. 1172–1175.

Scherer, S.; Glodek, M.; Schwenker, F.; Campbell, N. & Palm, G. (2012). “Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data”. *ACM TiiS* 2.1, pp. 111–144.

Schick, T. & Vaughn, L. (2002). *How to think about weird things: critical thinking for a New Age*. Boston, USA: McGraw-Hill Higher Education.

Schimmack, U. (May 1997). “The Berlin Everyday Language Mood Inventory (BELMI): Toward the content valid assessment of moods”. *Diagnostica* 43 (2), pp. 150–173.

Schlosberg, H. (1954). “Three dimensions of emotion”. *Psychol Rev* 61 (2), pp. 81–88.

Schmidt, J. E. (2001). “Bausteine der Intonation (Components of intonation)”. In: *Neue Wege der Intonationsforschung*. Vol. 157-158. Germanistische Linguistik. Hildesheim, Germany: Georg Olms Verlag, pp. 9–32.

Schmidt, T. & Wörner, K. (2009). “EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research”. *Pragmatics* 19 (4), pp. 565–582.

Schmidt, T. & Schütte, W. (2010). “FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction”. In: *Proc. of the 7th LREC*. Valletta, Malta, pp. 2091–2096.

Schmitt, N. (1996). “Uses and abuses of coefficient alpha”. *Psychol Assessment* 8 (4), pp. 350–353.

Schröder, M. (2003). “Experimental study of affect bursts”. *Speech Commun* 40 (1-2), pp. 99–116.

Schukat-Talamazzini, E. G. (1995). *Automatische Spracherkennung. Grundlagen, statistische Modelle und effiziente Algorithmen*. Braunschweig, Wiesbaden: Vieweg.

Schuller, B.; Seppi, D.; Batliner, A.; Maier, A. & Steidl, S. (2007a). “Towards More Reality in the Recognition of Emotional Speech”. In: *Proc. of the IEEE ICASSP-2007*. Vol. 4. Honolulu, USA, pp. 941–944.

Schuller, B.; Vlasenko, B.; Arsic, D.; Rigoll, G. & Wendemuth, A. (2008a). “Combining Speech Recognition and Acoustic Word Emotion Models for Robust Text-Independent Emotion Recognition”. In: *Proc. of the 2008 IEEE ICME*. Hannover, Germany, pp. 1333–1336.

Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G. & Wendemuth, A. (2009a). "Acoustic Emotion Recognition: A Benchmark Comparison of Performances". In: *Proc. of the IEEE ASRU-2009.* Merano, Italy, pp. 552–557.

Schuller, B.; Vlasenko, B.; Eyben, F.; Wollmer, M.; Stuhlsatz, A.; Wendemuth, A. & Rigoll, G. (2010a). "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies". *IEEE Trans. Affect. Comput.* 1 (2), pp. 119–131.

Schuller, B.; Steidl, S.; Batliner, A.; Burkhardt, F.; Devillers, L.; Mueller, C. & Narayanan, S. (2010b). "The INTERSPEECH 2010 Paralinguistic Challenge". In: *Proc. of the INTERSPEECH-2010.* Makuhari, Japan, pp. 2794–2797.

Schuller, B.; Steidl, S.; Batliner, A.; Schiel, F. & Krajewski, J. (2011a). "The INTERSPEECH 2011 Speaker State Challenge". In: *Proc. of the INTERSPEECH-2011.* Florence, Italy, pp. 3201–3204.

Schuller, B.; Steidl, S.; Batliner, A.; Nöth, E.; Vinciarelli, A.; Burkhardt, F.; Son van, v.; Weninger, F.; Eyben, F.; Bocklet, T.; Mohammadi, G. & Weiss, B. (2012a). "The INTERSPEECH 2012 Speaker Trait Challenge". In: *Proc. of the INTERSPEECH-2012.* Portland, USA, s.p.

Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Weninger, F.; Eyben, F.; Marchi, E.; Mortillaro, M.; Polychroniou, H. S. andA.; Valente, F. & Kim, S. (2013). "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism". In: *Proc. of the INTERSPEECH-2013.* Lyon, France, pp. 148–152.

