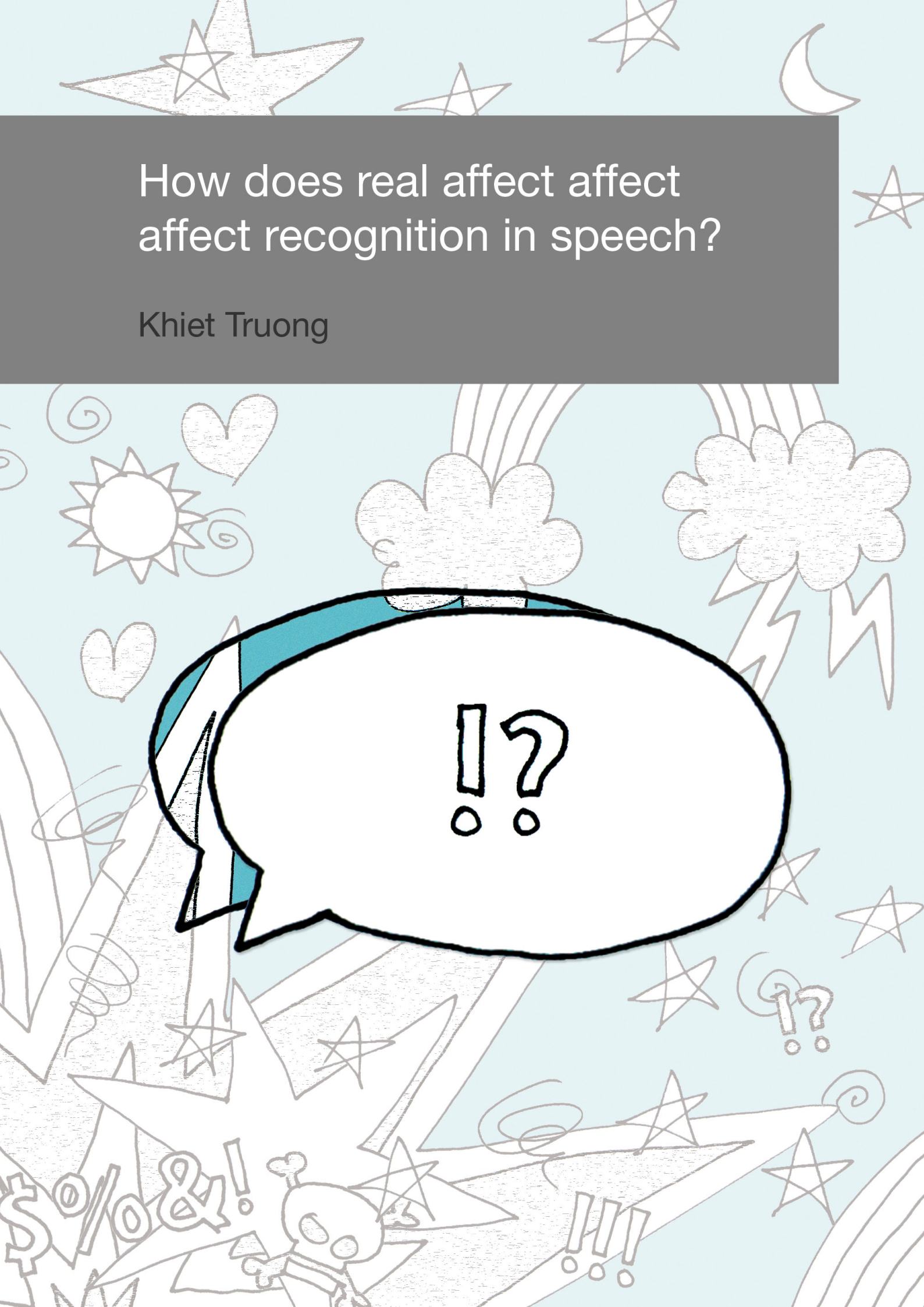
How does real affect affect affect recognition in speech?

Khiet Truong



How Does Real Affect Affect Affect Recognition In Speech?

Khiet Truong

PhD dissertation committee:

Chairman and Secretary:

Prof. dr. ir. A. J. Mouthaan, University of Twente, NL

Promotores:

Prof. dr. F. M. G. de Jong, University of Twente, NL

Prof. dr. ir. D. A. van Leeuwen, Radboud University Nijmegen/TNO, NL

Members:

Prof. dr. ir. A. Nijholt, University of Twente, NL

Prof. dr. M. Pantic, University of Twente, NL

Prof. dr. M. A. Neerincx, Delft University of Technology, NL

Prof. dr. M. G. J. Swerts, Tilburg University, NL

Prof. dr.-ing. E. Nöth, Friedrich-Alexander University Erlangen-Nuremberg, D

Prof. dr. N. Campbell, Trinity College Dublin, IRL

CTIT Dissertation Series No. 09-152
Center for Telematics and Information Technology (CTIT)
P.O. Box 217 – 7500AE Enschede – the Netherlands
ISSN: 1381-3617

MultimediaN

The research reported in this thesis has been supported by MultimediaN, a Dutch BSIK project.

TNO Defence, Security, and Safety

The research reported in this thesis has been carried out at the department of Human Interfaces at TNO Defence, Security, and Safety, Business Unit Human Factors in Soesterberg.

SIKS Dissertation Series No. 2009-33

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

ISBN: 978-90-365-2880-1
ISSN: 1381-3617, No. 09-152

HOW DOES REAL AFFECT AFFECT AFFECT RECOGNITION IN SPEECH?

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. H. Brinksma,
on account of the decision of the graduation committee
to be publicly defended
on Thursday, August 27, 2009 at 16:45

by

Khiet Phuong Truong
born on September 10, 1980
in Apeldoorn, The Netherlands

Promotores: Prof. dr. F. M. G. de Jong
Prof. dr. ir. D. A. van Leeuwen

ISBN: 978-90-365-2880-1

Acknowledgments

What a journey. Writing a PhD dissertation has been a great experience for me, and I would like to thank the people who I have met along this journey, and who have helped making all of this a wonderful experience. First of all, I would like to thank David van Leeuwen, my promotor and daily supervisor at TNO, who has been a great supervisor, supporting and encouraging me during my research. I have learned a great deal of him, from drawing nice DET curves, to carrying out good research, to removing ugly white spaces in LaTeX. Secondly, Franciska de Jong, my promotor, is thanked for her supervision. I valued her questions which made me think deeper about my research and which made me formulate and structure my work in a better way. I would also like to thank the dissertation committee for reading my thesis, and for providing me valuable comments.

I thank Arjan van Hessen for pointing out this PhD-job to me, and for encouraging me to do this.

TNO (Soesterberg) and the project MultimediaN have supported my research for which I am thankful. The project leaders of MultimediaN, first Adelbert Bronkhorst, and then Mark Neerincx, are thanked for their commitment to this project. Mark is also thanked for his encouragement and support throughout my research. I have enjoyed the talks and discussions with my TNO colleagues. Willem, Ronald, en Judith were fine roommates and discussion partners. Thanks to Wouter, Rosemarijn (my (2)52-bus buddies), and Johan for the chitchats. Thanks to Paul Merkx who recorded the database discussed in this thesis.

I also had train buddies. Thanks to Esther Janse for the talks about work and random topics; these talks made the train journey far less boring. Iwan de Kok made the train journeys in the last few months of my PhD-time less boring, thanks for that (and also for the games of ping-pong).

I appreciated the continued cooperation with Helmer Strik, Catia Cucchiarini, Febe de Wet, and Ambra Neri, even after I finished my internship. You were the first who introduced me to practising science, thank you for that.

Another collaboration which I have enjoyed was with Theresa Wilson and Stephan Raaijmakers. I also enjoyed the talks with Stephan about life and work and everything else.

At the University of Twente, I would like to thank Mannes, Boris, and Dirk, for their cooperation and the conversations. Ronald, Marijn, and Dennis provided me tips and help on (administrative) PhD and dissertation stuff, providing me all kinds of templates, and I also enjoyed talking to them about random topics. Ronald is also thanked for his humor and jokes. Charlotte and Alice are thanked for their

administrative support.

If it weren't for Johan van Balken, I would have had a very boring cover. I thank him for his time and effort spent on designing and illustrating this amazing cover.

I have a few more thankyous left. My family is thanked for their continuous support. My siblings Phuong, Cuong, and Tuyet are the best. I am grateful to my brave parents, Truong Phuc and Luu Phuoc Nga, who have traveled a long way to make this all possible.

Khiet Truong

張潔芳

Apeldoorn, July 2009

Contents

Chapter 1

Introduction

From Terminator 2 (1991):

———

The Terminator: "Why do you cry?"
John Connor: "You mean people?"
The Terminator: "Yes."
John Connor: "I don't know. We just cry. You know, when it hurts."
The Terminator: "Pain causes it?"
John Connor: "No, it's when there's nothing wrong with you, but you cry anyway. You get it?"
The Terminator: "No."

———

In the dialogue displayed above, the Terminator, a cyborg from the future, talks to a human. This cyborg appears to have acquired natural language processing skills and therefore is very human-like: it produces grammatically correct sentences and it reacts coherently to the human's utterances. However, the Terminator is not completely indiscernible from humans because one of the elements that it still lacks is emotional intelligence: the cyborg does not seem to understand why people cry. This is where affective computing can step in to make the cyborg emotionally intelligent. Affective computing is a relatively young multidisciplinary research area where disciplines like psychology, speech technology, computer vision, and machine learning meet. Psychology provides us ways to describe, model, understand and regulate emotions. Speech technology, computer vision and machine learning provides us methods to recognize and synthesize vocal and facial expressions. In addition to vocal and facial expressions, affect can also be expressed and measured through gestures or physiological measures like heart rate or respiration rate. Although affective computing is a relatively broad research area that is the interface between affect modeling and technology, and although affect can be expressed and measured through multiple modalities, we narrow our focus to the automatic recognition of affect in speech.

In this Chapter, we explain the basic 'ingredients' that are needed to develop speech-based affect recognition systems. First, in Section 1.1, we explain how affec-

tive computing is becoming increasingly important in people's lives, and we motivate our choice to focus on affective speech analysis (rather than e.g., analysis of physiological measurements). In Section 1.2, we describe some popular theories and models of emotion. We identify and describe challenges in Section 1.3 that one can encounter when one would like to develop affect recognition systems. Finally, we formulate our research questions in Section 1.4 and we give an outline of the content of this thesis.

1.1 Motivation for speech-based affect recognition

1.1.1 Affective Computing

Affective computing can be defined as a research area that aims at designing and developing systems that can **recognize**, **interpret** and **synthesize** human emotional states. Why would one want to develop these systems? It is an undeniable fact that computers are becoming increasingly embedded in our daily life. Technology is everywhere and one needs to **interact** with it. Affective computing can enhance the ways people interact with technology. For example, the way people play video games has evolved from sitting behind a computer screen or TV to standing or dancing or playing tennis in front of the TV. Imagine how the gaming experience could be enhanced when the gameplay is adapted to one's emotional state? Emotion recognition can add a new dimension in multimedia content analysis. Movies or TV broadcasts can be searched by types or various levels of emotion, such as excitement. In computer-aided learning, an affective component can help to maintain or increase the student's motivation. For instance, when the virtual tutor detects frustration with the student, the virtual tutor can give the student encouraging comments or it can slow down the pace. And if the virtual tutor detects that a student is getting bored, it can challenge the student by bringing up more complex exercises. Decision-making systems can improve their decision-making processes when emotional states are taken into account. For example, a system can decide to allocate fewer tasks to an operator who is recognized as being in stress. Interaction with machines, robots or spoken dialog systems in call centers, will feel much more natural and will be much more effective if human emotions can be recognized. Some research communities aim at developing humanoid robots that must have human-like capabilities such as emotion recognition and synthesis (unlike The Terminator who does not understand what causes the human to cry). Emotion recognition can also be employed in call centers for monitoring purposes: if the emotion recognition system recognizes an angry caller, the system can decide to route this caller to a more friendly and cooperative human employee. One of the most well-known examples of emotion recognition is that of an 'affective mirror' as proposed by Rosalind Picard [134]. This 'affective mirror' would be 'an agent that interacts with a person, helping him/her to see how he/she appears to others in various situations', and can be used to practice job interviews or presentations. In addition to these application-oriented contributions that open up many more research (and business) opportunities, research in affective computing also contributes to a better understanding of how emotion is produced and perceived by humans. It is clear that with the increasing amount of computers and technology embedded in our daily life, the need for a more **a/e**ffective and natural way of interaction increases.

1.1.2 Affect in speech

Vocal expressions, facial expressions, gestures, body postures and the ANS (autonomic nervous system, e.g., heart rate, diameter pupil, respiration rate etc.) are all ways of means through which emotions can be expressed and measured. The way these multiple modalities interact with each other is not yet clearly understood. A well-known study by Mehrabian [116] is an example of how multiple modalities can interact with each other. Mehrabian investigated the relative importance of verbal and nonverbal messages in expressing feelings and attitudes. He states that there are three elements in face-to-face communication: *words*, *tone of voice*, and *body language*. According to Mehrabian, each element has its relative importance in determining how *likeable* the person is who expresses his/her feelings:

$$\text{Total Liking} = 7\%\ \text{Verbal Liking} + 38\%\ \text{Vocal Liking} + 55\%\ \text{Facial Liking}$$

However, this rule has only been validated in specific situations. Many researchers have misinterpreted this rule by generalizing it to *all* situations. The rule is only valid when the verbal and non-verbal communications are *incongruent*. An example of incongruent verbal and non-verbal communication is:

Verbal:	"It's OK, I don't mind!"
Non-verbal:	avoids eye contact, looks anxious etc.

Only in cases where the communication is incongruent, the receiver of the message is more likely to trust the non-verbal message. Hence, in all other communications (that are not incongruent), the interaction between verbal and non-verbal communication is not understood yet.

Although emotion can be measured and expressed through many different modalities, this thesis focuses on the vocal channel of emotion expression. One of the main reasons for choosing speech is that speech measurements can be made in a relatively unobtrusive way. Attaining physiological measurements, such as heart rate or EEG signals, usually requires more effort and is usually more obtrusive for the subject, although nowadays, wearable measuring equipment is available which reduces the amount of effort and obtrusiveness. Secondly, speaking is a very natural way of interaction. Speech-enabled interaction will become increasingly important as the number of multitasking processes in daily life increases (e.g., making telephone calls while driving), and as interest in (humanoid) robots grows steadily. The third reason that we focus on speech is because we are interested in speech as an information carrier. Affect is only one of the types of information that is 'hidden' in speech. In addition to the verbal content, i.e., the words that are spoken, speech carries a lot of other (meta) information that helps the receiver (i.e., the listener) to decode what the message is that the sender (i.e., the speaker) wants to convey. Information that is 'hidden' in the voice of the speaker can tell the receiver something about the speaker's identity, the speaker's age, the speaker's gender, the speaker's regional accent or the speaker's emotional state. Technologies are being developed that enable the automatic extraction of these types of speaker information. For the recognition of the verbal content, **what** is said, *automatic speech recognition* systems (ASR) are available. Automatically recognizing **who** said something is undertaken in *speaker recognition*. Accent and language recognition involves finding out what foreign or regional accent, or language

this person speaks. In *emotion recognition*, the goal is to detect the emotion of the speaker: **how** something is said. These different types of speaker information are also referred to as **paralinguistic information**: all the non-verbal elements in speech that convey something about the speaker (e.g., laughter).

Prosody is considered the main (auditory) contributor to the conveyance of affect in speech (prosody can also be used for coding semantic and lexical information). Prosodic behavior in speech can usually be described in terms of speech characteristics such as rhythm, loudness, pitch, and tempo (Lexicon of Linguistics [1]). Other ways of expressing affect in speech are so-called 'affect bursts', see Scherer [161], Schröder [167]. As defined by Scherer [161], these are "very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events". Laughter, cries and sneezes are examples of affect bursts but verbal interjections like "Heaven!" are not. Although the emotional meaning of affect bursts may not be immediately apparent (laughter can have different types of meanings and functions), they have an important social, communicative and affective role in human conversation. The words chosen to communicate are obviously also cues to affect in speech. However, the main focus in speech-based affective recognition has traditionally been on an acoustic analysis of affective speech, without taking into account the lexical content. One of the reasons is that for lexical analyses, a transcription of what is said is needed, obtained either manually or automatically which is a hard problem itself, and not always available. Further, the choice of words is to an extend domain-dependent.

1.2 Theory and models of emotion

One of the first things we do when we perform science, is defining things in order to create a consensual working space. However, the notorious question 'What are emotions?' gives rise to a wide range of possible answers. As Scherer [163] puts it nicely, one of the major problems in emotion research is "the lack of a consensual definition of emotion and of qualitatively different types of emotions". There is no generally accepted methodology for describing emotions, and hence, there is no agreed taxonomy of emotional states, although the literature does offer some inexhaustive, possible taxonomies that are relatively frequently used. One well-known structuring of emotions is a structuring along the temporal dimension, see Table 1.1. On this dimension, 'emotion' is on one end of the scale while 'attitude' and 'personality traits' are on the opposite end. Emotions that are relatively brief in duration and very distinctive are also referred to as 'full- blown' emotions. Examples of 'full-blown' emotions are the well-known 'basic, universal emotions', see Ekman [56]: Anger, Disgust, Fear, Happiness, Sadness, and Surprise.

Definitions of emotions are related to theories and models of emotion. We will shortly describe three theories and models of emotions that have been influential in emotion research (for a richer and more comprehensive description of emotion theories, the reader is referred to Scherer [162]):

Componential emotion theory: Scherer has proposed a componential model of emotion, see Scherer [160, 164]. A leading concept in these componential models is that emotions are regulated by a cognitive evaluation of eliciting events and sit-

Short description	Duration	Rapidity of change	Intensity
Emotion: relatively brief episode of synchronized responses by all or most organismic subsystems to the evaluation of an external or internal event as being of major significance (e.g., Anger, Sadness, Joy, Fear, Shame, Pride, Elation, Desperation)	+	+++	++→ +++
Mood: diffuse affect state, most pronounced as change in subjective feeling, of low intensity but relatively long duration, often without apparent cause (e.g., cheerful, gloomy, irritable, depressed)	++	++	+→++
Interpersonal stances: affective stance taken toward another person in a specific interaction, coloring the interpersonal exchange in that situation (e.g., distant, cold, warm, supportive, contemptuous)	+→++	++	+→++
Attitudes: relatively enduring, affectively colored beliefs, preferences, and predispositions toward objects or persons (e.g., liking, loving, hating, valuing, desiring)	++→+++	0→+	0 →++
Personality traits: emotionally laden, stable personality dispositions and behavior tendencies, typical for a person (e.g., nervous, anxious, reckless, hostile, envious, jealous)	+++	0	0→+

Table 1.1: *Affective states taxonomy adopted from Scherer [162],* 0 *indicates absence,* +++ *indicates highest degree,* → *indicates hypothetical range.*

uations. These evaluation processes determine the relevance of the event and its consequences: if the eliciting event is not relevant to the major concerns of the organism, then there is no need to be emotional. The patterning of the responses in different domains (e.g., physiology, expression) are determined by the outcome of these evaluation processes. Componential models thus aim at making the link between the elicitation of emotion and the response patterning more explicit. Scherer's component process model states that different emotions are produced by a sequence of cumulative stimulus evaluation or appraisal checks with emotion-specific outcome profiles. Moreover, the model assumes that there are as many different emotional states as there are differential patterns of appraisal results. One of the advantages of componential models is the emphasis on the variability of different emotional states that are produced by different appraisal events which presumably makes the emotion-voice relation testable by concrete hypotheses.

Discrete emotion model: One of the most popular description of emotion is based on the assumption that there is a small number of universal or fundamental discrete emotion categories. Most of the discrete emotion theories stem from Darwin ([47]) who observed that a large number of emotional phenomena are universal, and who placed strong emphasis on the expression of emotion in face, body and voice. Inspired by Darwin, psychologists like Tomkins ([183]) and Ekman ([57]), who were mainly working in the field of facial expressions, theoritized that there are a number of basic emotions that are characterized by very specific response patterns in physiology, and facial and vocal expressions as well. A well-known set of basic emotions is termed "the Big Six" which are Anger, Disgust, Fear, Joy, Sadness and Surprise. Major drawbacks of this model are that 1) usually, these archetypical 'basic' emotions are not very much part of everyday life emotions, and 2) the set of emotions is very small. The Big Six basic emotions are based on Ekman's observations that members of a Stone Age culture are able to recognize this list of emotions which suggests that there are at least some emotions that are universal.

Dimensional emotion model: Another model that has gained much attention in emotion research is the dimensional approach to emotion. Several 'flavors' of this approach are possible: some use 2 dimensions while others use 3 emotion dimensions, and some position emotions in a circular way. Wundt ([212]) was one of the first who suggested that emotional states can be mapped in a 2 or 3-dimensional space. He proposed that emotions can be positioned by three dimensions: pleasantness – unpleasantness, rest – activation, and relaxation – attention. In 1954, Schlosberg [165] derived three similar dimensions: pleasantness – unpleasantness, attention – rejection, and sleep – tension. Osgood et al. [129] showed that almost all (non-)linguistic concepts could be placed in a three dimensional space (positive – negative, active – passive, degree of power) with respect to their meaning. So, researchers seem to agree on the existence of 2 or 3 emotion dimensions along which emotion concepts can be described. Furthermore, there is evidence that emotion concepts are (mentally) placed in a circular order by people. Russell [153] showed that affective concepts fall in a circle where similar emotions lie close to each other while opposite emotions lie 180 degrees apart from each other in a two dimensional map (arousal-sleepiness, pleasure-displeasure): pleasure ($0°$), excitement ($45°$), arousal ($90°$), distress ($135°$), displeasure ($180°$), depression ($225°$), sleepiness ($270°$) and relaxation ($315°$), see Fig. 1.1. Plutchik [136, 138] also proposed a circular model of emotion in which emotions are conceptualized in a color wheel where similar emotions lie close together. He added a third dimension, intensity, such that the three dimensional emotion model is shaped like a cone, see Fig. 1.2.

A dimensional emotion model is attractive since it has the ability to cover a large amount of varied emotions in a relatively simple way. The first main emotion dimension is positive (pleasure) vs. negative (displeasure) and is also known as Valence (or Evaluation). The emphasis of emotion research has usually been on Valence: people are simply more interested in discriminating positive from

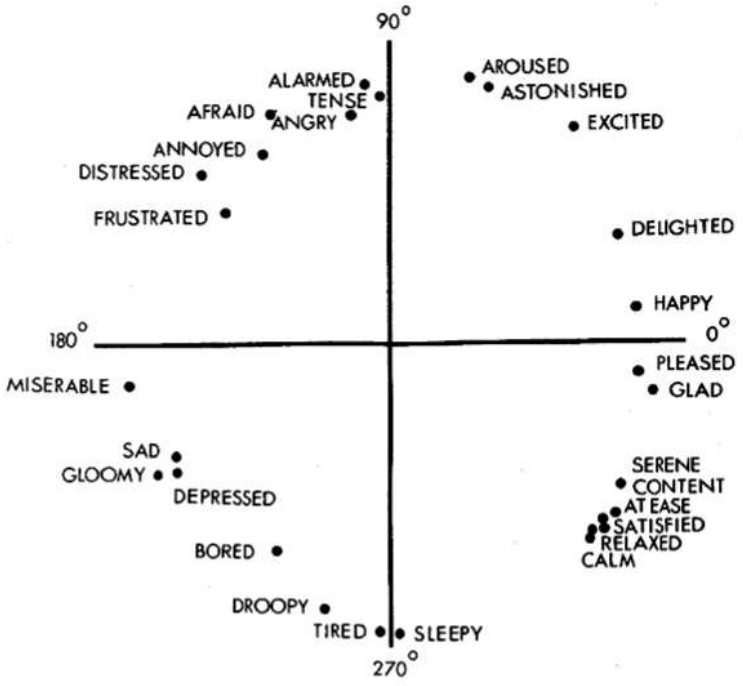


Figure 1.1: *The circular order of emotions as proposed by Russell [153] (figure adopted from Russell [153]).*

> negative emotions, e.g., detection of frustration with customers calling to a call center or detection of aggression in public environments. The second dimension is active (aroused) vs. passive (sleepy) and is also known as the Arousal dimension. For example, is this person bored or very excited? The third dimension represents a degree of power or control, e.g., dominance vs. submissiveness. In the literature, the Arousal and Valence dimensions are the most frequently used ones; mostly because most of the emotion concepts can be sufficiently described in terms of Arousal and Valence.

We have described three emotion theories and models that are relatively frequently adopted by the affective computing community. In our research, we will mostly work with discrete emotion categories and a dimensional model of emotion. As we will be tackling and discussing a broad range of various types of discrete emotion categories and emotion dimensions, we will view emotion in this thesis as a very broad concept. As a working definition for 'emotion' throughout this work, the following view on emotion that is stated in Cowie and Schröder [44] and the technical annex of the HUMAINE project [1] (an EU-funded network of excellence) is retained. Emotion, in this thesis, is considered

> in an inclusive sense rather than in the narrow sense of episodes where a strong rush of feeling briefly dominates a person's awareness . . . emotion in the broad sense pervades human communication and cognition. Human beings have positive or negative feelings about most things, people,

[1] http://emotion-research.net/projects/humaine/aboutHUMAINE/ technical_annex_public.pdf

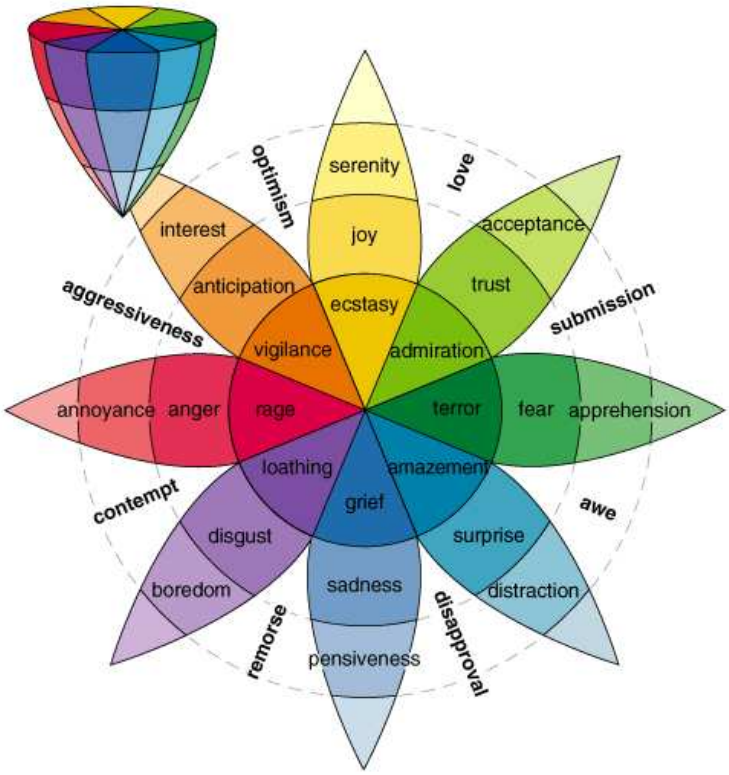


Figure 1.2: *Plutchik's circular model of emotion (figure adopted from Plutchik [138]).*

events and symbols. These feelings strongly influence the way they attend, behave, plan, learn and select.

Terms like 'affect' or 'emotional state' will be interchangeably used to refer to 'emotion' in its broader sense.

1.3 Challenges in speech-based affect recognition

In this Section, challenges that one can encounter in the development of a speech-based affect recognizer are identified. The challenges are divided into three development phases of an affect analyzer: data acquisition and annotation, feature extraction and learning, and performance evaluation.

1.3.1 The development phases of speech-based affect recognizers

For the development of (speech-based) affect recognizers, roughly three phases can be distinguished. Fig. 1.3 summarizes the development in a scheme. The first phase deals with data acquisition and annotation. It is not sufficient to have the data alone, the data also needs labeling: what emotion is associated with this particular speech signal? The second phase deals with feature extraction and model learning: the speech signals need to be described in terms of speech features that serve as input

for the learning algorithm. A (machine) learning algorithm must be chosen that can learn the mapping between the features and the emotion classes. And finally, in the third phase, in order to find out how good this mapping works, the recognizer needs to be evaluated in a proper way.

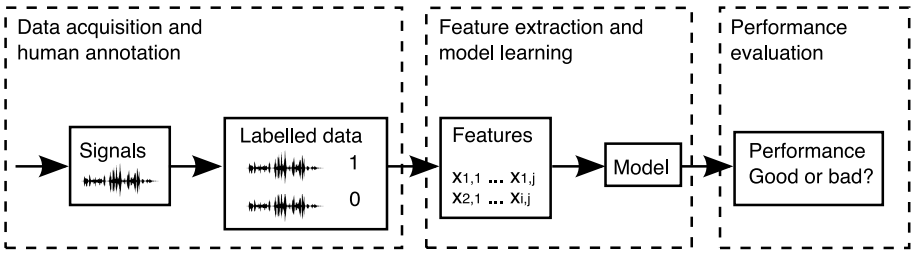


Figure 1.3: *The development phases of an affect recognizer.*

This chain of development shown in Fig. 1.3 looks straightforward. However, in each phase there are challenges and issues identifiable that need further attention and discussion.

1.3.2 Challenges in data acquisition and annotation

The machine learning techniques used to train models to recognize emotion require a lot of labeled data. To give an idea of how much data sometimes is needed to train a system: a speaker or language recognition system is usually trained with hundreds of hours of speech data. Labeled emotional speech data is sparse, which is a notorious problem in the emotion research community: there is a lack of annotated spontaneous emotional speech data. Filling this shortage of natural emotional speech data with acted emotional speech data is somewhat dangerous since several studies have shown that there are (large) differences between acted and natural emotional speech, e.g., Wilting et al. [210], and it decreases the ecological validity of the study. However, to an extent, the use of acted emotional speech can be supported by arguing that natural emotions are to a certain extent portrayals of emotions that are expressed in a controlled manner, so the question can be reversed: how natural are real-life emotional expressions (Banse and Scherer [12])?

Acquiring a substantial amount of spontaneous emotional speech data in the field has appeared to be a difficult process. A large percentage of real-life emotion situations occur in a social-interactive context in which people adhere to social conversational dialogue rules (e.g., Levinson [106]). Due to these implicit conversation rules, and due to the Observer's Paradox (the influence of the presence of the observer/investigator on the experiment, see Labov [100]), people suppress their emotions to a certain degree when they converse with each other while knowing that they are being recorded and observed. For example, Ekman [56] found in one of his experiments that Japanese people masked their negative expressions with a smile when a scientist sat with them as they watched films. Without the scientist sitting next to the subject, the masking was less frequent. As Ekman [56] suggests: "in private, innate expressions; in public, managed expressions". Furthermore, speaking

is a highly controlled and regulated process. Vocalizations that are less controlled are usually triggered by physiological changes that are caused by relatively extreme events. When we want to elicit such vocalizations, we should also consider ethics which is an aspect that must not be underestimated. For example, with respect to data acquisition and distribution, many people (e.g., companies, call centers) are reluctant to give away their data, even if it is for research purposes, because of privacy issues which is understandable but unfortunate for researchers.

When we have collected real-life, natural emotional speech data, the next challenge is to describe these naturally occurring emotions. It appears difficult to label naturally occurring emotions, especially when the context in which the emotional situation took place is unknown. In addition, the production and perception of emotion is to a certain degree person-dependent. Some people are intrinsically more expressive than others. Moreover, people disagree on the description and nature of the emotion perceived. One way to obtain reliable "ground truth" labels for emotional speech data, is to have multiple persons annotate parts of the data and to analyze how much people agree with other (inter-annotator agreement): when multiple annotators agree with each other on a specific label of a segment, then this can be considered more or less "ground truth". Intra-annotator agreement, the consistency/quality of the rater him/herself, may also play a role. Hence, post-processing the data recorded is a very time consuming and effort consuming process.

In short, we have identified some challenges to acquire natural emotional speech data that is suitable for the development of speech-based emotion recognizers:

- Due to suppression or masking of emotions in a natural social-interactive context, the emotions expressed are subtle and non-frequent.
- Natural, real-life emotions are difficult to label and may be mixed: there is no consensus on how to describe these emotions methodologically.
- The production and perception of emotion is mostly person-dependent which complicates the emotion annotation procedure and the development of a general affect recognition system.

1.3.3 Challenges in feature extraction and model learning

From the literature, it is clear that some acoustic features (e.g., F_0, energy, speech rate) are important for discrimination between emotions. Most of the features appear to correlate relatively well with the Arousal dimension: for example, according to our studies (see Chapter 4), Anger (=high Arousal) can be relatively well discriminated from Sadness (=low Arousal) acoustically. This is not the case with the Valence dimension: it appears that e.g., Anger and Joy are acoustically very easily confused with each other by emotion classifiers (e.g., Truong and van Leeuwen [189]). Although significant acoustic differences have been found between the expression of positive and negative emotions, in practice, these differences do not turn out to be predictive enough for automatic discrimination. Therefore, the strategy that is usually adopted is to extract as many features as possible from the speech signal and feed these features to an algorithm that selects the features that are highly discriminative.

In contrast with other research areas, such as ASR or facial expression recognition, in which well-established features and methods exist (e.g., Facial Action Coding System by Ekman and Friesen [58], Active Appearance Modeling by Cootes et al. [43]), the search in speech-based emotion recognition for a set of acoustic features in combination with an algorithm that achieves high performance, is still ongoing. In general, with the current set of features and algorithms, it appears difficult to capture subtle emotion expressions that are often encountered in natural emotional speech. Extreme emotions on the other hand, can be better discriminated from each other, at least when the extremes lie on the Arousal dimension.

In order to boost performance, multimodal approaches to emotion recognition have been employed and are becoming increasingly popular. Acoustic features are often combined with facial features, lexical features or other physiological features. How to combine and synchronize these different sources of information is an ongoing question and a research area on its own.

In short, some challenges in feature extraction and learning that can be encountered are the following:

- It is difficult to establish acoustic profiles for specific emotions.
- Discriminative acoustic features for Valence discrimination are hard to find.
- The speech features and technology commonly used have trouble recognizing subtle emotion expressions.

1.3.4 Challenges in performance evaluation

In many technologies, such as automatic speaker recognition and automatic language recognition, there exist international benchmark tests that enable the researchers to assess and compare the performances of their systems on an international level (carried out by the National Institute of Standards and Technology [2]). This is only possible when there are clear tasks, shared data sources and evaluation protocols defined and provided. For a relatively new research area such as speech-based emotion recognition, this does not exist yet. This is one of the reasons why it is difficult to read, compare and interpret the performances reported in the large number of studies, see also Table 2.4.