Schuller, B. & Batliner, A. (2013). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing.* Hoboken, USA: John Wiley & Sons.

Schuller, B.; Arsic, D.; Rigoll, G.; Wimmer, M. & Radig, B. (2007b). "Audiovisual Behavior Modeling by Combined Feature Spaces". In: *Proc. of the IEEE ICASSP-2007.* Honolulu, USA, pp. 733–736.

Schuller, B.; Zhang, X. & Rigoll, G. (2008b). "Prosodic and spectral features within segment-based acoustic modeling". In: *Proc. of the INTERSPEECH-2008.* Brisbane, Australia, pp. 2370–2373.

Schuller, B.; Müller, R.; Eyben, F.; Gast, J.; Hörnler, B.; Wöllmer, M.; Rigoll, G.; Höthker, A. & Konosu, H. (2009b). "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application". *Image Vision Comput* 27 (12), pp. 1760–1774.

Schuller, B.; Steidl, S. & Batliner, A. (2009c). "The INTERSPEECH 2009 Emotion Challenge". In: *Proc. of the INTERSPEECH-2009.* Brighton, UK, pp. 312–315.

Schuller, B.; Valstar, M.; Eyben, F.; McKeown, G.; Cowie, R. & Pantic, M. (2011b). "AVEC 2011 –The First International Audio/Visual Emotion Challenge". In: *Affective Computing and Intelligent Interaction.* Ed. by D'Mello, S.; Graesser, A.; Schuller, B. & Martin, J.-C. Vol. 6975. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 415–424.

Schuller, B.; Batliner, A.; Steidl, S. & Seppi, D. (Nov. 2011c). “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge”. *Speech Commun* 53 (9-10), pp. 1062–1087.

Schuller, B.; Valstar, M.; Cowie, R. & Pantic, M. (2012b). “AVEC 2012: The Continuous Audio/Visual Emotion Challenge - an Introduction”. In: *Proc. of the 14th ACM ICMI'12*. Santa Monica, USA, pp. 361–362.

Scott, W. A. (Sept. 1955). “Reliability of Content Analysis: The Case of Nominal Scale Coding”. *Public Opin Quart* 19 (3), pp. 321–325.

Selting, M.; Auer, P.; Barth-Weingarten, D.; Bergmann, J. R.; Bergmann, P.; Birkner, K.; Couper-Kuhlen, E.; Deppermann, A.; Gilles, P.; Günthner, S.; Hartung, M.; Kern, F.; Mertzlufft, C.; Meyer, C.; Morek, M.; Oberzaucher, F.; Peters, J.; Quasthoff, U.; Schütte, W.; Stukenbrock, A. & Uhmann, S. (2009). “Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)”. *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 10, pp. 353–402.

Seppi, D.; Batliner, A.; Steidl, S.; Schuller, B. & Nöth, E. (2010). “Word Accent and Emotion”. In: *Proc. of the 5th Speech Prosody.* Chicago, USA, s. p.

Sezgin, M.; Gunsel, B. & Kurt, G. (2012). “Perceptual audio features for emotion detection”. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 (1), pp. 1–21.

Shahin, I. M. A. (2013). “Gender-dependent emotion recognition based on HMMs and SPHMMs”. *International Journal of Speech Technology* 16 (2), pp. 133–141.

Siegel, S. & Castellan jr., N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences.* 2nd ed. New York, USA: McGraw-Hill.

Siegert, I. (2014). *Results and Significance Test for Parameter Tuning, Classification Experiments on Speaker Group Dependent Modelling, and Discourse Particles as Interaction Patterns.* Tech. rep. IIKT, Otto-von-Guericke University Magdeburg.

Siegert, I.; Böck, R.; Philippou-Hübner, D.; Vlasenko, B. & Wendemuth, A. (2011). “Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins”. In: *Proc. of the 2011 IEEE ICME.* Barcelona, Spain, s.p.

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

– (2012b). “The Influence of Context Knowledge for Multimodal Annotation on natural Material”. In: *Joint Proceedings of the IVA 2012 Workshops.* Santa Cruz, USA, pp. 25–32.

Siegert, I.; Hartmann, K.; Philippou-Hübner, D. & Wendemuth, A. (2013a). “Human Behaviour in HCI: Complex Emotion Detection through Sparse Speech Features”. In:

Human Behavior Understanding. Ed. by Salah, A.; Hung, H.; Aran, O. & Gunes, H. Vol. 8212. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 246–257.

Siegert, I.; Hartmann, K.; Glüge, S. & Wendemuth, A. (2013b). "Modelling of Emotional Development within Human-Computer-Interaction". *Kognitive Systeme* 1, s.p.

Siegert, I.; Böck, R.; Hartmann, K. & Wendemuth, A. (2013c). "Speaker Group Dependent Modelling for Affect Recognition from Speech". In: *ERM4HCI 2013: The 1st Workshop on Emotion Representation and Modelling in Human-computer-interaction-systems.* Berlin, Heidelberg, Germany: Springer, s.p.

Siegert, I.; Böck, R. & Wendemuth, A. (2013d). "The Influence of Context Knowledge for Multi-modal Affective Annotation". In: *Human-Computer Interaction. Towards Intelligent and Implicit Interaction.* Ed. by Kurosu, M. Vol. 8008. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 381–390.

Siegert, I.; Glodek, M.; Panning, A.; Krell, G.; Schwenker, F.; Al-Hamadi, A. & Wendemuth, A. (2013e). "Using speaker group dependent modelling to improve fusion of fragmentary classifier decisions". In: *Proc. of 2013 IEEE CYBCONF.* Lausanne, Switzerland, pp. 132–137.

Siegert, I.; Haase, M.; Prylipko, D. & Wendemuth, A. (2014a). "Discourse Particles and User Characteristics in Naturalistic Human-Computer Interaction". In: *Human-Computer Interaction. Advanced Interaction Modalities and Techniques.* Ed. by Kurosu, M. Vol. 8511. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 492–501.

Siegert, I.; Böck, R. & Wendemuth, A. (2014b). "Inter-Rater Reliability for Emotion Annotation in Human-Computer Interaction – Comparison and Methodological Improvements". *Journal of Multimodal User Interfaces* 8 (1), pp. 17–28.

Siegert, I.; Prylipko, D.; Hartmann, K.; Böck, R. & Wendemuth, A. (2014c). "Investigating the Form-Function-Relation of the Discourse Particle "hm" in a Naturalistic Human-Computer Interaction". In: *Recent Advances of Neural Network Models and Applications.* Ed. by Bassis, S.; Esposito, A. & Morabito, F. C. Vol. 26. Smart Innovation, Systems and Technologies. Berlin, Heidelberg, Germany: Springer, pp. 387–394.

Siegert, I.; Philippou-Hübner, D.; Hartmann, K.; Böck, R. & Wendemuth, A. (2014d). "Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech". *Cognitive Computation*, s.p.

Sijtsma, K. (2009). "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha". *Psychometrika* 74 (1), pp. 107–120.

Siniscalchi, S. M.; Yu, D.; Deng, L. & Lee, C.-H. (2013). "Speech Recognition Using Long-Span Temporal Patterns in a Deep Network Model". *IEEE Signal Process. Lett.* 20 (3), pp. 201–204.

Smart, J. (1984). "Ockham's Razor". In: *Principles of Philosophical Reasoning.* Ed. by Fetzer, J. H. Lanham, USA: Rowman & Littlefield, pp. 118–128.

Smith, C. A. (1989). “Dimensions of appraisal and physiological response in emotion.” *J Pers Soc Psychol* 56, pp. 339–353.

Snell, R. C. & Milinazzo, F. (1993). “Formant location from LPC analysis data”. *IEEE Speech Audio Process.* 1 (2), pp. 129–134.

Soeken, K. L. & Prescott, P. A. (Aug. 1986). “Issues in the Use of Kappa to Estimate Reliability”. *Med Care* 24 (8), pp. 733–741.

Steidl, S. (2009). “Automatic classification of emotion-related user states in spontaneous children’s speech”. PhD thesis. FAU Erlangen-Nürnberg.

Stevens, S. S.; John, V. & Newman, E. B. (1937). “A scale for the measurement of the psychological magnitude pitch”. *J Acoust Soc Am* 8 (3), pp. 185–190.