It is arguable whether the evaluation approach undertaken in the majority of the studies shown in Table 2.4 reflects the 'true' task of the emotion classifier. To what extent do the performance figures reflect the real performance of the targeted application when applied in the real-world? Emotion recognition is a multi-class classification task that can be approached in various ways. A lot of studies have used a relatively small set of basic emotions in their classification experiments. An example of a popular set of emotions is Anger, Disgust, Fear, Joy, Sadness, Boredom and Neutral. This emotion recognition problem can be approached as a classification task, conforming to the 'traditional' forced-choice classification evaluation paradigm. Given a sample, the task is to choose one of the emotion classes available: is it Anger or Disgust or Fear or etc. As Banse and Scherer [12] already suggested, since the number of emotion classes is small, this task does not really reflect *recognition* which

is what we actually want: it rather reflects *discrimination* between a small number of emotion classes. In addition, in such configuration, we should acknowledge that it is impossible to model each possible emotion. Hence, we should also acknowledge the possibility that in real-life, the emotion classifier can encounter 'new' emotions that have not been 'learned' by the emotion classifier. Associated with the traditional classification evaluation framework is the classification accuracy defined as the number of correctly classified cases defined by the total number of cases. While this performance figure is sensitive to skewed class distributions which make its interpretation non-transparent and less comparable, the classification accuracy is still often reported as a single main performance figure although alternatives are available.

In short, challenges involving performance evaluation of affect recognizers are the following:

- The lack of shared data sources and evaluation protocols makes it difficult to compare performances between studies.
- The current evaluation methodology can be improved in terms of soundness.

1.4 About this thesis

1.4.1 Goals and research questions

Traditionally, emotion recognition has been carried out with clean data that was acquired in a controlled way, meaning that acted emotional speech was used that usually contained extreme, basic universal emotions, i.e., *not-so-real* affect. These studies have formed the basis of the current emotion recognition research. However, it is clear that, in order to develop advanced affect recognition systems, the use of *real affect* is a must. Hence, the central aim in this thesis is the following:

> to develop speech-based affect recognition systems that can deal with *real affect*.

The challenges associated with the aim to develop speech-based affect recognition systems that can deal with real affect (described in Section 1.3) give rise to several interesting research questions that are answered in this thesis.

Researchers have come to realize that the gap between affect recognition in the lab and in the field is a significant one and that it is a problem that should be addressed. Hence, we designed our affect recognition experiments such that aspects of reality, naturalness and validity during all phases of development of our speech-based affect recognition systems are addressed. We believe that the link between the experimental setting, in which the affect recognition experiments are carried out, and the targeted affect application needs to be strengthened. This has some consequences for the way automatic affect recognition systems traditionally are developed.

We hypothesize that the character of the speech material available plays a leading role in the development of an affect recognizer, more than in other similar recognition technologies, such as e.g., language recognition. The naturalness and the intensity of the emotions expressed, and the way these expressions are annotated in the speech

data are all aspects that heavily influence the task and performance of the recognizer. Hence, we can formulate the following three research questions:

Research question 1 (RQ1): How does the speech data's level of naturalness used in speech-based affect recognition affect the task and performance of the recognizer?

Research question 2 (RQ2): How does the description and annotation of emotional speech data that is used in speech-based affect recognition, affect the task and performance of the recognizer?

Research question 3 (RQ3): What features and modeling techniques can best be used to automatically extract information from the speech signal about the speaker's emotional state?

Since affect is such a broad term, we have made decisions about what type of emotions to focus on. Firstly, to allow for comparison with previous studies, we performed emotion recognition on acted emotional speech data containing the six basic universal emotions (see Chapter 4). Using recognition technology and a detection framework adopted from related research areas such as language recognition, we show how basic, extreme emotions can be detected and discriminated from each other under fairly clean conditions.

Subsequently, we shifted towards the use of more natural affective speech data. For example, we have used speech data recorded during meetings and emotion data elicited from people who were playing a videogame. As a consequence, our focus has moved to the detection of non-verbal vocal expressions that are somehow related to affect. Laughter is such a non-verbal vocal expression. Until recently, the automatic detection of laughter has not gained much attention: in the ASR (automatic speech recognition) community for example, laughter was simply seen as non-speech that one should get rid of first. Our laughter study presented in Chapter 5 was one of the first studies that investigated the automatic detection of laughter in meetings in a systematic way, comparing several feature types and learning algorithms with the eventual goal to apply laughter detection in affective computing. In addition to laughter, we decided to focus on another emotionally colored phenomenon present in meetings, namely the recognition of sentiments and opinions (i.e., subjectivity). We assume that when people express their sentiments and opinions, people are more expressive (both vocally and textually) than when they express factual statements. Moreover, the recognition of subjectivity may help to identify so-called hot spots in meetings, which can be described as moments with increased involvement of multiple participants. Subjectivity recognition has traditionally been investigated on textual level. To the best of our knowledge, our experiments presented in Chapter 5 are one of the first to use both acoustic and textual features for the recognition of opinion clauses, and the polarity (positive or negative opinion) of these opinion clauses. Using combinations of these features, we show what the contribution of acoustic information can be to subjectivity and polarity recognition.

As an intermediate between emotions that are acted or natural, we used spontaneous material containing affective vocal and facial expressions that are elicited

through gaming. This is material that we have collected ourselves at TNO with the aims to 1) compare 'felt' (annotations from the subjects playing the game themselves) and 'perceived' emotion annotations (annotations from observers), 2) develop affect recognizers that can predict Arousal and Valence scalar values rather than emotion categories, and 3) compare human performance to machine performance. The effect of 'felt' vs. 'perceived' emotion annotations on the task and performance of an affect recognizer has previously not been investigated yet (to the best of our knowledge). One advantage of using separate Arousal and Valence scales is that recognizers for these emotion dimensions can be developed and optimized separately from each other. We used acoustic and lexical features for the prediction of Arousal and Valence, and compared their performances. The description of the spontaneous emotion material collected and the results of the prediction experiments and analyses are presented in Chapter 6.

The insights attained during the development of all these different types of recognizers, using speech material containing emotions ranging from acted, to elicited, to natural, provide answers to RQ1, RQ2, and RQ3.

Although the focus is on the use of spontaneous emotional speech material, there is one Chapter in this thesis that involves a speech database containing acted basic, universal emotions. These types of databases have been used frequently in the past, and many recognizers were developed with these databases. Main reasons for using acted emotional speech is that this type of material is much easier to acquire than spontaneous emotional speech data, and the emotion labeling is straightforward. However, one major objection against the use of acted and basic emotions is that the classification experiments performed with these datasets are not very representative of real-life situations; in other words, the ecological validity of these classification experiments is relatively low. Obviously, one partial solution is to use natural emotional speech data. That is exactly what we have done in Chapter 5 and Chapter 6. Alternatively, we can try to bridge the gap between lab and field emotion classification experiments by proposing more appropriate ways of evaluation that will better reflect real-life situations:

Research question 4: How can the current evaluation methodology for affect recognition in the lab be improved to match more closely the real-life, field situation in which affect occurs?

In contrast with other similar recognition technologies such as language recognition (given a speech sample, what is the language spoken?), the relatively young research area of emotion recognition (given a speech sample, what is the emotion?) does not seem to have a common evaluation framework. In Chapter 4, we show how the detection framework, that is commonly used in language recognition, can be adopted in emotion recognition. We will show that this framework offers many advantages which can make the traditional emotion classification experiments (slightly) more ecological valid. We also propose an 'open-set' detection evaluation methodology which addresses RQ4.

1.4.2 Outline

In the next Chapters, we describe several experiments that we have performed to investigate the research questions mentioned previously. All these experiments involve the development of speech-based affect recognition systems. First, in Chapter 2, we give an overview of past speech-based affect recognition studies and describe what data, features, methods, and evaluation metrics were frequently used in these studies. In addition, we provide an overview of all the materials and methods used in our current experiments.

Acquiring non-acted affective speech material is a well-known issue in affective computing research. In our studies (see Truong et al. [192]), we have undertaken efforts to acquire natural affective speech data in the field. We have tried to measure real affect in speech during emergency situations on a naval ship, during crisis meetings, and while people are playing a virtual reality game. In Chapter 3 (based on Truong et al. [192]), we describe what difficulties we have encountered (and the implications thereof) in our efforts to collect emotional speech data in the field.

Since labeled natural emotional speech data is sparse, it is very convenient to be able to use existing databases that contain acted emotional speech. Additional advantages are that we can relatively quickly and easily test new recognition technologies, we have few worries about the labeling of the emotions, and we can adopt techniques and evaluation procedures from similar recognition technologies such as automatic language recognition. In Chapter 4, we describe how we used state-of-the-art recognition technology to develop emotion detectors that can detect acted basic, universal emotions. One of the key elements in developing these detectors is that we adopt a detection framework which has not frequently been used in emotion recognition, but which offers many advantages over the classical classification paradigm that is traditionally used in emotion recognition. For example, within this detection framework, we have designed an 'open-set' evaluation that simulates an open-set situation (see Truong and van Leeuwen [188], van Leeuwen and Truong [195]), i.e., the possibility that the detector encounters new emotion categories that have not been 'seen' before by the detector (that were not included in its training set). The 'open-set' simulation was introduced with the goal to make the results of lab emotion classification experiments more representative of real-life situations.

It is commonly agreed that the use of acted emotional speech in affect recognition is very convenient, but it is not very ecologically valid. Hence, the experiments described in Chapter 5 and Chapter 6 involve natural emotional speech and elicited emotional speech respectively. In Chapter 5, we present detection experiments performed on spontaneous meeting data with the goal to detect emotionally colored behavior in meetings. In the first part of Chapter 5 (based on our work published in Truong and van Leeuwen [186, 187, 190]), we explain how we developed automatic laughter detectors. In the second part of Chapter 5 (published as Raaijmakers, Truong, and Wilson [143]), we explain how we developed detectors for the recognition of sentiment and opinions in meetings: we detect whether an utterance is subjective or not, and if it is subjective, whether it is positive or negative subjective (i.e., polarity detection).

As an intermediate between natural and acted emotions, we have also experi-

mented with emotional speech data that we elicited from people who were playing video games (see Merkx, Truong, and Neerincx [118]). Part of the data is annotated by the gamers themselves *and* observers. Emotion prediction experiments were carried out with this data to compare the use of self annotations to observers' annotations (see Truong et al. [191]), and to compare the use of acoustic and lexical features for Arousal and Valence recognition (see Truong and Raaijmakers [185]). Rather than to classify emotion categories, the detectors were developed to predict Arousal and Valence scalar values. The elicitation and recording procedures of this corpus, and the results of the emotion prediction experiments are presented in Chapter 6.

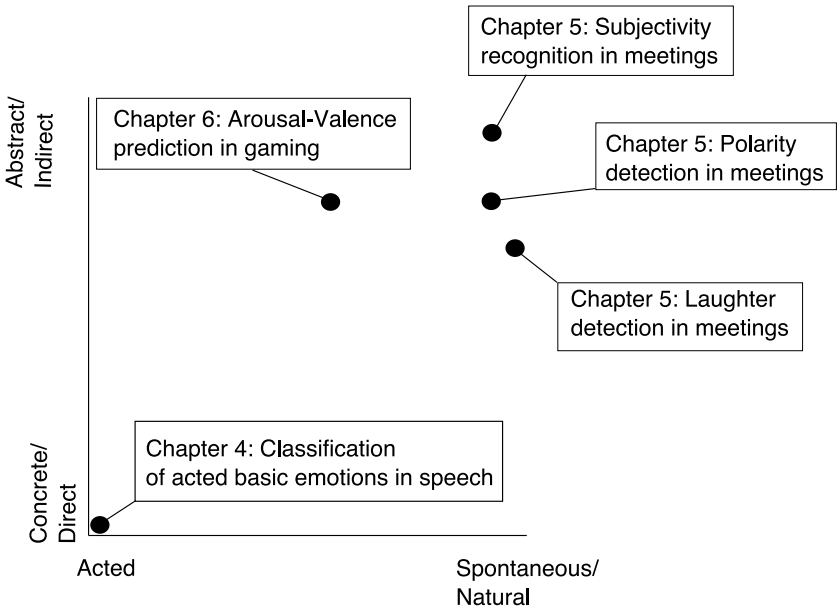


Figure 1.4: *Detection experiments described in this work.*

We can place all detection experiments that we performed along two scales. The first one ranges from acted data to spontaneous/natural data: we have performed detection experiments with acted, elicited and natural emotional speech data. The second one ranges from concrete/direct emotion modeling to abstract/indirect modeling. It seems as if we progress towards the use of natural emotion data, the modeling of emotion becomes more abstract: for instance, in using natural meeting speech data, the focus has shifted to the detection of subjectivity which can be linked to affective expressiveness, but it is not considered a specific emotion category. This also applies to laughter: the expression of laughter can be an affective event, but it is not always immediately clear what the emotional meaning is of that laughter event. When we place our detection experiments in a 2-dimensional plot, the chapters can be arranged as in Fig. 1.4.

Finally, in Chapter 7, we draw conclusions from the experiments performed and discuss these in the light of the research questions. Furthermore, we give recommendations for future research.

Chapter 2

Automatic affect recognition in speech: past and current affairs

In recent years, due to a growing interest for affective computing, an increased amount of literature has become available on the investigation of automatic emotion recognition (and synthesis) in speech. The first studies on emotional speech focused on finding acoustic correlates of emotional speech. Furthermore, also in the areas of psychology, researchers started to investigate the perception of emotion, and human's ability to recognize emotions in speech. Subsequently, with the rapid development of recognition technology, the first studies on automated analyses of emotional speech began to appear. In this Chapter, that is divided in two parts, we provide an introduction into the research area of automatic emotion recognition in speech, and we introduce the materials, methods, features and performance metrics used to develop our speech-based affect recognizers presented in this thesis. First, we carried out a literature study on past speech-based affect recognition studies. In Section 2.1, we describe some acoustic characteristics of emotional speech as found in past studies. In Section 2.2, we briefly describe how well humans can classify emotions in speech. An overview of past speech-based affect recognition studies is given in Section 2.3. Finally, in the second part of this Chapter, we give a description of the materials, methods, features, and performance metrics used in the current study.

2.1 Acoustic characteristics of emotional speech

Early studies on the acoustics of emotional speech originate from the seventies, carried out by Williams and Stevens [206, 207]. In Williams and Stevens [206], the emotional states of pilots during flight were studied. In Williams and Stevens [207], acoustic correlates of emotional speech, originating from actors and originating from a real-life situation were investigated and compared. The sound sample used in [207] is a good example (and one of the first) of a naturalistic emotional speech sample that is collected in the field. The sound sample is that of a radio announcer who was describing the landing of the Hindenburg zeppelin that suddenly burst into flames and crashed. The radio announcer, who witnessed the crash and sounded deeply affected, continued reporting. An acoustic analysis was carried out on this sample of emotional

speech (see Fig. 2.1). Among other acoustic parameters investigated, Williams and Stevens [207] concluded that the fundamental frequency (F_0) was the most important predictor of emotion.

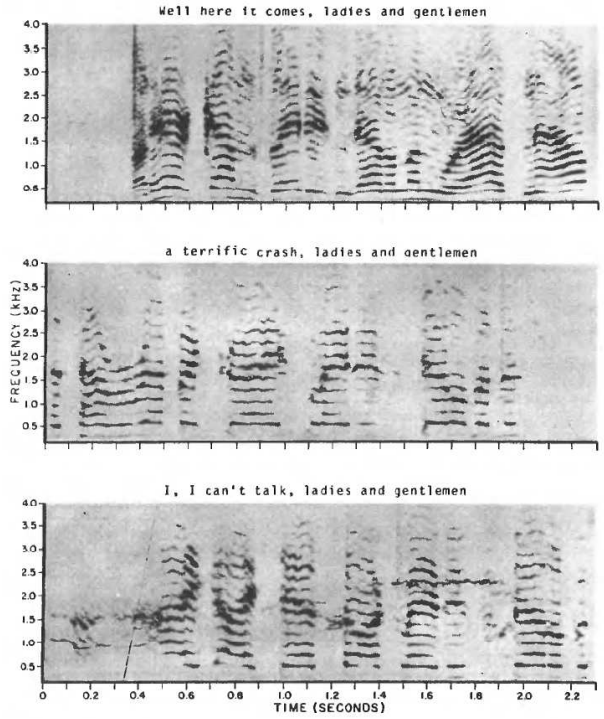


Figure 2.1: *Narrow-band spectrograms of the radio announcer's speech during his report on the Hindenburg crash (from Williams and Stevens [207]).*

Also in the seventies, Scherer and colleagues developed interests for the study of the relationship between personality and voice characteristics, and vocal expression of emotions. During that time, emotional speech researchers took up observations made in studies on facial expressions, mainly led by scientists like Tomkins [183], Friesen and Ekman [59, 56]. As a consequence, most of the classical emotional speech studies employed the popular "basic, universal emotion categories" rather than "a dimensional model" as suggested by Schlosberg [165]. The recognizability and generalizability of basic, universal emotions, and the relative easiness with which these emotion data could be portrayed and collected (by hiring actors) also contributed to the popular use of basic emotions in emotional speech research.

In general, the studies on the acoustics of discrete basic emotions (e.g., Banse and Scherer [12], Murray and Arnott [123]) seem to provide a consistent view, except for a few inconsistencies. Most inconsistencies may be contributed to differences in manifestations or portrayals of the basic emotion. For example, the acoustic characteristics of Anger described in Table 2.1 are associated with Hot Anger rather than Cold Anger; it is not always clear what type of Anger was used in a particular study. Although the studies seem to agree with each other, the evidence for these emotion-specific vocal patterns is not at all conclusive. In Banse and Scherer [12], three major causes are given for this observation, which affect the interpretation of these studies

and the development of speech-based affect recognizers: 1) most of the studies on the acoustics of emotional speech only employ a small, restricted set (3–6) of emotion classes, consequently, the acoustic descriptions are more likely to be specific to this set of emotions and contrastive with respect to each other rather than generic, 2) the limited number of acoustic parameters (F_0, energy) used in previous studies may have obscured the existence of other vocal profiles of emotions that manifest themselves through other acoustic parameters, 3) the atheoretical nature of much of the research makes cumulativeness of the empirical findings and hypotheses hard. These are valid points made in Banse and Scherer [12], which are gradually being taken up by researchers.

	Anger (Hot)	Sadness	Joy	Fear	Disgust
Speech rate	+	-	+/-	++	- -
Pitch average	+++	-	++	+++	- - -
Pitch range	++	-	++	++	+
Intensity	++	- -	++	=	- -
High-frequency energy	++	-	+	++	+

Table 2.1: *Acoustic characteristics for some basic emotions (partly adopted from Ververidis and Kotropoulos [200], Murray and Arnott [123], Scherer [159]).*

Table 2.1 summarizes the behavior of some frequently used acoustic features for a number of discrete basic emotion categories. The summary shows that Anger and Sadness are very distinct emotions, while Anger and Joy and Fear appear to be very similar acoustically.

In a dimensional approach to emotion, statements on acoustic profiles of emotions can be made in a broader and generic context namely in terms of the two or three emotion dimensions Arousal, Valence and Dominance. Murray and Arnott [123] noted that the Arousal dimension is correlated with the auditory variables which implies that the activity of emotional meaning can be carried by the relatively simpler acoustic parameters of F_0 and energy. Many of the studies using 'traditional' acoustic features such as F_0, energy, duration and speech rate, have found that these features are characteristic for emotions that differ in Arousal level, for instance Anger vs. Sadness. Valence, on the other hand, is probably communicated through much more subtle and complex vocal patterns and parameters that are less auditory evident and measurable. Emotions that differ on the Valence scale, for instance, Anger vs. Happy, may be more characterized by source and articulation characteristics which manifest themselves in voice quality (e.g., creaky, harshness, breathy) and spectral features (e.g., formants, MFCCs, energy distribution in spectrum). In the literature, it is agreed upon that the usual acoustic variables investigated show indeed stronger correlations with the Arousal dimension than the Valence dimension, e.g., Banse and Scherer [12], Scherer [159, 163], Ververidis and Kotropoulos [200], Schröder et al. [169].

Whether the findings here about the acoustic characteristics of emotional speech are also valid in spontaneous emotional speech, remains debatable. The acoustic characteristics partly seem to overlap, however, several studies have found indications

that there are indeed significant differences in the acoustics of acted vs. spontaneous emotional speech. Wilting et al. [210] have found differences in the production and perception of acted vs. spontaneous speech which may also reflect in the acoustics. Vogt and André [203] compared feature sets for acted and spontaneous speech. They found, by performing feature selection, that for acted speech, pitch-related features and pauses are very important, whereas for spontaneous speech, Mel-Frequency Cepstrum Coefficients were most important. In addition, they found that there was few overlap between the feature sets of acted and spontaneous speech. In Schaeffler et al. [158], vocal parameters in spontaneous and posed child-directed speech was investigated. It appeared that voice quality parameters are more used in mothers' child-directed speech (presumably spontaneous affective speech) than in speech from non-mothers directed to imaginary children (presumably acted affective speech), although it remains unclear whether the factor mother or non-mother may have also played a role. These studies have shown that there are indeed important acoustic differences between acted and spontaneous speech.

Note that the acoustic characteristics of emotional speech have also been investigated from a speech synthesis view, e.g., Schröder [166], Murray and Arnott [123]. However, there is no one-to-one mapping between emotional speech synthesis features and emotional speech recognition features, although there are similar modeling difficulties. For example, Valence *also* appears to be difficult to convey in synthetic speech (Schröder [168]).

2.2 Human classification of emotions in speech

Prior to the rise of *machine* classification of emotions in speech, it was investigated by e.g., Banse and Scherer [12], Van Bezooijen [193] how good *humans* can recognize emotions in speech. These studies actually involve *discrimination* rather than *recognition*: the subject is usually forced to choose between a relatively small number of emotion classes. Furthermore, subjects are usually asked to classify acted, discrete, basic emotions. One large study on the human perception of Dutch emotional speech was carried out by Van Bezooijen [193] in 1984. In a forced choice perception experiment, Dutch subjects were asked to classify the acoustic emotional stimuli (produced by actors) into one of the 10 discrete emotion categories. The stimuli consisted of Dutch sentences that were produced in different emotions. Banse and Scherer [12] used a larger number of emotion categories, namely 14, and carried out a similar perception experiment with German listeners. The carrier sentences were two meaningless, nonsense utterances that were composed of phonemes of several Indo-European languages. Burkhardt et al. [25] used 7 discrete emotion categories in his perception experiment, and offered German emotional utterances to German listeners. In Figure 2.2, the recognition rates of these three human recognition studies are plotted against each other to see whether there is agreement among several studies on the recognizability of various emotions. From this figure, it can be seen that there is indeed a common trend visible between the studies: of all the emotions offered, Disgust and Shame are worst recognized by humans whereas (Hot) Anger is best recognized.

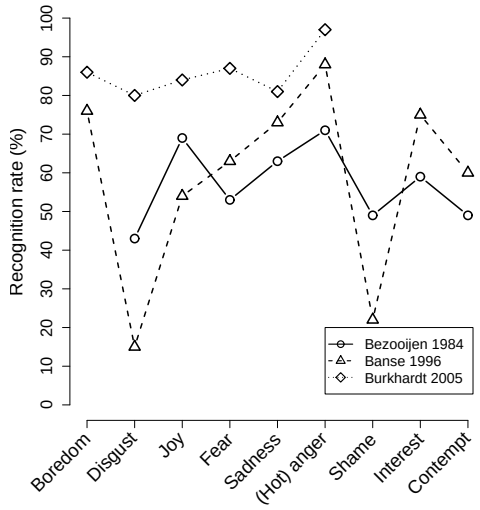
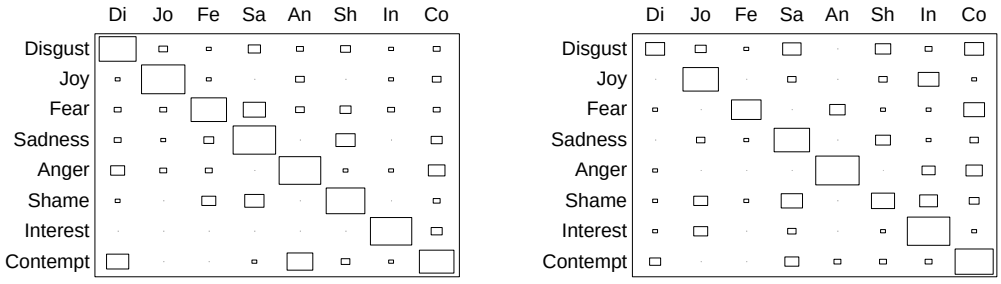


Figure 2.2: *Recognition rates (%) of human recognition experiments of emotions in speech - comparison between several studies.*



(a) *Van Bezooijen study [193].*

(b) *Banse and Scherer study [12].*

Figure 2.3: *In these $r \times c$ matrices, row r is classified as column c: the larger the square, the higher the recognition rate.*

In Figure 2.3, erroneous confusions between emotions made by humans are shown. Humans appear to be very good in discriminating between basic emotions that lie on opposite sides of the Arousal and Valence dimensions: it can be seen from Fig. 2.3 that Anger and Joy are seldom mistaken for each other, and Anger and Sadness are never confused with each other. The rest of the erroneous confusions do not seem to show a pattern.

2.3 Machine classification of emotions in speech

In the studies on acoustic correlates of emotional speech and human perception of emotion, the basis was laid to pursue automatic classification of emotions in speech. From the nineties on, a large number of automatic speech-based emotion classifica-

tion studies have been carried out. Given the amount of variation between these studies along various dimensions, it is difficult to develop a concise and consistent view about the state of affairs in this research area. However, there are certain developments visible that are geared towards a more consistent view and approach to speech-based emotion classification. In Table 2.4, a brief summary of several speech-based emotion classification studies is given. Each study can be characterized by a number of 'parameters' within each development process that can vary between emotion recognition studies as shown in Fig. 1.3 and Table 2.2.

Development process	**Parameters**	**Examples**
Data acquisition and annotation	Nature of data	acted, WOZ, spontaneous
	Number of speakers	
	Number of emotion classes	
	Type of emotion/annotation	discrete categories, dimensions, basic emotions
Feature extraction	Unit of analysis	phoneme, syllable, word, utterance
	Short-term ASR spectral	Mel-Frequency Cepstrum Coefficients, (Rasta-) Perceptual Linear Prediction
	Other (long-term)	pitch-related, energy-related, energy in spectrum-related, voice quality
Learning	Probability density function (pdf) modeling	Gaussian Mixture Models, Hidden Markov Models
	Kernel methods	Support Vector Machine, Support Vector Regression
	Other	Neural Networks, Decision Trees, Boosting, K-Nearest Neighbor
Evaluation	Protocol	K-fold-cross validation, person dependent/independent, detection, classification
	Metrics	classification accuracy, F_1, Equal Error Rate, Cost of Detection

Table 2.2: *Variations along several parameters in emotion recognition studies.*

2.3.1 Data acquisition and annotation

As Table 2.2 and Fig. 1.3 show, the first development process is that of data acquisition and description. Data acquisition and description can be varied along several parameters. The first parameter involves **the nature of the data**. With the nature of data,

we mean whether the data is acted or spontaneous. One of the problems in emotional speech research is the lack of annotated, natural emotional speech data. Hence, we can observe from Table 2.4 that most of the studies have used acted (posed) emotional speech data, e.g., Banse and Scherer [12], Petrushin [133], Tato et al. [182], Nwe et al. [125], Yacoub et al. [214], Ververidis and Kotropoulos [199], Schuller et al. [171], Clavel et al. [40], Schuller and Rigoll [170], Hu et al. [81], Vlasenko et al. [202]. For these type of databases, actors (or non-actors) are hired to act out a given emotion and utter the same sentences in various (usually basic) emotions. Clearly, there are advantages to the use of acted emotional speech when one wants to perform automatic classification experiments. Using actors offers a quick way to collect emotional speech uttered by various speakers in various emotions under controlled conditions. The speech signal is clean and the amount of work needed to post-process the data and signals is relatively small, e.g., no segmentation and emotion annotation is required. It can be useful to perform a rating (perception) study to verify the naturalness and recognizability of the emotions expressed by the actors. A major drawback of using acted emotional speech is that it, to a certain extent, lacks ecological validity: emotions acted out by actors do not per se reflect natural emotions that occur in real-life. Acted emotional speech tend to sound more exaggerated and less natural to listeners. Several studies have proven that there are significant differences between acted and spontaneous emotional speech, e.g., Vogt and André [203], Wilting et al. [210], Schaeffler et al. [158]. However, acquiring natural emotional speech data suitable for machine classification is a complex and very time- and effort consuming process. In addition, the quality of the signal recorded can be degraded and in general, there are less controllable parameters. Localizing and labeling the emotion in the speech data recorded involves a lot of human labor. Despite these factors, an increasing number of researchers is undertaking efforts to acquire emotion data in a natural environment which is also reflected in the growing number of classification studies using natural emotional speech data, see e.g., Fernandez and Picard [62], Vidrascu and Devillers [201], Devillers and Vidrascu [52], Neiberg et al. [124], Graciarena et al. [67], Truong and van Leeuwen [187], see Table 2.4.

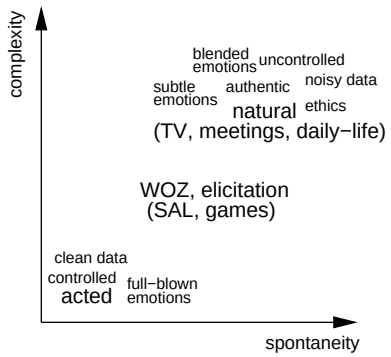


Figure 2.4: *Tension between acted and natural emotional speech.*

As an intermediate, elicitation and Wizard-Of-Oz methods (WOZ) can be used to collect (semi-)spontaneous emotional speech. Elicitation methods include watching movie clips (e.g., Lang [102]), listening to music (e.g., Wagner et al. [204]), and

playing games (e.g., Kim et al. [91], Johnstone [88], Yildirim et al. [216], Merkx et al. [118], Truong et al. [191]). Wizard-Of-Oz methods include interaction with a virtual character (e.g., Cox [46]) or spoken dialog system, see e.g., Ang et al. [6], Batliner et al. [16, 15]. In Fig. 2.4, a graph is plotted that reflects the tension between the use of spontaneous material and the level of complexity. In Table 2.3, a number of spontaneous emotional speech databases is listed.

Database	**Number speakers**	**Nature data**	**Types of emotions**	**Description, annotation of emotion data**
Belfast Natural (Douglas-Cowie et al. [53])	31m, 94f (English)	natural	wide range of affective speech, neutral, angry, sad, pleased, happy, amused, worried	clips taken from television chatshows, current affairs programs, and interviews conducted by research team, unscripted interactive discourse, annotation with Feeltrace and in categories
SAL (pilot database, Cox [46])	20 subjects (English)	semi-natural	wide range of emotion related states, not very intense	interaction with virtual characters, each of whom have different personalities, interactive discourse, annotation with Feeltrace
Smartkom (Steininger et al. [181])	45 subjects (German)	semi-natural	joy, gratification, anger, irritation, helplessness, pondering, reflecting, surprise, neutral	human-machine WOZ dialogues, solving tasks with system, interactive discourse, annotation in categories and intensity (weak, strong)
AIBO (Batliner et al. [17])	81 children (51 German, 30 English)	semi-natural	joyful, surprised, emphatic, helpless, angry, motherese	human-robot interaction (robot pet), annotation in categories
SUSAS (Hansen and Bou-Ghazale [75])	13f, 19m (English)	acted, natural	anger, stress (fear, anxiety), task load stress	partly read speech, speech recorded during task executions, or rollercoaster rides in amusement park
Vera am Mittag (Grimm et al. [70])	44m, 60f (German)	natural	wide range of primarily negative and neutral emotions, fewer positive emotions	clips from talk shows, annotation on Arousal, Valence and Dominance scales with Self-Assessment-Manikin (SAM)

Table 2.3: *Various spontaneous emotional speech databases that have frequently been used by the research community.*

The second and third parameters that can be varied within the development process of data acquisition and description is that of **number and types of emotion**. It can be seen from Table 2.4 that the number of emotions in these studies varies from two to fourteen; the majority of the studies have two to seven emotion classes which

is relatively small. Also, the majority of the studies covers discrete stereotypical 'basic' emotions, e.g., Tato et al. [182], Nwe et al. [125], Ververidis and Kotropoulos [199], Schuller et al. [171], Datcu and Rothkrantz [48], Hu et al. [81], Vlasenko et al. [202]. This is one of the points of criticism that is brought forward by Banse and Scherer [12]:

> It is doubtful whether studies using 4–6 response alternatives in a vocal emotion recognition study actually recognition or whether, more likely, the psychological process involved is *discrimination* among a small number of alternatives . . . Obviously, real life requires true emotion recognition rather than emotion discrimination . . . In consequence, the ecological validity of recognition rates can be expected to increase with the number of alternatives.

The quote above also applies to automatic emotion classification. The performance of the emotion classifier depends on the number and types of emotion classes used in the classification experiment, so the performance figures should be read and interpreted with care. For example, an emotion classifier that achieves 87% accuracy in a three-class classification problem with Anger, Neutral, and Sadness, can give a somewhat flattering picture of the situation, e.g., in Yacoub et al. [214], the 87% accuracy dropped when a fourth class Happiness was added.

Reasons for classifying 'basic, universal' emotions are simply that the data is relatively easily acquired, and/or that the data is publicly available. The choice for emotions to be classified seems to be therefore somewhat data-driven rather than application-driven. Hence, in some cases, from an application point of view, the emotion classification experiment performed does not make much sense. For example, a forced choice classification experiment between Anger, Sadness, Disgust and Happiness does not seem to provide very useful information about Angriness detection if the targeted application is an Angriness detector in a call center environment since 1) Sadness and Disgust are probably not frequently encountered in such environments, and 2) we are less interested in how Anger can be discriminated from Sadness, Disgust or Happiness, but more interested in how Anger can be discriminated from Not-Anger.

In a dimensional approach to emotion recognition the number of emotion classes is usually reduced to 3–5, characterized by the two emotion dimensions or the "four quadrants" in the Arousal-Valence space. The four quadrants are usually called Positive-Active (PA), Positive-Passive (PP), Negative-Active (NA), and Negative-Passive (NP). Discrimination can also be performed along the dimensions: Positive vs. Negative, and Active vs. Passive. More recently, Grimm et al. [69, 68] predicted emotion on continuous scales of Arousal and Valence. Using Support Vector Regression, Grimm et al. [69] outputs scalar values on Arousal and Valence scales without the use of discrete categories. The advantage of such an approach is that language-dependent emotion labels that describe the emotion categories have become superfluous.