Sullivan, H. S. (1953). *The interpersonal theory of psychiatry.* New York, USA: Norton.

Tamir, M. (Apr. 2009). “Differential preferences for happiness: Extraversion and trait-consistent emotion regulation”. *J Pers* 77 (2), pp. 447–470.

Tan, W. Y. (June 1982). “On comparing several straight lines under heteroscedasticity and robustness with respect to departure from normality”. *Commun Stat A-Theor* 11 (7), pp. 731–750.

Tarasov, A. & Delany, S. J. (2011). “Benchmarking classification models for emotion recognition in natural speech: A multi-corporal study”. In: *Proc. of the 9th IEEE FG.* Santa Barbara, USA, pp. 841–846.

Teixeira, J. P.; Oliveira, C. & Lopes, C. (2013). “Vocal Acoustic Analysis –Jitter, Shimmer and HNR Parameters”. *Procedia Technology* 9 (0), pp. 1112–1122.

Thun, F. Schulz von (1981). *Miteinander reden 1 - Störungen und Klärungen.* Reinbek, Germany: Rowohlt.

Tohkura, Y. (1987). “A weighted cepstral distance measure for speech recognition”. *IEEE Trans. Acoust., Speech, Signal Process.* 35 (10), pp. 1414–1422.

Tolkmitt, F. J. & Scherer, K. R. (1986). “Effect of experimentally induced stress on vocal parameters”. *J Exper Psychol Hum Percept Perform* 12 (3), pp. 302–313.

Torres-Carrasquillo, P. A.; Singer, E.; Kohler, M. A. & Deller, J. R. (2002). “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features”. In: *Proc. of the INTERSPPECH-2002.* Denver, USA, pp. 89–92.

Truong, K. P.; Neerincx, M. A. & Leeuwen van, D. A. (2008). “Assessing Agreement of Observer- and Self-Annotations in Spontaneous Multimodal Emotion Data”. In: *Proc. of the INTERSPEECH-2008.* Brisbane, Australia, pp. 318–321.

Truong, K. P.; David Leeuwen van, A. & Jong de, F. M. G. (Nov. 2012). “Speech-based recognition of self-reported and observed emotion in a dimensional space”. *Speech Commun* 54 (9), pp. 1049–1063.

Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R. & Pantic, M. (2013). "AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge". In: *Proc. of the 3rd ACM AVEC '13.* Barcelona, Spain, pp. 3–10.

Veer van der, G. C.; Tauber, M. J.; Waem, Y. & Muylwijk van, B. (1985). "On the interaction between system and user characteristics". *Behav Inform Technol* 4 (4), pp. 289–308.

Vergin, R.; Farhat, A. & O'Shaughnessy, D. (1996). "Robust Gender-Dependent Acoustic-Phonetic Modelling In Continuous Speech Recognition Based On A New Automatic Male/Female Classification". In: *Proc. of the ICSLP-1996.* Philadelphia, USA, pp. 1081–1084.

Ververidis, D. & Kotropoulos, C. (2006). "Emotional speech recognition: Resources, features, and methods". *Speech Commun* 48 (9), pp. 1162–1181.

Veth, J. de & Boves, L. (Feb. 2003). "On the Efficiency of Classical RASTA Filtering for Continuous Speech Recognition: Keeping the Balance Between Acoustic Pre-processing and Acoustic Modelling". *Speech Commun* 39 (3-4), pp. 269–286.

Vinciarelli, A.; Pantic, M. & Bourlard, H. (Nov. 2009). "Social Signal Processing: Survey of an Emerging Domain". *Image Vision Comput* 27 (12), pp. 1743–1759.

Viterbi, A. J. (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Trans. Inf. Theory* 13 (2), pp. 260–269.

Vlasenko, B. (2011). "Emotion Recognition within Spoken Dialog Systems". PhD thesis. Otto von Guericke University Magdeburg.

Vlasenko, B. & Wendemuth, A. (2013). "Determining the Smallest Emotional Unit for Level of Arousal Classification". In: *Proc. of the 5th IEEE ACII.* Geneva, Switzerland, pp. 734–739.

Vlasenko, B.; Schuller, B.; Wendemuth, A. & Rigoll, G. (2007a). "Combining frame and turn-level information for robust recognition of emotions within speech". In: *Proc. of the INTERSPEECH-2007.* Antwerp, Belgium, pp. 2249–2252.

Vlasenko, B.; Schuller, B.; Wendemuth, A. & Rigoll, G. (2007b). "Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing". In: *Affective Computing and Intelligent Interaction.* Ed. by Paiva, A. C. R.; Prada, R. & Picard, R. W. Vol. 4738. LNCS. Berlin, Heidelberg, Germany: Springer, pp. 139–147.

Vlasenko, B.; Prylipko, D.; Böck, R. & Wendemuth, A. (2014). "Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications". *Comput Speech Lang* 28 (2), pp. 483–500.

Vogt, T. & André, E. (2005). "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition". In: *Proc. of the 2005 IEEE ICME.* Amsterdam, The Netherlands, pp. 474–477.

– (2006). “Improving automatic emotion recognition from speech via gender differentiation”. In: *Proc. of the 5th LREC.* Genoa, Italy, s.p.

Wagner, J.; André, E.; Lingenfelser, F. & Kim, J. (Oct. 2011). “Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data”. *IEEE Trans. Affect. Comput.* 2 (4), pp. 206–218.

Wahlster, W. (ed.). *SmartKom: Foundations of Multimodal Dialogue Systems.* Heidelberg, Berlin: Springer.

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

Ward, N. (2004). “Pragmatic functions of prosodic features in non-lexical utterances”. In: *Proc. of the 2nd Speech Prosody.* Nara, Japan, pp. 325–328.

Watson, D.; Clark, L. A. & Tellegen, A. (June 1988). “Development and validation of brief measures of positive and negative affect: the PANAS scales”. *J Pers Soc Psychol* 54 (6), pp. 1063–1070.

Watzlawick, P.; Beavin, J. H. & Jackson, D. D. (1967). *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes.* Bern, Switzerland: Norton.

Weinberg, G. M. (1971). *The psychology of computer programming.* New York, USA: Van Nostrand Reinhold.

Wendemuth, A. (2004). *Grundlagen der stochastischen Sprachverarbeitung.* Munich, Germany: Oldenbourg.

Wendemuth, A. & Biundo, S. (2012). “A Companion Technology for Cognitive Technical Systems”. In: *Cognitive Behavioural Systems.* Ed. by Esposito, A.; Esposito, A.; Vinciarelli, A.; Hoffmann, R. & Müller, V. Vol. 7403. LNCS. Berlin Heidelberg, Germany: Springer, pp. 89–103.

Wilks, Y (2005). “Artificial companions”. *Interdisciplinary Science Reviews* 30 (2), pp. 145–152.

Wolff, J. G. (2006). “Medical diagnosis as pattern recognition in a framework of information compression by multiple alignment, unification and search”. *Decis Support Syst* 42 (2), pp. 608–625.

Wong, E. & Sridharan, S. (2002). “Utilise Vocal Tract Length Normalisation for Robust Automatic Language Identification”. In: *Proc. of the 9th SST.* Melbourne, Australia, s.p.

Wundt, W. M. (1919). *Vorlesungen über die Menschen- und Tierseele.* 6th ed. Leipzig: L. Voss.

Wöllmer, M.; Eyben, F.; Schuller, B.; Douglas-Cowie, E. & Cowie, R. (2009). "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks". In: *Proc. of the INTERSPEECH-2009*. Brighton, UK, pp. 1595–1598.

Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B. & Narayanan, S. (2010). "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling". In: *Proc. of the INTERSPPECH-2010*. Makuhari, Japan, pp. 2362–2365.

Yang, Y.-H.; Lin, Y.-C.; Su, Y.-F. & Chen, H. H. (2007). "Music Emotion Classification: A Regression Approach". In: *Proc. of the 2007 IEEE ICME*. Beijing, China, pp. 208–211.

Young, S.; Evermann, G.; Gales, M.; Hain, T.; Kershaw, D.; Liu, X.; Moore, G.; Odell, J.; Ollason, D.; Povey, D.; Valtchev, V. & Woodland, P. (2006). *The HTK Book (for HTK Version 3.4)*. Cambridge, UK: Cambridge University Engineering Department.