Naturally, the choice for certain types of emotions depends on the targeted application in mind. The affect recognizers developed in Petrushin [133], Yacoub et al. [214], Vidrascu and Devillers [201], Devillers and Vidrascu [52] aim to detect emotions of Agitation, Frustration, Anger, and Fear in medical, financial and customer-service call centers. Fear is also detected in the context of surveillance and safety

(Clavel et al. [40]). Frustration, Annoyance and Anger can also occur when spoken dialogue systems that provide a service to a customer fail to fulfill the need of the customer, e.g., Batliner et al. [16, 15], Ang et al. [6], or when students get upset and frustrated because a spoken-tutoring-dialogue system fails to understand and help them, e.g., Liscombe et al. [108], Hua et al. [83], Litman and Forbes-Riley [109]. Stress has been extensively investigated by Zhou et al. [219], Fernandez and Picard [62], Kwon et al. [99] in the context of car-driving and pilots: in environments where critical situations are likely to occur, stress can be useful to detect. Some 'exotic' emotions such as motherese (i.e., child directed speech) and emphatic (Batliner et al. [20], Kwon et al. [99]), deceptive speech (Graciarena et al. [67]), depressed, suicidal speech (Yingthawornsuk et al. [217]), and fatigue, sleepiness in speech (Krajewski and Kröger [96]) have also been addressed. Hotspot detection for meeting summarization and/or meeting browsing (i.e., localization of events with a high level of activity in a meeting) has recently gained interest (Neiberg et al. [124], Wrede and Shriberg [211]). Salway and Graham [157], Hanjalic and Xu [74] are modeling emotion in the context of information-retrieval like applications, such as movie browsing. In general, negative emotions receive more attention from researchers than positive emotions; there simply is a greater need for applications that detect negative emotions than positive emotions. However, we would also like to mention a few studies that have analyzed non-negative emotional speech. Rosenberg and Hirschberg [151] investigated the acoustics of charismatic speech by analyzing presidential speech, and Chen et al. [38] investigated the acoustics of friendly speech.

Finally, the fourth parameter in data acquisition and description is that of **number of speakers**. The number of speakers partly determines how generizable the results of the classification experiments are. The more speakers, the better generalizable the results will be. More importantly, when classification experiments are performed speaker-independently, the generalizability of the results also increases.

2.3.2 Feature extraction

In feature extraction, an important parameter along which one can vary is the **unit of analysis** for feature analysis. The choice for a certain unit of analysis is related to the choice for the type of speech features: **spectral** or **prosody-oriented** features.

Since it is not clear yet which acoustic features describe what emotions best, most researchers use a rather crude, though effective strategy to feature extraction: a large number of acoustic features ($>$1000) is extracted and a feature selection algorithm is applied to reduce this set to a smaller number of most powerful features. From the literature, we know that prosodic features, including F_0 (pitch), energy, speech rate, and the distribution of energy in the spectrum are among the most important features for emotion classification. These types of features are usually measured *suprasegmentally*, that is, measured over a unit larger than a phoneme. Prosody usually occurs in a hierarchy of higher levels of an utterance (although prosodic analysis at phoneme-level does exist). Short-term (e.g., each 16 ms) prosody measurements are usually not meaningful, since it is the prosodic behavior over time that is important for emotion recognition. This means that the unit of analysis (or analysis window) in prosodic analyses can be as large as a syllable, or a word, or a sentence. However, note that lin-

guistic units such as syllables and words require an ASR system that can deliver these units' boundaries. Usually statistical measures such as mean, standard deviation, the range etc. are calculated from these series of measurements and used as input features. Examples of various levels of measurements of prosody can be given when prosody is used in a linguistic sense rather than in a paralinguistic sense. Prosody is not only utilized for the expression of paralinguistic information, it also has a linguistic function (which makes affect recognition even more complex and difficult). For example, in some languages, like Dutch, different syllables in words can be stressed to mark differences in meaning. In the Dutch minimal stress-pair, 'kaNON vs. KAnon', lexical stress makes a meaningful distinction between the words 'cannon' and 'canon'. Stress on word-level may mark given information vs. new information in a sentence: 'MARY bought a book' ('book' is given information, 'mary' is new information) vs. 'Mary bought a BOOK' ('mary' is given information, 'book' is new information).

On the other hand, there are features that can be more meaningful when extracted on a short-term level. Spectral features as used in automatic speech recognition or speaker recognition, such as Mel Frequency Cepstrum Coefficients or Perceptual Linear Prediction coefficients, are typically measured with frame rates around 10–20 ms and window lengths around 20–40 ms. Formants can also be used as features. The majority of studies, see Table 2.4, employ a combination of pitch, energy, duration and spectral features for the classification of emotion: usually a combination of features extracted on different levels yields the best performances (e.g., fusion between spectral and prosodic features).

2.3.3 Learning

The machine learning algorithm learns how to map input features to specific emotion classes. There are several learning algorithms that are very popular among emotional speech researchers, mainly because these have proven to give good performance. In Table 2.4, it can be observed that Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM) are among the most frequently used ones.

Support Vector Machine (Vapnik [197]), a **kernel method**, is one of the most popular method used in emotional speech recognition, e.g., Schuller et al. [171], Schuller and Rigoll [170], Devillers and Vidrascu [52]. A Support Vector Machine aims at finding the best separating hyperplane between groups of datapoints that maximizes the margins between these groups (see Section 2.4.3, and for a more detailed and mathematical description, see Vapnik [197]), and is also known as a *discriminative* method.

Hidden Markov Models and Gaussian Mixture Models are *generative* methods that model **probability density functions (pdfs)**. HMMs have successfully been employed to model sequential, temporal patterns: current automatic speech recognition technology is based on the use of HMMs. An HMM models these sequential patterns with stochastic processes, complying to the Markov property that given a state, future states only depend on the current state and not on the past states. In short, an HMM can be specified by five elements: 1) the number of states in the model, 2) the observation space per state which can consist of a finite number of elements (discrete HMM) or which can be infinite, multidimensional and continuous (continu-

ous HMM), 3) the state transition probability distribution, 4) the observation symbol probability distribution in a certain state, and 5) the initial state distribution. The observation probability distribution in a certain state can be modeled by a Gaussian Mixture Model. Hence, a GMM is sometimes also referred to as a 1-state HMM. For a more detailed description of HMMs, the reader is referred to Rabiner and Juang [144], Rabiner [145]. GMMs are briefly described in section 2.4.3.

The training of HMMs and GMMs usually requires a lot of speech data in order to accurately model the probability distributions: speaker or speech recognition systems that employ HMMs or GMMs are usually trained with hundreds hours of audio. Typically, HMMs and GMMs perform best when trained with short-term spectral features while SVMs perform best when trained with static acoustic vectors computed over the whole utterance (e.g., Batliner et al. [18], Schuller et al. [173], Truong and van Leeuwen [187]).

2.3.4 Evaluation

Through evaluation, the final development process in Table 2.2, we can assess the performances of the affect recognition systems developed: how well does the affect recognition system perform (in comparison with other systems)? During evaluation, it is standard practice to divide the data in three mutually disjoint sets: one for training, one for development and one for testing. Most importantly, the testing set should be a held-out set containing samples that are not present in the training set. Often, to compensate for lack of data, a K-fold cross validation procedure can be used. A special case of K-fold cross validation in which each fold samples from a certain speaker is held out for testing is also referred to as 'leave-one-out' or 'leave-one-speaker-out' (LOSO). Leave-one-speaker-out evaluation is useful to asses the speaker-independency of emotion classifiers. If shared datasets, common evaluation protocols and common performance metrics are used, then comparisons between different studies can be properly made. However, in speech-based emotion recognition research, there is a lack of well-defined standards: this is the main reason why the studies in Table 2.4 cannot be properly compared to each other.

As a performance metric, the most commonly used metric according to Table 2.4 is classification accuracy which is defined by the number of correct classifications divided by the total number of test samples. However, classification accuracy is not a very appropriate performance measure when the balance between the classes in the test set is imbalanced. Several studies have proposed to use a performance measure defined as the uniformly weighted harmonic of the classification accuracy and the per-class recognition rate, e.g., Batliner et al. [16, 18], Schuller et al. [172], while others borrow performance measures from other research areas. For example, emotion recognition can be approached as an information retrieval problem: given a set of speech samples that contain different emotions, how can we retrieve the speech samples with the specific target emotion that was queried. For these types of tasks, performance can be expressed as the weighted harmonic mean between precision and recall (e.g., Neiberg et al. [124]). In the context of this type of emotion search system, precision can be defined as the number of correctly retrieved speech samples divided by the total number of speech samples retrieved. Recall can be defined as the num-

ber of correctly retrieved speech samples divided by the number of speech samples that should have been retrieved. Furthermore, borrowed from speaker recognition and signal detection theory, Equal Error Rate (the point where the false alarm rate is equal to the miss rate) can also be applied in emotion recognition (e.g., Clavel et al. [39]). These evaluation metrics will be discussed in more detail in section 2.4.4.

Obviously, the use of all these different performance metrics, the lack of standards and shared datasets do not contribute to the transparency of the large number of emotion recognition studies.

Study	Type emotions (number of emotions)	Number of speakers	Nature data	SI/
Banse 1996 [12]	Hot+Cold Anger, Joy, Disgust, Sadness etc. (14)	12	Acted	?
Petrushin 1999 [133]	Agitation, Calm (2)	18	acted telephone messages	?
Zhou 2001 [219]	Neutral, Anger, Loud, Lombard (4)	?	simulated, real (SUSAS)	SI/
Ang 2002 [6]	Annoyance, Frustration (2)	?	wizard- of-oz	?
Tato 2002 [182]	Anger, Happy, Sad, Bored, Neutral (5)	14	acted	SI
Nwe 2003 [125]	Anger, Disgust, Fear, Joy, Sadness, Surprise (6)	12	acted	SD
Batliner 2003 [16]	Joy, Surprised, Neutral, Helpless, Anger (7)	86	wizard of oz (Smartkom)	?
Fernandez 2003 [62]	Stress (4)	4	spontaneous (in-car)	SD
Kwon 2003 [99]	Anger, Lombard, Loud, Neutral (4)	?	partly simulated, partly real (SUSAS)	?
Kwon 2003 [99]	Anger, Bored, Neutral, Sad, Happy (5)	51	spontaneous (AIBO)	?
Yacoub 2003 [214]	Anger, Neutral (2)	8	acted	SI
Ververidis 2004 [199]	Anger, Happiness, Sadness, Surprise, Neutral (5)	4	acted	?
Vidrascu 2005 [201]	Positive, Negative (2)	404	real-life medical	SI
Schüller 2005 [171]	Joy, Anger, Disgust, Fear, Sadness, Surprise, Neutral (7)	35	movies	SI

Devillers 2006 [52]	Relief, Anger, Fear, Sadness (4)	690	real-life financial, medical	SI
Clavel 2006 [39]/Clavel 2007 [40]	Fear, Neutral (2)	400	movies	SI
Batliner 2006 [18]	Anger, Motherese, Emphatic, Neutral (4)	51	spontaneous (AIBO)	SI
Neiberg 2006 [124]	Positive, Neutral, Negative (3)	92	spontaneous meetings	?
Schuller 2006 [170]	Anger, Fear, Joy, Disgust, Sadness, Boredom, Neutral (7)	10	acted (Berlin)	?
Graciarena 2006 [67]	Deceptive, non-deceptive (2)	32	spontaneous (interviews)	SD
Datcu 2006 [48]	Anger, Fear, Joy, Disgust, Neutral, Sadness, Boredom (7)	10	?	act
Schuller 2007 [172]	Anger, Motherese, Emphatic, Neutral (4)	51	spontaneous (AIBO)	?
Hu 2007 [81]	Anger, Fear, Happiness, Sadness, Neutral (5)	8	acted	?
Grimm 2007 [69]	Activation, Valence, Dominance (3)	47	spontaneous tv (VAM)	?
Yingthawornsuk 2007 [217]	Depressed, Suicidal (2)	20	partly spontaneous	?
Vlasenko 2007 [202]	Anger, Fear, Disgust, Joy, Neutral, Boredom, Sadness (7)	10	acted (Berlin)	SI
Batliner 2008 [20]	Intimacy, Neutral (2)	24	spontaneous (child-directed speech)	?

Table 2.4: *Various emotion classification studies in speech briefly summarized. LDA=Linear Disc*
Model, QDA=Quadratic Discriminant Analysis, SVM=Support Vector Machine, RF=Random Forest

2.4 Materials and methods used in current study

As an introduction to the experiments presented in this thesis, a short overview of all databases, speech features, learning algorithms and evaluation metrics used in our experiments, is given in this chapter.

2.4.1 Databases

For the emotion recognition experiments performed in this thesis, various speech databases were employed. Five different databases were used, of which four contain spontaneous speech, i.e., ICSI, AMI, CGN, and TNO-GAMING. The TNO-GAMING corpus was newly collected and is presented in this thesis (for a detailed description of this corpus see Chapter 6). These databases are shortly described here, and summarized in Table 2.5.

Database	**Number speakers**	**Size**	**Nature data**	**Types of emotion**	**Description, annotation of emotion data**
ICSI (Janin et al. [85])	53 (English, non-native English)	72 h	non-scripted, natural	wide range of emotionally colored behavior	emotional behavior in meetings, (discrete) annotations of dialog acts and laughter, (dis-)agreement
AMI (Carletta [35])	171 (English, non-native English)	100 h	scripted, natural	wide range of emotionally colored behavior	emotional behavior in meetings, (discrete) annotations of dialog acts and subjectivity, laughter, (dis-)agreement
CGN (Oostdijk [127])	4251 (Dutch)	800 h	natural, scripted spontaneous, read speech	no particular emotion	face-to-face, spontaneous telephone, interviews, lectures, broadcast, read speech, no particular emotion annotations
TNO-GAMING (Merkx et al. [118])	28 (Dutch)	appr. 78 m – 186 m	spontaneous, elicited	Frustration, Amusement, Excitement, Surprise etc.	audiovisual recordings of subjects playing videogames, discrete and continuous dimensional emotion annotation from gamers themselves and observers
BERLIN (Burkhardt et al. [25])	10 (German)	appr. 25 m	acted	Anger, Joy, Disgust, Fear, Neutral, Sadness, Boredom	discrete

Table 2.5: *Short overview of (emotional) speech databases used in current study.*

ICSI The ICSI Meeting Recorder Corpus (Janin et al. [85]) contains 72 hours of audio recorded during 75 meetings held at ICSI (International Computer Science Institute). These are 'natural' meetings, in the sense that these meetings would have taken place anyway. The recordings were made with head-worn microphones (near-field) and desktop microphones (far-field). In total, there are 53 unique speakers of which 23 are non-native speakers of English and 28 are native speakers of English. For each meeting, speech transcripts are available at different information levels. In addition to words, other information is also transcribed such as word fragments, restarts, filled pauses, contextual comments (e.g., "while whispering") and non-lexical events such as laughter and coughs. In addition, dialog act information is also available through the MRDA corpus (the ICSI Meeting Recorder Dialog Act corpus, Shriberg et al. [176]), which includes hot spot labeling (Wrede and Shriberg [211]). Emotion annotation is available in the form of hot spot and laughter annotation. The speech transcriptions in the ICSI corpus allow for research on diverse topics such as speech activity detection, overlap analysis, hot spot analysis, agreement vs. disagreement detection etc.

AMI The AMI Meeting Corpus (Carletta [35], McCowan et al. [114]) contains 100 hours of multimodal meeting data. Of the 100 hours, 35 hours contain non-scripted meetings and 65 hours contain scripted meetings. In the scripted meetings, the participants (4 per meeting) play the roles of employees of an electronics company who are part of a design team whose task is to develop a new remote control. Since not all participants are native speakers of English, the database consists of a mix of native and non-native English. The AMI (Augmented Multi-party Interaction) project is concerned with research and development of technology to support human interaction in meetings, and to improve the effectiveness of the ways meetings are run and documented. Speech and face recordings are available, as well as captured images of the slides, shared documents, and electronic whiteboard output used during the meetings. In addition, annotations of different types of meta-data are available, e.g., speech transcriptions, dialog act annotation, emotion annotation, topic segmentation, gesture annotation, subjectivity, etc. These multimodal, synchronized recordings and annotations allow for multidisciplinary research in audio and visual processing and recognition, for example, speech recognition, emotion recognition, information retrieval, gesture recognition etc.

CGN **(Corpus Gesproken Nederlands, in English: Dutch Spoken Corpus)** The Spoken Dutch Corpus (Oostdijk [127]) contains approximately 800 hours of Dutch speech, recorded in The Netherlands and Flanders. Recordings consist of various types of human communications such as spontaneous face-to-face conversations, telephone dialogues, interviews, lectures, broadcasts, read speech and more. The corpus includes speech transcriptions and annotations on different levels: orthographic transcription, POS (Part-Of-Speech) tagging and lemmatization, and lexicon coupling for multi-word units. A part of the corpus is also enriched with syntactic annotations, phonetic transcriptions, word segmenta-

tions and prosodic annotations.

TNO-GAMING The TNO-GAMING database (Merkx et al. [118]) is a newly collected database that is presented in more detail in Chapter 6. The database contains approximately 186 minutes spontaneous audiovisual data of subjects who where playing a videogame against each other. The data was collected at TNO in Soesterberg, The Netherlands. In 7 sessions, 28 participants played a videogame in teams of 2 by 2 against each other. High quality speech and face recordings are available, as well as a speech transcriptions on word-level. Emotions were elicited by inserting surprising events in the game and by hampering keyboard and mouse controls. The gamers annotated their own emotions in emotion categories, and on Arousal and Valence emotion dimensions. A part of the database is also annotated by external observers. One of the interesting aspects of this database is that it contains emotion annotations of the people playing the videogame themselves, *and* of observers.