Young, S. (2008). "HMMs and Related Speech Recognition Technologies". In: *Springer Handbook of Speech Processing*. Ed. by Benesty, J.; Sondhi, M. M. & Huang, Y. Berlin, Heidelberg, Germany: Springer.

Yu, C; Aoki, P. M. & Woodruff, A. (2004). "Detecting user engagement in everyday conversations". In: *Proc. of the INTERSPEECH-2004*. Jeju, Korea, pp. 1329–1332.

Yuan, J. & Liberman, M. (2010). "Robust speaking rate estimation using broad phonetic class recognition". In: *Proc. of the IEEE ICASSP-2010*. Dallas, USA, pp. 4222–4225.

Zeng, Z.; Pantic, M.; Roisman, G. I. & Huang, T. S. (2009). "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions". *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (1), pp. 39–58.

Zhan, P. & Waibel, A. (1997). *Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition*. Tech. rep. CMU-CS-97-148. Carnegie Mellon University.

Zhang, T.; Hasegawa-Johnson, M. & Levinson, S. E. (2004). "Children's emotion recognition in an intelligent tutoring scenario". In: *Proc. of the INTERSPEECH-2004*. Jeju, Korea, pp. 1441–1444.

Zhang, Z.; Weninger, F.; Wöllmer, M. & Schuller, B. (2011). "Unsupervised learning in cross-corpus acoustic emotion recognition". In: *Proc. of the IEEE ASRU-2011*. Waikoloa, USA, pp. 523–528.

List of Authored Publications

Articles in International Journals

1 I. Siegert, D. Philippou-Hübner, K. Hartmann, R. Böck and A. Wendemuth. “Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech”. *Cognitive Computation*, pp 1-22, 2014.

2 D. Prylipko, D. Rösner, I. Siegert, S. Günther, R. Friesen, M. Haase, V. Vlasenko and A. Wendemuth. “Analysis of significant dialog events in realistic human–computer interaction”, *Journal on Multimodal User Interfaces* 8(1), pp. 75–86, 2014.

3 I. Siegert, R. Böck and A. Wendemuth. “Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements”. *Journal on Multimodal User Interfaces* 8(1), pp. 17–28, 2014.

Articles in National Journals

4 I. Siegert, K. Hartmann, S. Glüge and A. Wendemuth. “Modelling of Emotional Development within Human-Computer-Interaction”. *Kognitive Systeme* 1, 2013, s.p.

Contributions in Book Series and International Conferences

5 K. Hartmann, I. Siegert and D. Prylipko. “Emotion and Disposition Detection in Medical Machines: Chances and Challenges”. In S.P. Rysewyk and M. Pontier (eds.). *Machine Medical Ethics*. Intelligent Systems, Control and Automation: Science and Engineering series, V. 74, Springer International Publishing, 2015, pp 317–339.

6 I. Siegert, D. Prylipko, K. Hartmann, R. Böck and A. Wendemuth. “Investigating the Form-Function-Relation of the Discourse Particle “hm” in a Naturalistic Human-Computer Interaction”. In S. Bassis and A. Esposito F.C. Morabito (eds.). *Recent Advances of Neural Network Models and Applications*. Smart Innovation, Systems and Technologies series, V. 26, Springer, 2014, pp 387–394.

7 D. Prylipko, O. Egorow, I. Siegert and A. Wendemuth. “Application of Image Processing Methods to Filled Pauses Detection from Spontaneous Speech”. In *Proceedings of the INTERSPEECH 2014*. 2014, pp. 1816-1820.

8 I. Siegert, M. Haase, M., D. Prylipko, A. Wendemuth. “Discourse Particles and User Characteristics in Naturalistic Human-Computer Interaction”. In Kurosu, M. (eds.). *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*. Lecture Notes in Computer Science series, V. 8511, Springer Berlin, Heidelberg, 2014 pp. 492–501.

9 I. Siegert, R. Böck, K. Hartmann, and A. Wendemuth. "Speaker Group Dependent Modelling for Affect Recognition from Speech". In *Proceedings of ERM4HCI 2013: The 1st Workshop on Emotion Representation and Modelling in Human-computer-interaction-systems*. Sydney, Australia, December 2013, s.p.

10 I. Siegert, M. Glodek, A. Panning, G. Krell, F. Schwenker, A. Al-Hamadi and A. Wendemuth. "Using speaker group dependent modelling to improve fusion of fragmentary classifier decisions". In *IEEE International Conference on Cybernetics (CYBCONF)*. 2013, pp. 132–137.

11 I. Siegert, K. Hartmann, D. Philippou-Hübner and A. Wendemuth. "Human Behaviour in HCI: Complex Emotion Detection through Sparse Speech Features". In A.A. Salah, H. Hung, O. Aran and H. Gunes (eds.). *Human Behavior Understanding*. Lecture Notes in Computer Science series, V. 8212, Springer International Publishing, 2013, pp. 246-257.

12 R. Böck, S. Glüge, I. Siegert and A. Wendemuth. "Annotation and Classification of Changes of Involvement in Group Conversation". In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*. September 2013, pp. 803–808.

13 K. Hartmann, I. Siegert, D. Philippou-Hübner and A. Wendemuth. "Emotion Detection in HCI: From Speech Features to Emotion Space". In S. Narayanan (ed.). *Analysis, Design, and Evaluation of Human-Machine Systems*. V. 12/1, 2013, pp. 288–295.

14 I. Siegert, R. Böck and A. Wendemuth. "The Influence of Context Knowledge for Multi-modal Affective Annotation". In M. Kurosu (ed.). *Human-Computer Interaction. Towards Intelligent and Implicit Interaction*. Lecture Notes in Computer Science series, V. 8008, Springer Berlin, Heidelberg, 2013, pp. 381–390.

15 R. Böck, K. Limbrecht-Ecklundt, I. Siegert, S. Walter and A. Wendemuth. "Audio-Based Pre-classification for Semi-automatic Facial Expression Coding". In M. Kurosu (ed.). *Human-Computer Interaction. Towards Intelligent and Implicit Interaction*. Lecture Notes in Computer Science series, V. 8008, Springer Berlin, Heidelberg, 2013, pp. 301–309

16 D. Schmidt, H. Sadri, A. Szewieczek, M. Sinapius, P. Wierach, I. Siegert and A. Wendemuth. "Characterization of Lamb wave attenuation mechanisms". In *Proceedings of SPIE Smart Structures and Materials+ Nondestructive Evaluation and Health Monitoring*. V. 8695, 2013, pp. 869503–869510.

17 G. Krell, M. Glodek, A. Panning, I. Siegert, B. Michaelis, A. Wendemuth and F. Schwenker. "Fusion of Fragmentary Classifier Decisions for Affective State Recognition". In F. Schwenker, S. Scherer and L.P. Morency (eds.). *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Lecture Notes in Artificial Intelligence series, V. 7742, Springer Berlin, Heidelberg, 2013, pp. 116–130.

18 I. Siegert, R. Böck and A. Wendemuth. “Modeling users’ mood state to improve human-machine-interaction”. In A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann and V. C. Müller (eds.). *Cognitive Behavioural Systems*. Lecture Notes in Computer Science series, V. 7403, Springer Berlin, Heidelberg, 2012, pp. 273–279.

19 I. Siegert, R. Böck and A. Wendemuth. “The Influence of Context Knowledge for Multimodal Annotation on natural Material”. In R. Böck, F. Bonin, N. Campbell, J. Edlund, I. Kok, R. Poppe and D. Traum (eds.). *Joint Proceedings of the IVA 2012 Workshops*. September 2012, pp. 25–32.

20 A. Panning, I. Siegert, A. Al-Hamadi, A. Wendemuth, D. Rösner, J. Frommer, G. Krell and Bend Michaelis. “Multimodal Affect Recognition in Spontaneous HCI Environment”. In *Proceedings of 2012 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. 2012, pp. 430–435.

21 J. Frommer, B. Michaelis, D. Rösner, A. Wendemuth, R. Friesen, M. Haase, M. Kunze, R. Andrich, J. Lange, A. Panning and I. Siegert. “Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus”. In N. Calzolari, K. Choukri, T. Declerck, M. Doğan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis (eds.). *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*. Mai 2012, pp. 3064–3069.