BERLIN The BERLIN Emotional Speech database[1] (Burkhardt et al. [25]) contains emotional speech from 10 actors (5 male and 5 female speakers). The emotions expressed are Anger, Disgust, Joy, Fear, Neutral, Sadness and Boredom (note that Surprise is not included which is considered one of the 6 basic universal emotions by Ekman [56], and that Boredom is not considered a basic emotion by Ekman [56]). As validation of the emotions expressed, 20 subjects were asked to recognize the correct emotion and to rate the emotion expressed on level of naturalness. From the original 800 sentences, 494 sentences remain after selecting speech samples that are rated as natural by at least 60% of the listeners and that are correctly recognized by at least 80% of the listeners.

2.4.2 Speech features

Several types of segmental (frame-based) and suprasegmental speech features were used in several experiments described in this thesis. First, we have employed typical ASR and speaker recognition frame-based speech features, like PLP and MFCC, for speech-based emotion recognition. Furthermore, suprasegmental features, like pitch and energy, will also briefly be described.

Frame-based (segmental) features

Perceptual Linear Prediction Coefficients (PLP) Perceptual Linear Prediction Coefficients (PLP, Hermansky [77]) have been successfully applied in speaker and speech recognition as representatives of the speech signal (e.g., Hermansky [77], Matejka et al. [113], Kajarekar et al. [89]). PLP coding is similar to Linear Predictive Coding (LPC) analysis in that it is based on the short-term spectrum of speech with the advantage that PLP coding is more consistent with human hearing; it modifies the short-term spectrum of speech by several psychophysically based transformations. The basic idea of LPC is that a speech sample at the

[1]The whole database, including additional (acoustic) analyses can be downloaded from *http://www.expressive-speech.net*

current time can be approximated as a linear combination of past speech samples. The RASTA filter (Relative Spectral), proposed by Hermansky and Morgan [78] makes PLP analysis more robust against slowly varying linear spectral distortions (that can be caused by channel distortions). Rasta-perceptual linear prediction coefficients are also referred to as RPLP. Figures 2.5 and 2.6 summarize the processes involved in the computation of these features.

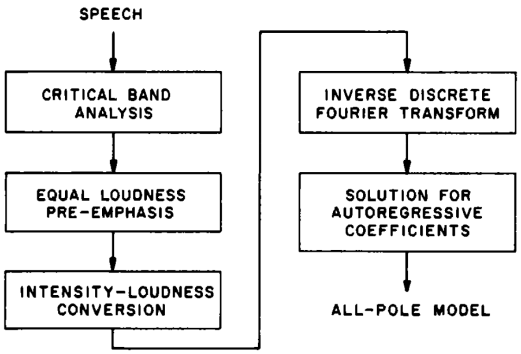


Figure 2.5: *Processes involved in the computation of PLP features (figure adopted from Hermansky [77]).*

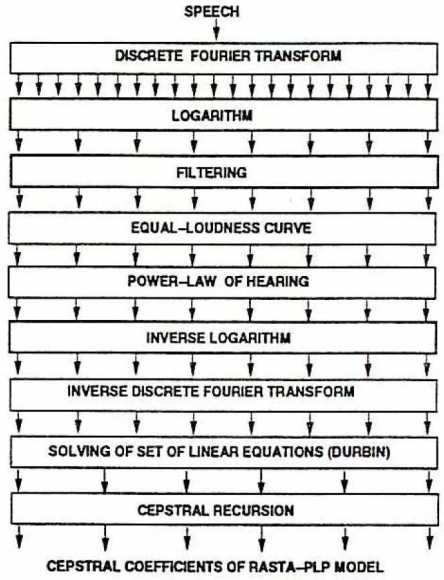


Figure 2.6: *Processes involved in the computation of RPLP features (figure adopted from Hermansky and Morgan [78]).*

Mel-Frequency Cepstrum Coefficients (MFCC) Mel-Frequency Cepstrum Coefficients are the most popular speech features used in automatic speech recognition (e.g., Davis and Mermelstein [49], Morgan and Bourlard [119]). Similar to PLP and RPLP analysis, MFCCs are representations of the short-term power spectrum, except that the spectrum is averaged over neighboring frequency bands

and the frequencies are scaled according to a scale that imitates psychoacoustic properties of the human ear. In this case, the Mel-scale is used which is a scale of pitches judged by listeners to be equal in distance from one and another; the Mel-scale is linear in the lower frequency area and logarithmic in the higher frequency area. The coefficients are obtained after a Discrete Cosine Transform of the spectrum; hence, this process is also known as 'taking the spectrum of a spectrum' and results in a representation that is also known as a *cepstrum*. Fig. 2.7 shows the processes involved in the computation of MFCC features.

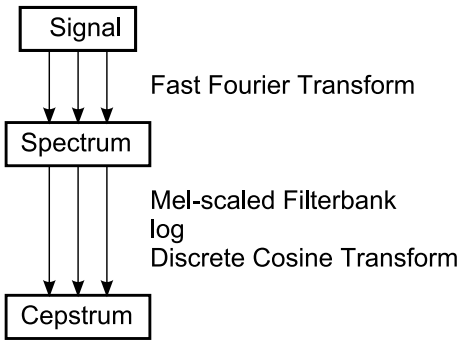


Figure 2.7: *Processes involved in the computation of MFCC features.*

Suprasegmental features

In a substantial number of studies on the characteristics of emotional speech, measurements of pitch, energy, speech rate, the distribution of energy in the long-term averaged spectrum (LTAS), and statistics thereof (e.g., standard deviation, mean, minimum, maximum), have proven to be relatively good descriptors of emotional speech (e.g., Schuller et al. [172], Vlasenko et al. [202], McGilloway et al. [115], Banse and Scherer [12], Pittam et al. [135]). Inspired by these studies, we have mainly focused on pitch-related, energy-related, and the long-term averaged spectrum-related suprasegmental properties of the speech signal. As Table 2.2 shows, in feature extraction, one has to decide how large the analysis window is for feature extraction. The use of linguistic meaningful units as analysis windows, such as words, was not always possible because word segmentation was not always available. Therefore, in some cases, we used the whole utterances as an analysis unit, and in some cases, we used voiced units that were determined by a simple voiced-unvoiced detection algorithm. Furthermore, when the acoustic features were extracted per word in an utterance, we aggregated these word-level features to utterance-level because some machine learning models perform better with static feature vectors. Aggregation was simply done by taking the mean, minimum and maximum over all word-level features, per feature per utterance. Table 2.6 gives an summary of acoustic features that were frequently employed in this study. More details about the extraction of these features are given in the Chapters separately. Here, we briefly give short descriptions of the speech features used. Most of these features could be extracted with Praat (Boersma and Weenink [23]).

Features extracted over each $word_1 \dots word_{\mathrm{N}}$ **or whole utterance**	**Aggregation functions for each feature over all the** N **features of an utterance** $\{featureA_1 \dots featureA_{\mathrm{N}}\}$
standard deviation pitch, mean pitch, range pitch, mean absolute slope pitch, standard deviation intensity, mean intensity, range intensity, mean root-mean- square, mean absolute slope intensity, slope LTAS, Hammarberg index, center of gravity of spectrum, skewness of spectrum, speech rate	mean, minimum, maximum

Table 2.6: *Some acoustic features that we have frequently used in this thesis, the aggregation functions only apply when features are extracted on word-level.*

Pitch It is generally acknowledged and shown (e.g., Banse and Scherer [12], McGilloway et al. [115], Murray and Arnott [123]) that pitch plays an important role in the expression of vocal affect. As shown in Table 2.1, mean pitch and the range of pitch tend to increase with an increase in Arousal. Sadness and Boredom are usually associated with monotonous melody contours which are reflected in e.g., low pitch range values and small standard deviations. Anger and Joy are often associated with increased pitch values. Pitch is actually the perception of a physical property of a signal namely the *fundamental frequency* (F_0). To measure F_0, Praat (Boersma and Weenink [23]) uses an algorithm that performs an acoustic periodicity detection on the basis of an accurate autocorrelation method as described in Boersma [22].

Intensity/energy Similar to pitch, intensity is deemed important for the expression of affect in the voice. Highly Aroused speech can be associated with high intensity values, and vice versa (see Table 2.1). In Praat (Boersma and Weenink [23]), intensity values are based on an intensity contour that is calculated at linearly spaced time points $t_i = t_1 + (i-1)dt$.

Energy distribution in (long-term averaged) spectrum It is known that with increased vocal effort (that is related to perceived loudness), the amount of energy in the higher frequency regions of the spectrum increases (relative to the lower frequency regions, see e.g., Sluijter and Heuven [177]). Since vocal effort can also be related to affect, the difference between energy in higher and lower frequency bands of the (long-term averaged) spectrum is also often used as a cue to different speaking styles and emotions (e.g., Banse and Scherer [12], Schröder et al. [169]). The Hammarberg index (Hammarberg et al. [73]) is an example of a measure that measures this difference of energy: it is defined as the difference between the energy maximum in the 0–2000 Hz frequency band and in the 2000–5000 Hz band. We have also used the slope of the long-term averaged spectrum as a measure: the expectation is that this negative slope will be less steep with increasing vocal effort. The center of gravity is a measure for

how high the frequencies in a spectrum are on average, and the skewness is a measure of how much the shape of the spectrum below the center of gravity is different from the shape above the mean frequency (both can be measured with Praat).

Speech rate A higher speech rate is associated with high Arousal, while low speech rate is associated with low Arousal, see Table 2.1. We approximate speech rate by dividing the number of words (or voiced units) spoken by the total amount of time used to speak these words (optionally, the pauses can be subtracted from the total time).

2.4.3 Machine learning methods

Several machine learning methods have been employed for affect recognition in this work. The choice for a specific machine learning model was mainly motivated by proven successes achieved in other, past studies or other similar recognition technologies. We discuss Gaussian Mixture Models, Support Vector Machines, and the AdaBoost algorithm that we have used in our affect recognition experiments.

Gaussian Mixture Modeling (GMM) GMMs (e.g., Alpaydin [5], Reynolds and Rose [148]) form the basis of Hidden Markov Models used in current ASR technology. GMMs are also referred to as 1-state HMMs. A GMM models the distribution of observed data through Gaussian Probability Density Functions (PDFs). Each pdf can be defined by its parameters: its mean μ and its covariance matrix Σ. The general assumption of an GMM is that there are K components; each component generates data from a Gaussian with mean μ_k and covariance matrix Σ_k. A GMM is a weighted average of the K components (hence the 'mixture'), where the sum of w_k amounts to 1. In learning, these unknown parameters are optimized through the Expectation-Maximization (EM, see Dempster et al. [50]) algorithm. In testing, it is calculated what the likelihood is that the GMM has generated the test datapoint. GMM is also known as a 'generative' method.

Support Vector Machines (SVM) The basic principle of a Support Vector Machine (Vapnik [197]) is that the algorithm aims to find the best separating hyperplane, $\mathbf{w} \cdot \mathbf{x} - \mathrm{b} = 0$, between groups of datapoints $\mathbf{x}$ that maximizes the margins between these two groups labeled by -1 or 1, i.e., to maximize the distance between the data points of each class that are nearest to each other. The goal of an SVM is to find $\mathbf{w}$ and b so that the margin between the parallel separating hyperplanes is maximized such that $\mathbf{w} \cdot \mathbf{x}_\mathrm{i} - \mathrm{b} \geq 1$ for $\mathbf{x}_\mathrm{i}$ of the first class, or $\mathbf{w} \cdot \mathbf{x}_\mathrm{i} - \mathrm{b} \leq -1$ for $\mathbf{x}_\mathrm{i}$ of the second class. This can be formulated as an optimization problem that can be solved by *quadratic programming*. The vectors found, that uniquely determine the largest margin between the two classes, are called the *support vectors*. Hence, an SVM is also known as a *maximum-margin classifier* and is also called a 'discriminative' method. In some cases, the data is linearly separable. However, in general, non-linear classifiers are needed to deal with low dimensional, real-world, noisy data. SVMs can use a (possibly non-linear) kernel function that implicitly maps the data to a higher dimensional

feature space. The idea behind this is that separation may be easier in higher dimensions.

Support Vector Regression (SVR, Smola and Schölkopf [178]) is based on the same principles as SVM: SVR is a kernel-based method and allows the use of the kernel trick to transform the original feature space to a higher-dimensional feature space through a (non-linear) kernel function. In the case of SVR, a margin ϵ is introduced and SVR tries to construct a discriminative hyperplane that has at most ϵ deviation from the original training samples. Similar to linear regression, SVR finds relations between a set of independent variables and dependent variables. In contrast with SVM, these dependent variables are scalar values rather than discrete categories.

Recently, researchers have combined the concepts of the 'generative' GMMs and 'discriminative' SVMs into a method known as 'GMM supervector based SVM' (Campbell et al. [33, 34]). The basic idea behind this method is to use the means of a GMM for SVM classification. In short, the mean of each Gaussian component of each speech sample is stacked in a 'supervector' and forms the input for SVM classification. We used this method for the development of a speech-based affect recognizer described in Chapter 4; hence, more details about this method are given in that Chapter.

AdaBoost AdaBoost (Freund and Shapire [64]) is one of the many boosting algorithms. Boosting is in fact a meta-algorithm that can be used in combination with many other learning algorithms to improve their performance. It is an iterative algorithm that is based on the principle of combining many simple, weak classifiers (learners) into a single, strong classifier. These weak learners can be any type of classifier; for example, one implementation of boosting called Boostexter (Shapire and Singer [175]) employs one-level decision trees as weak learners. In a series of rounds, weak learners are repeatedly called to produce weak hypotheses. AdaBoost is an *adaptive* boosting algorithm in that it adapts to the error rates of the individual weak hypotheses. As the boosting process progresses, importance weights increase for training samples that are hard to predict and decrease for training samples that are easy to classify. In this way, future weak learners concentrate on those examples that are hardest to classify.

2.4.4 Evaluation metrics

Various performance metrics have been used in emotion classification studies. The most widely metric used in emotion classification is probably classification accuracy, although this metric is sensitive to inhomogeneously distributed evaluation classes. Classification accuracy is suited for a *classification task*: how well can a number of different classes (where this number can be larger than 2) be discriminated from each other? In a classification task, the question can be of the kind "Does the speech of this person belong to the classes Anger, Sadness, or Frustration?". Confusion matrices are very handy for investigating classification errors, see Fig. 2.9. In our evaluation, we mainly adopt performance metrics from the detection framework. *Detection*, however,

is slightly different from classification. In detection, one wants to answer the question: "Does the speech of this person belong to the class Anger or not?". Furthermore, detection is a binary-decision task while classification can be multiclass. The types of error a binary classifier can make are given are shown in Fig. 2.8. Unfortunately, few researchers have evaluated their emotion recognizers in the detection evaluation framework. Therefore, in order to allow other researchers to compare their performances to our performances, we have used a range of different performance metrics which will be described in this section.

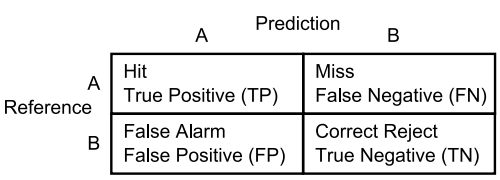


Figure 2.8: *Error types for a 2 class problem.*

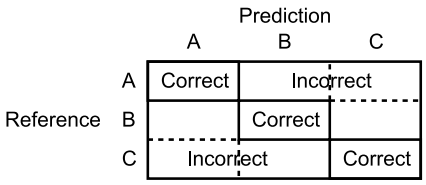


Figure 2.9: *Confusion matrix for a multi-class (>2) problem.*

Classification accuracy The (overall) classification accuracy is a widely used measure that expresses how well a classifier works, and that can be computed relatively easy. It can be defined as the number of correct classification divided by the total number of test samples. In a $K \times K$ class confusion matrix (Fig. 2.9), the diagonal represents the number of correctly classified samples, while the off-diagonals contain numbers of incorrectly classified samples. Hence, a confusion matrix gives insight in the type of confusions that are made between classes. Classification is therefore more diagnostic of character.

$\mathbf{F}_1$ The F_1 score is often used in information retrieval to measure the performance of a search system. It is defined as the harmonic mean between precision and recall.

$$\mathrm{F}_1 = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{2.1}$$

In the context of information retrieval, precision and recall can be explained as follows. Precision is the number of relevant documents retrieved divided by the total number of *retrieved* documents. Recall has a different denominator: recall is the number of relevant documents retrieved divided by the total number of relevant documents that *should* have been retrieved. In the context of classification, precision and recall can be defined in terms of different error types. In 2-class classification problems, there are four types of classifications possible, see Fig. 2.8. In this context, precision and recall can be defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{P}_{\mathrm{miss}} \tag{2.2}$$

Affect recognition can thus also be approached as a retrieval problem (see also section 2.3): given a query for a specific emotion, the search system's task is to retrieve all speech samples that contain this emotion.

$\mathbf{F}_{\text{emo}}$ In Batliner et al. [19], Schuller et al. [172, 173] a performance measure was used that is defined as the harmonic mean of the overall classification accuracy and the averaged per-class recognition rate. This measure will be referred to as F_{emo} throughout this thesis and is only reported to allow for comparison.

Equal Error Rate (EER) In several research areas, e.g., speaker, language recognition, it is very common to evaluate classifiers in a so-called *detection* evaluation framework. Equal Error Rate is one of the popular metrics used in this framework. In a 2-class classification problem, there are 2 classes that need to be discriminated from each other: one *target* class and one *non-target* class. The assumption that is made is that when a classifier outputs scores, e.g., decision scores, likelihood ratios, the higher scores are associated with *target* samples whereas lower scores are associated with *non-target* samples. In order to make a final decision on class membership a threshold must be placed in this distribution of scores. When a threshold is placed, errors can be count, see Fig. 2.10. The classifier can make two types of errors as shown in Fig. 2.8:

1. False alarms, i.e., a non-target sample is classified as belonging to the target class. The false alarm rate is computed as $\frac{FP}{FP+TN}$.
2. Misses, i.e., a target sample is classified as belonging to the non-target class. The miss rate is computed as $\frac{FN}{FN+TP} = 1 - \text{Recall}$.

The tradeoff between the false alarm rate and miss rate can be made visible in a Detection Error Tradeoff (DET) curve which is a helpful visual evaluation tool. By stepping through all the scores which serve as 'temporal' thresholds, error rates can be computed and plotted which result in a DET curve. The DET curve can be summarized into one single performance figure: EER represents the operating point where the false alarm rate is equal to the miss rate. The disadvantage is that EER is based on a decision threshold that is set a posteriori: what would the false alarm and miss rate be if the threshold was set at this value?

Cost of Detection ($\mathbf{C}_{\text{det}}$) When a decision threshold is set a priori (obtained through a difficult process known as threshold calibration), the performance can be measured through the Detection Cost Function (DCF) that is defined as:

$$C_{\text{det}} = C_{\text{miss}} P_{\text{miss}} P_{\text{tar}} + C_{\text{fa}} P_{\text{fa}} (1 - P_{\text{tar}}) \tag{2.3}$$

C_{miss} and C_{fa} are the costs that one can attribute to a certain type of error. P_{tar} is the a priori probability that a target sample occurs. And P_{miss} and P_{fa} are the miss and false alarm rates respectively. In a specific case where the cost parameters C_{miss} and C_{fa} are equally 1, and where there is equal prior probability $P_{\text{tar}} = 0.5$, C_{det} is also called Half Total Error Rate (HTER) (which is in fact the mean of P_{miss} and P_{fa}, and hence the name HTER is misleading).

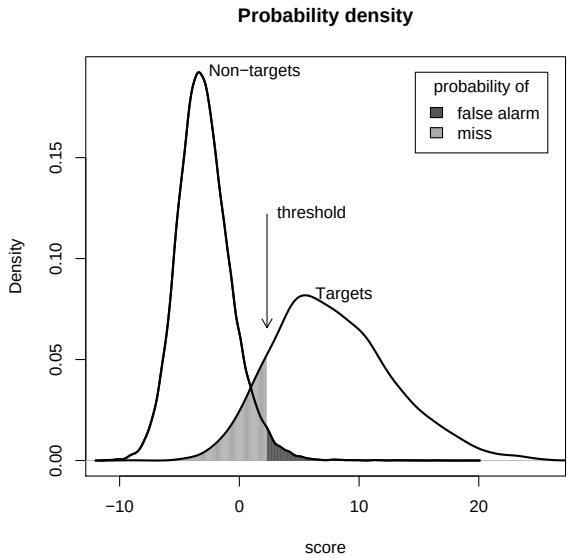


Figure 2.10: *Score distributions for target and non-target samples. The grey areas left and right of the threshold represent* P_{miss} *and* P_{fa} *respectively (reproduced from van Leeuwen and Brümmer [196]).*

2.5 Conclusions

In this Chapter, we have given a concise introduction into the research area of speech-based affect recognition. First, we gave an overview of speech-based affect recognition studies and described development processes that are involved in the development of an affect recognizer. We identified development processes namely 'data acquisition and annotation', 'feature extraction and model learning', and 'evaluation'. Each of these processes were described and terms and concepts associated with these processes were explained. Finally, we presented our own materials, methods, features, and performance metrics used in our current study during these development processes. It is clear that automatic speech-based affect recognition is a relatively young research area. One of the developing research directions is the use of more realistic, natural emotion data that is one of the topics under investigation in this thesis.

Chapter 3

Capturing and measuring real affect in the field

In the emotion recognition research community, there is a growing need for labeled natural emotion data. In addition to capturing the natural emotion data, the emotions in the data also need to be measured and described: these are difficult tasks which will be discussed in this Chapter. This Chapter is about ways to capture and measure people's emotions in the *field*. When actors are used in the lab, capturing and measuring emotion is rather straightforward since actors can be hired and directed to express a specific emotion: actors can be directed to act angry, act happy etc. and the emotional label can thus be directly acquired. But when one wants to capture and measure natural emotions in the field, one needs to *find* and *specify* the emotion. Since the annotation of discrete emotion categories is rather straightforward, the majority of the tools that have been developed to measure emotion are more flexible and offer ways to describe and annotate emotion according to a dimensional model of emotion. The most popular measures and tools are described in Section 3.1, and can be applied to natural emotion data. In Section 3.2, we describe three attempts that we have undertaken to capture and measure natural emotional expressions in the field. Natural emotions were measured on board of a naval ship, during time-pressured crisis meetings, and during a virtual reality game. We discuss our experiences and 'lessons learned' acquired from these efforts[1].

3.1 Measures of affect

Here, we describe some popular instruments/tools to measure affect. These measurements of emotion are in fact all variants from each other and differ mostly in the way these measures are presented. Some of these measures are more verbally oriented while others are more graphically oriented.

Osgood's Semantic Differential Osgood's Semantic Differentials (Osgood [128]) are rating scales that were originally developed to measure the "meaning" of particular concepts. The idea of Semantic Differentials is that the "meaning" of each

[1]The work described here was (partly) previously discussed in Truong et al. [192]

concept can be characterized by a number of 7-point scales, each of which has bipolar adjectives, as shown in Fig. 3.1. The set of adjectives used in the scales can be large since almost each adjective can be used as a descriptor of a 'concept'. In agreement with other researchers, Osgood found that there were three recurring factors that people use to evaluate words: evaluation, potency, and activity. Evaluation can be rated by the adjective pair 'good vs. bad', potency can be rated by 'strong vs. weak', and activity can be rated by 'active vs. passive'. These three dimensions have frequently been applied to emotion research to describe and measure emotion.

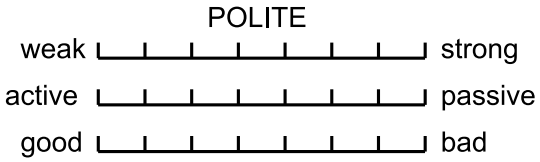


Figure 3.1: *Example of Semantic Differential scales for measuring the "meaning" of the concept 'polite'.*

Feeltrace Feeltrace is an emotion annotation instrument based on the two emotion dimensions of Arousal (also known as Activity) and Valence (also known as Evaluation) developed by Cowie et al. [45], see Fig. 3.2. Feeltrace was designed to let observers track the emotional content as they perceive it in time. It incorporates the notion that the emotion space is naturally circular (Plutchik [137]). The circumference is defined by states that are at the limit of emotional intensity. These states are equally distant from an emotionally neutral point, i.e., the center which represents neutrality. In the circle, landmarks that mark key emotional states, are drawn to help the observers' orientation in this Arousal-Valence space. When the mouse button is held down, the trace of the mouse movements in time in the circle is drawn and recorded. The trace consists of little circles that are associated with the mouse cursor's position on the screen and the dimension of time: the circles gradually shrink in the course of time, leaving behind a trail of diminishing and vanishing circles. The circles are colored according to a color scheme derived from Plutchik [137] which subjects find reasonably intuitive. The circle is colored pure red when the cursor is at the most negative position, and neutral in Arousal. It is pure green when the cursor is at the most positive position, and neutral in Arousal. It is pure yellow or pure blue when the cursor is in most active or passive position respectively, and neutral in Valence. The circle is white in neutral position.

Affect Grid The Affect Grid (Russell et al. [154]) was designed as a quick means to measure affect along the dimensions of pleasure vs. displeasure (Valence) and arousal vs. sleepiness (Arousal). The aim was to have an instrument that would be short and easy to fill out and that could, therefore, be used rapidly and repeatedly. In contrast with Feeltrace, the Affect Grid is square and does not track affect in time, see Fig. 3.3. The dimensions pleasure-displeasure and arousal-sleepiness are considered independent from each other: the two dimensions are

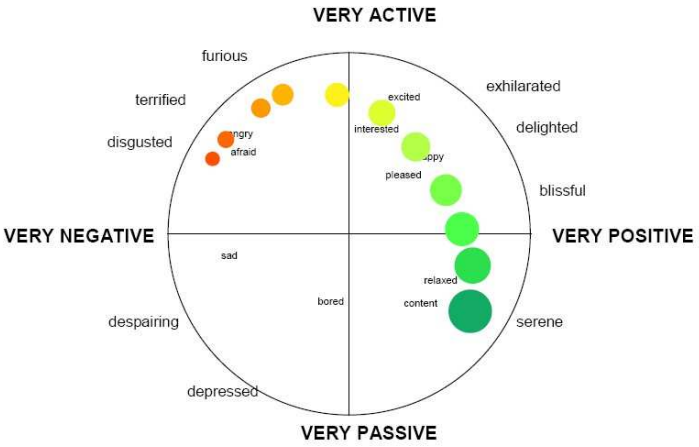


Figure 3.2: *Example of a Feeltrace tracking session: the circles represent a person's emotional state who is gradually going from a disgusted state to a state of serenity (figure adopted from Cowie et al. [45]).*

conceptually separate.

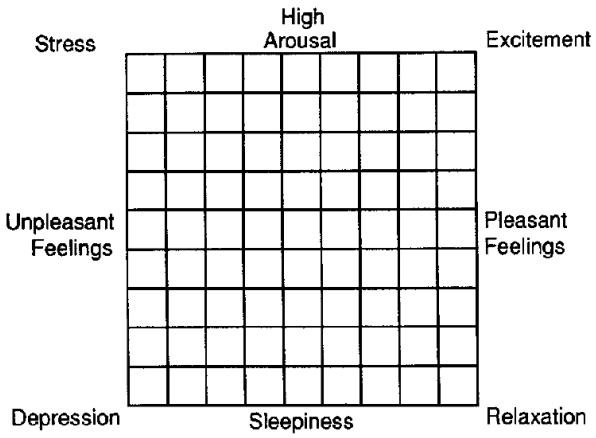


Figure 3.3: *Affect Grid: in this* 9×9 *square grid, subjects can place an 'X' to mark a specific affect in this 'map of feelings'.*

SAM The Self Assessment Manikin (SAM, Lang [101]) is a tool to measure affect on the scales of happy vs. unhappy (Valence), excited vs. calm (Arousal), and control vs. controlled (Dominance). Instead of using adjectives to describe the scales, SAM uses graphical, symbolic representations of affect, see Fig. 3.4. For the Valence scale, a broadly smiling face is gradually changing to a tragic mask. The Arousal scale is represented by figures that change from having their eyes wide open and showing rapidly body movement to figures that have their eyes closed. The Dominance scale is represented by figures that grow from tiny to gigantic. The SAM method has been used relatively frequently

to assess perceived emotion. The main advantage of the SAM is the symbolic representation of the three emotion dimensions. This makes the SAM method a language-independent instrument that can also be used with children.

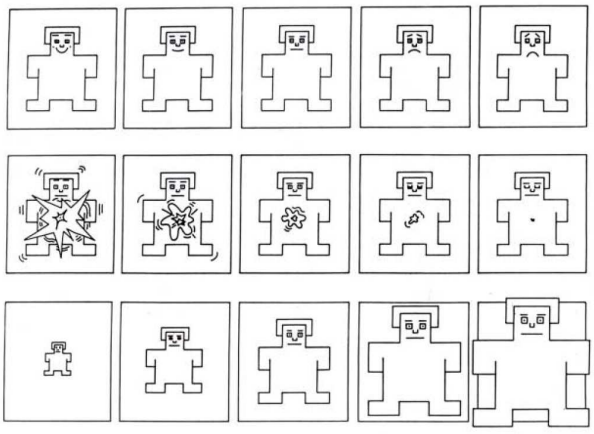
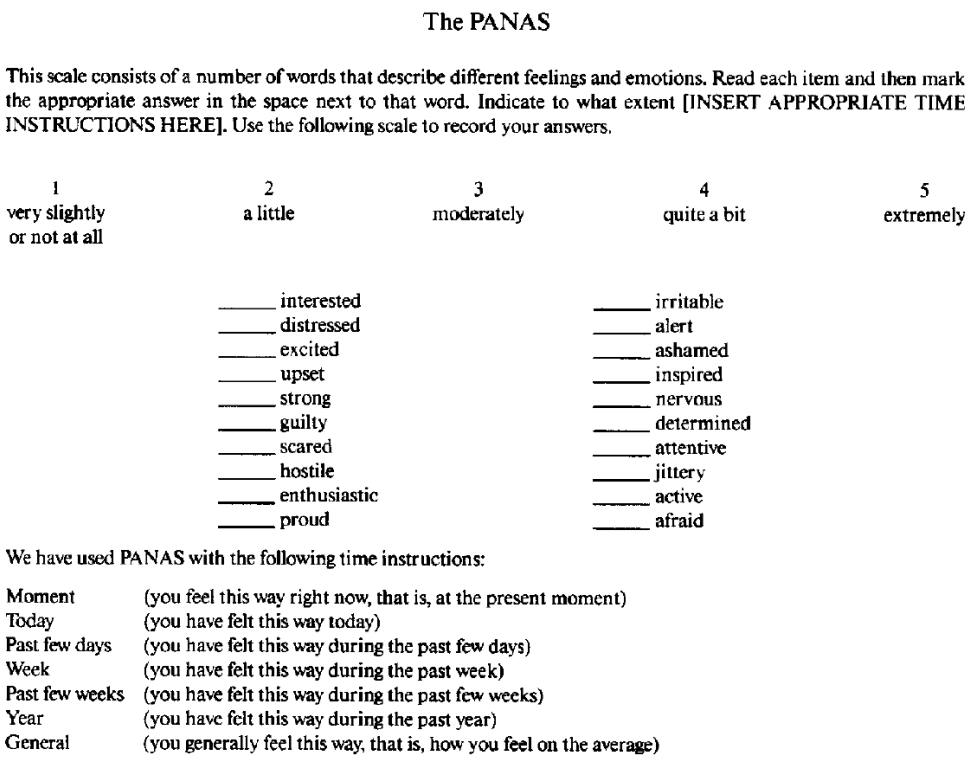


Figure 3.4: *Self Assessment Manikin: the top row of figures represents the Valence scale, the middle row represents the Arousal scale, and the bottom row represents the Dominance scale.*

The PANAS scales Watson et al. [205] have developed two scales with textual descriptions of different feelings and emotions for the measurement of *mood*, see Fig 3.5. Although mood is currently not investigated in this thesis, we nevertheless briefly describe these scales because they show some similarities with the other affect measures. The PANAS scales are based on two mood factors that been used more extensively in the self-report mood literature. The first scale measures the mood factor Positive Affect (PA). PA can be described as the extent to which a person feels enthusiastic, active, and alert. High PA is a state of full energy, full concentration, and pleasurable engagement, whereas low PA is characterized by sadness. The PA scale is described by 10 terms: *attentive, interested, alert, excited, enthusiastic, inspired, proud determined, strong* and *active*. NA can be described as unpleasurable engagement that subsumes a variety of aversive mood states, including anger, contempt, disgust, guilt, fear, and nervousness, with low NA being a state of calmness and serenity. The NA scale is described by the following 10 terms: *distressed, upset (distressed), hostile, irritable (angry), scared, afraid (fearful), ashamed, guilty, nervous* and *jittery*. The two scales proved to be a reliable, valid, and efficient means for measuring Positive Affect and Negative Affect of mood Watson et al. [205].

The tools described here offer ways to describe natural emotion in verbal or symbolic descriptions. Although these tools have proven their success in other studies, the annotation of natural emotion occurring in the field with these tools is not at all a straightforward process. It remains difficult to specify the naturally occurring emotions observed and to fit these into annotation schemes offered by the tools described.



The PANAS

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent [INSERT APPROPRIATE TIME INSTRUCTIONS HERE]. Use the following scale to record your answers.

1	2	3	4	5
very slightly or not at all	a little	moderately	quite a bit	extremely

_____ interested	_____ irritable
_____ distressed	_____ alert
_____ excited	_____ ashamed
_____ upset	_____ inspired
_____ strong	_____ nervous
_____ guilty	_____ determined
_____ scared	_____ attentive
_____ hostile	_____ jittery
_____ enthusiastic	_____ active
_____ proud	_____ afraid

We have used PANAS with the following time instructions:

Moment	(you feel this way right now, that is, at the present moment)
Today	(you have felt this way today)
Past few days	(you have felt this way during the past few days)
Week	(you have felt this way during the past week)
Past few weeks	(you have felt this way during the past few weeks)
Year	(you have felt this way during the past year)
General	(you generally feel this way, that is, how you feel on the average)

Figure 3.5: *The PANAS scales for measuring mood on Positive Affect and Negative Affect dimensions.*

3.2 Acquiring natural emotion data in the field

This Section elaborates on how to measure naturally occurring emotions in the field and how to establish a description of these emotions with the aim to develop automatic emotion recognizers: three data collection attempts in three different real-world, natural environments will be presented. Three experiments were carried out by TNO, DECIS lab and the V2 institute who all had their own goals defined, and did not specifically have the intention to capture vocal and facial expressions for the purpose of developing an affect recognizer. Since the scenarios used in these experiments could evoke a substantial amount of naturally occurring emotional behavior (although these experiments were not specifically designed to evoke affect), we decided to participate in these experiments and record vocal (and where possible facial) expressions. The disadvantage was that we had to adhere to the original experimental setup of the hosts of the experiments, and hence, no additional changes could be made upon our request. For example, using headsets with a close-talk microphone to make high-quality individual speech recordings instead of using a desktop microphone was not an option in the experiment about crisis meetings. We describe the experiments and discuss our experiences and 'lessons learned' in acquiring emotion data in the field.