22 K. Hartmann, I. Siegert, S. Glüge, A. Wendemuth, M. Kotzyba and B. Deml. “Describing Human Emotions Through Mathematical Modelling”. In *Proceedings of the MATHMOD 2012*. Februar 2012, s.p.

23 R. Böck, I. Siegert, M. Haase, J. Lange and A. Wendemuth. “ikannotate - A Tool for Labelling, Transcription, and Annotation of Emotionally Coloured Speech”. In S. D’Mello, A: Graesser, B. Schuller and J.C. Martin (eds.). *Affective Computing and Intelligent Interaction*. Lecture Notes in Computer Science series, V. 6974, Springer Berlin, Heidelberg, 2011, pp. 25–34.

24 I. Siegert, R. Böck, D. Philippou-Hübner, V. Vlasenko and A. Wendemuth. “Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins”. In *Proceedings of the 2011 IEEE International Conference on Multimedia & Expo*. 2011, s.p.

25 V. Vlasenko, D. Philippou-Hübner, D. Prylipko, R. Böck, I. Siegert and A. Wendemuth. “Vowels Formants Analysis Allows Straightforward Detection of High Arousal Emotions”. In *Proceedings of the 2011 IEEE International Conference on Multimedia & Expo*. 2011, s.p.

26 R. Böck, I. Siegert, V. Vlasenko, A. Wendemuth, M. Haase and J. Lange. “A Processing Tool for Emotionally Coloured Speech”. In *Proceedings of the 2011 IEEE International Conference on Multimedia & Expo*. 2011, s.p.

27 S. Scherer, I. Siegert, L. Bigalke and S. Meudt. “Developing an Expressive Speech Labeling Tool Incorporating the Temporal Characteristics of Emotion”. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner and D. Tapias (eds.). *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*. Mai 2010, s.p.

Editorship

28 R. Böck, N. Degens, D. Heylen, S. Louchart, W. Minker, L.-P. Morency, A. Nazir, F. Schwenker and I. Siegert (eds.). *Joint Proceedings of the 2013 T2CT and CCGL Workshops*. Otto von Guericke University Magdeburg, 2013.

Contributions in National Conferences

29 M. Kotzyba, I. Siegert, Tatiana Gossen, A. Nürnberger and A. Wendemuth. “Exploratory Voice-Controlled Search for Young Users : Challenges and Potential Benefits”. In A. Wendemuth, M. Jipp, A. Kluge and D. Söffker (eds.). *Proceedings 3. Interdisziplinärer Workshop Kognitive Systeme: Mensch, Teams, Systeme und Automaten*. März 2013, s.p.

30 I. Siegert, R. Böck, D. Philippou-Hübner and A. Wendemuth. “Investigation of Hierarchical Classification for Simultaneous Gender and Age Recognitions”. In *Proceedings of the 23. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2012)*. August 2012, pages 58–65.

31 R. Böck, K. Limbrecht, I. Siegert, S. Glüge, S. Walter and A. Wendemuth. “Combining Mimic and Prosodic Analyses for User Disposition Classification”. In *Proceedings of the 23. Konferenz Elektronische Sprachsignalverarbeitung (ESSV 2012)*. August 2012, pages 220–227.

32 M. Kotzyba, B. Deml, Hendrik Neumann, S. Glüge, K. Hartmann, I. Siegert, A. Wendemuth, Harald Traue and S. Walter. “Emotion Detection by Event Evaluation using Fuzzy Sets as Appraisal Variables”. In N. Rußwinkel, U. Drewitz and H. Rijn (eds.). *Proceedings of the 11th International Conference on Cognitive Modeling (ICCM 2012)*. April 2012, pages 123–124.

33 T. Grosser, V. Heine, S. Glüge, I. Siegert, J. Frommer and A. Wendemuth. “Artitificial Intelligent Systems and Cognition”. In *Proceedings of 1st International Conference on What makes Humans Human*. 2010, s.p.

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 30. 03. 2015

Dipl.-Ing. Ingo Siegert