3.2.1 Measuring task load during emergency situations on a naval ship

In Grootjen et al. [71], an experiment is described in which the goal was to measure cognitive task load by processing several measurements, including vocal and facial expressions. The task of the operator in the ship control center was to deal correctly with the emergency situations that occurred on the naval ship, e.g., fire or platform system failures. The type and frequency of these emergencies were controlled via several scenarios that were designed to evoke low, medium or high task load with the operators. High-quality webcams and head-mounted microphones were used to record video and audio. After each scenario, the operators (each minute) had to rate task complexity and subjective effort on a five point scale. The idea was to find correlations between task load (or stress) and vocal and facial measurements. However, this was easier said than done. Firstly, the audio and video signals were initially recorded for the purpose of monitoring, not for speech processing or facial analysis (with the goal to develop affect recognizers), and were therefore noisier than expected. The speech signal was sometimes clipped and contained interfering background noises and crosstalk (i.e., softer speech from other speakers). The recordings of the facial expressions could not always be processed with automatic facial recognition software, such as the FaceReader (see den Uyl and van Kuilenberg [51] and [61]). Lighting conditions, moving head poses, clutter in the background etc. made it difficult for the FaceReader to analyze these faces reliably, see Fig. 3.6. Secondly, after viewing a couple of the recorded sessions that were each approximately 15 minutes long, it became clear that there were very few vocal and facial emotional expressions observed. In addition, with a rate of 1 rating per minute, the ratings were difficult to relate to vocal and facial expressions which can occur and change with a much higher rate. Therefore, the recordings of this experiment were rendered unfit for the development of automatic affect recognizers.

Figure 3.6: *A video still of a facial expression during an experiment on task load, processed by the FaceReader.*

3.2.2 Measuring affect during time-pressured crisis meetings

In this exercise, that was set up by TNO and the DECIS lab[2] (Delft Cooperation on Intelligent Systems), the goal was to investigate how the stream of incoming and outgoing messages in different cooperating teams was handled by professionals in a crisis situation. The crisis situation was simulated with a 1-day scenario. The experiment took place in the city hall of a small town in The Netherlands. The crisis situation simulated was a flood disaster caused by a dyke breach. Emergency centers and protocols were activated and were led from a distance by a crisis policy team that resided in the city hall. The crisis policy team had to make time-pressured decisions and deal with issues that arose as a consequence of the dyke breach, e.g., flooded highways, drowning cattle, possible evacuations, worried citizens etc. The team consisted of 8 main participants including the mayor, heads of the fire and police departments, the head of public safety and an employee of the department of public relations and communication. Each half hour, the team members reported on updates from their departments in a 15-minute long meeting. The meetings were led by the mayor who made the final decisions. Five meetings were recorded using an 8-channel circular microphone array that was positioned at the middle of the meeting table. Due to the realistic setting of the exercise, there were a few conditions that had an impact on the recording of the audio: firstly, no head-mounted microphones could be used for the recordings, and secondly, real names of citizens were used during the exercise which limits the usage and the distribution of the data recorded. The author of this thesis was present to observe and to perform a rough, first step annotation of vocal emotional expressions.

In crisis situations where decisions are made under time pressure, we expected to find emotion-related expressions, e.g., stress or frustration, in the speech or facial expressions of the participants. However, this appeared to be less apparent than expected. Only a few mild instances of frustration or irritation or laughter were found, on the average less than 6 per meeting session. This is not sufficient for the development of an automatic affect recognizer or a reliable statistical vocal analysis. The meetings were firmly led by the mayor who was strict and clear. Obviously, the task of the leader is to lead the meeting in a structured manner and to make decisions based on the facts he/she has. There simply was no time for emotions. However, the data recorded could still be of interests for researchers working in the domain of discourse and dialog act analysis.

3.2.3 Measuring affect with players in a virtual reality game

In the context of the MultimediaN project, one of the project partners V2_lab (Institute for the Unstable Media[3]) created *Exercise in Immersion 4* in cooperation with the artist Marnix de Nijs[4]. *Exercise in Immersion 4* (EI4) is a virtual reality art-game played in an existing physical space. In order to achieve this mix of reality, each game-player wears

[2] http://www.decis.nl

[3] http://www.v2.nl

[4] Marnix de Nijs (see http://www.marnixdenijs.nl) is an artist who explores the dynamic clash between bodies, machines and other media. The interface between the body and technology forms an important basis for his work.

a VR (virtual reality) head-mounted display and a so-called crash suit, specifically designed for this game, see Fig. 3.7. The player is "in between" two worlds, the virtual one and the physical one: the objects in the game correspond to real physical objects. Gradually, the virtual environment in the game shifts and changes, which leads to an increasingly disorienting experience because the senses no longer correspond to each other. One would expect to find emotional behavior from players who undergo this experience and who immerse in the game. Therefore, during the DEAF07 festival where EI4 was showcased, participants who played the game were asked to wear a wireless head-mounted microphone (in addition to the head- mounted display) and to "think aloud" and express their feelings and emotions freely. Afterwards, the players filled in two questionnaires that were related to the emotions felt and the amount of presence (i.e., the subjective sense of being in a virtual environment) experienced. With respect to emotions, the participants were asked to fill in how aroused, and how positive or negative they felt. In total, 9 players participated in this experiment. Unfortunately, the number of spoken emotional expressions found with the players was relatively low. More importantly, the players did not sound natural due to the "think aloud" procedure, some of them reported that it was awkward to "think aloud", even though for other disciplines, e.g., psychology, this procedure appears to work.

Figure 3.7: *Exercise in Immersion 4: a virtual reality game.*

3.3 Summary and conclusions

In this Chapter, the experiences from three data collection efforts, situated in real-world environments in which affect was expected to play a role, were described. The data collected did not prove to be useful for the study of this thesis, but can be valuable as "lessons learned". The main conclusion is that in real-life professional working environments and situations such as a control center on a naval ship or a crisis meeting held at city hall, affect is not frequently overtly expressed, even though these are environments in which affect was expected to play a role. Possible reasons for this 'lack of affect' are that in professional working environments such as the ones

assessed, the people involved are professionals who are trained to deal with emergencies and crisis situations. In these situations, they are probably taught to suppress their emotions. In addition, the quality of the signals recorded in these environments is usually low which complicates the analysis of these signals. In the case of measuring user experience in a virtual game, the 'think aloud' procedure produced unnatural speech with the participants. As an alternative to these unfortunate attempts to collect naturally occurring speech data, we suggest that an intermediate method to collect emotion data, such as emotion elicitation or Wizard-Of-Oz experiments, could be a solution for solving the 'lack of affect' and the bad signal quality. As part of our goal to develop a speech-based affect recognizer with spontaneous emotion data, we collected spontaneous emotional speech data that was elicited with subjects playing videogames. This corpus is described in Chapter 6.

Chapter 4

Emotion recognition in acted speech: adopting the detection evaluation framework

The main goal of this Chapter is to present existing evaluation methodologies from similar recognition technologies and to apply these to the field of emotion recognition where a benchmark style evaluation is still an underexposed topic. We evaluate our emotion recognizers in a **detection** evaluation framework which is not very commonly used in the emotion recognition community, but that is commonly used in similar recognition technologies such as speaker and language recognition. Conceptually, a detection evaluation approach fits the 'real' task of an affect recognizer very well: as explained in section 2.3 and Banse and Scherer [12], an affect recognizer should aim at 'true' affect recognition rather than emotion discrimination among a small number of alternatives. Using a relatively commonly used database, the BERLIN emotional speech database (Burkhardt et al. [25]), and state-of-the-art detection technology, we show what the performances of today's emotion classifiers are when clean, acted emotional speech is used. We compare several types of acoustic features and several types of classifiers to each other and fuse the best performing classifiers in order to optimize detection performance.

The second goal of this Chapter is to improve the ecological validity of lab classification experiments that are traditionally carried out with non-exhaustive sets of discrete emotion categories containing acted emotional speech samples. We show how this improvement can be achieved in a detection evaluation framework.

The motivation to adopt the detection evaluation framework is explained in Section 4.1. In Section 4.3, the material used in the classification experiments is described. Several state-of-the-art machine learning methods used are described in Section 4.4. We explain the detection evaluation methodology adopted in this study and applied to emotion recognition in Section 4.5. The results of the classification experiments are presented in Section 4.6. We also touch upon a fairly new and emerging emotion recognition approach that aims at prediction of emotion in terms of the emotion dimensions Arousal, Valence and Dominance (Section 4.7). Furthermore, we present an 'open-set' evaluation procedure that simulates the occurrences of 'new, un-

known' emotions that are 'unseen' by the classifier (Section 4.8) which will render the performance results of this evaluation less dependent on the types of emotions available in the database. This procedure presents a way to evaluate an emotion recognizer in a manner that is closer to real-life situations where 'out-of-set' emotions (i.e., emotion classes that were not present in the database and hence, were not modeled) can occur. Finally, we take full advantage of the characteristics of the detection framework and show how similarities between emotions, as defined in the detection framework, can be visualized in a 'map of emotions' (Section 4.9).

*4.1 Motivation for emotion **detection***

Why are we so keen on using a detection framework for emotion classification? First of all, this is given in by the growing need from application domains to develop single-emotion detectors that are tuned to specific emotions. For example, call centers are interested in detecting frustrated or angry people. They often formulate their questions as binary choices such as "Is this person frustrated or not?" or "We want to know if this person is angry or not?" rather than "Is this person angry, sad, happy or neutral?". These types of binary tasks fit the detection framework very well. Secondly, these formulations of binary recognition tasks in the detection evaluation methodology have as advantage that these tasks better reflect the concept of 'true' recognition than the traditional classification paradigm. Rather than emotion *discrimination* between a small number of classes, where the outcome is much dependent on the types of emotion classes used, we want to move towards 'true' emotion *recognition* (see also section 2.3 and Banse and Scherer [12]). Furthermore, in detection, the prior is taken out of the problem. In classification, the prior is implicitly known. Thirdly, there is a need for a more sound and shared evaluation methodology. For similar recognition technologies such as speaker recognition ("Is this person Bill Clinton or not?") and language recognition ("Is this English or Mandarin or Dutch?"), shared evaluation protocols already exist and international benchmarks are being organized by NIST (the National Institute of Standards and Technology[1]). The task of language recognition in particular seems to be somewhat similar to that of emotion recognition: given a speech sample, detect the emotion/language where an N number of emotions/languages are possible. In essence, this is a multiclass classification problem, but in the detection framework, this problem can be presented in binary tasks as is done in language recognition. So by adopting this detection framework that provides sound and shared evaluation methods and tools, we hope to advance towards a more standardized and shared task for automatic emotion recognition. The evaluation procedure of our emotion classifiers is shortly explained in section 4.5, for a more detailed explanation of the detection evaluation methodology the reader is referred to van Leeuwen and Brümmer [196].

[1] http://www.nist.gov

4.2 Related work

We selected the BERLIN Emotional Speech database (Burkhardt et al. [25], and see Section 2.4.1) that has been used relatively frequently by other researchers as well to make comparison possible. We attempted to compare the performances of various speech-based emotion recognition studies using the same database, and summarized the results in Table 4.1. It still appears to be difficult to compare these performances since the data sizes differ and in some cases, information about the test protocol is missing. In Table 4.1, the accuracy is computed as the number of correct classification divided by the total number of trials. The F_{emo} in this table refers to a performance measure used in e.g., Schuller et al. [173], which is calculated as $2 * CL * RR/(CL + RR)$, where RR is the accuracy as defined above, and CL is the mean of the per-class-accuracy.

Study	Data size	SI	A	F_{emo}
Xiao et al. [213]	286	?	65.7	-
Vlasenko et al. [202]	494	yes	89.9	-
Schuller and Rigoll [170]	488	?	96.6	-
Schuller et al. [173]	494	yes	72.3	69.8
Shami and Verhelst [174]	494	?	75.5	-
Lugger and Yang [111]	694	yes	72.8	-

Table 4.1: *Speech-based emotion recognition studies that have used the Berlin Emotional Speech database, SI=Speaker-Independent, A=accuracy (number of correctly classified samples divided by total number of samples),* F_{emo}.

Vlasenko et al. [202] achieved an accuracy of 89.9%. Frame-level spectral features (MFCCs) were used in combination with GMMs. Turn-level features were used with SVMs. A combination of frame-level and turn-level information was accomplished by adding the final GMM scores as features to the turn-level feature vector. Finally, speaker normalization was applied which contributed to this relatively high upper benchmark accuracy.

On a different subset of the BERLIN database, Schuller and Rigoll [170] were able to achieve an accuracy as high as 96.6%. They compared different segmentation schemes and classifiers and constructed a 'super feature vector' from different levels of segmentations which appeared to work best. However, we do not know whether the classification experiments were performed speaker-independently.

Shami and Verhelst [174] extracted series of pitch, intensity, lowpass intensity, highpass intensity and MFCCs and calculated statistics over these series which resulted in a feature vector of 200 features. Employing these features in an SVM, an accuracy of 75.5% was achieved. However, there is no information about whether the classification experiments were carried out speaker-independently or not.

Schuller et al. [173] used approximately 4000 acoustic features extracted at utterance level in combination with Random Forests. With speaker-independent classification experiments, they achieved an accuracy of 72.3%, and an F_{emo} of 69.8%. The experimental setup in this study seems to be best comparable to our current study.

Campbell et al. [33] developed a method that employs a GMM 'supervector' for SVM learning and that was applied to speaker recognition. The GMM supervector based SVM system appeared to be very competitive with a standard GMM UBM system. In Hu et al. [81], the GMM supervector based SVM method was applied to emotion recognition in acted emotional speech. The GMM supervector based SVM method outperformed the standard GMMs substantially. In our study, we used this GMM supervector based SVM method in combination with RPLP features, see Section 4.4, and we adopted a detection approach to recognize acted emotions in speech.

4.3 Data used in experiments

The BERLIN Emotional Speech database (Burkhardt et al. [25] and Section 2.4.1) contains emotional speech (German) from five female and five male actors, uttered in seven different 'basic' emotions namely Anger (An), Joy (Jo), Fear (Fe), Disgust (Di), Boredom (Bo), Sadness (Sa), and Neutral (Ne). The whole database comprises 816 utterances. In a validation test (carried out by the makers of the database themselves), 20 subjects were asked to classify the utterances and to rate their naturalness. In our analyses, we used only those utterances that had a human recognition accuracy of more than 80% and a rated naturalness of more than 60%. This decreased the number of utterances used in the current study to 494 (see Table 4.2). We can observe in Table 4.2 that a large proportion of Disgust, Sadness, and Fear samples were judged less natural or not recognizable by human listeners. The averaged human recognition accuracy for this selected set of speech utterances ($N = 494$) is 94.9% (an F_{emo} of 94.2%). The averaged human recognition accuracy calculated over the whole database ($N = 816$) is 85.4% (F_{emo} of 85.3%).

Emotion	N_{orig}	N_{filtered}	$\%_{\text{removed}}$
Anger (An)	137	127	7.3
Joy (Jo)	115	64	44.3
Fear (Fe)	122	55	54.9
Disgust (Di)	105	38	63.8
Boredom (Bo)	112	79	29.5
Sadness (Sa)	121	53	56.2
Neutral (Ne)	104	78	25.0
Total	816	494	39.5

Table 4.2: BERLIN *Emotional Speech database,* N_{orig} *=original number of utterances,* N_{filtered} *=number of utterances filtered by criterion (more than 80% correct human recognition and more than 60% rated naturalness) and used in the current study,* $\%_{\text{removed}}$ *=percentage of utterances removed.*

4.4 Method and features

We employed several modeling techniques in combination with frame-level and utterance-level acoustic features. Frame-level features (spectral) were used in combination with

Gaussian Mixture Modeling (GMM), and utterance-level features in combination with SVMs. Linear Discriminant Analysis (LDA) served as either a back-end to the binary detectors or as a fuser. Another fuse method we applied was a linear combination of the separate scores. In Fig. 4.1, an overview of all methods and features used in this chapter is given. Finally, we explain how we performed multiclass classification with our binary emotion detectors developed in a detection framework.

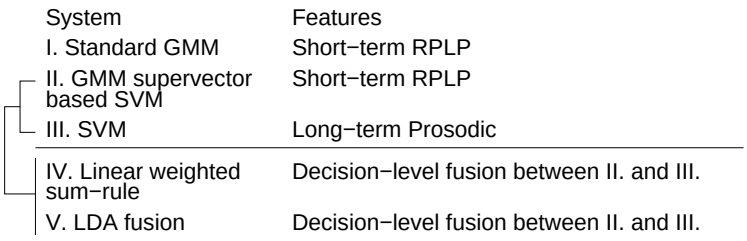


Figure 4.1: *Overview of methods used in this chapter to recognize basic emotions.*

4.4.1 Three 'single' systems

Here, we describe the three 'single' systems that were developed:

I. Standard GMMs using frame-level spectral features

Gaussian Mixture Modeling (GMMs, see also Section 2.4.3) is a very popular machine learning technique in speech technology and stands at the basis of many good results. The GMMs are trained using the Expectation Maximization (EM) algorithm (five iterations). A binary emotion classification scheme is adopted. For each *target* emotion, a pair of GMMs is trained: a target GMM (i.e., trained with the emotion that one would like to detect, e.g., Anger) and a non-target GMM (i.e., all the other emotions that are not the target emotion, e.g., not-Anger). In testing, the log ratio between the likelihoods of the two GMMs are used as 'soft decision' scores. A visualization of this method is given in Fig. 4.2.

As speech features, RASTA-PLP features (RPLP, Hermansky and Morgan [78], see also SPRACHCORE [3] and Section 2.4.2) are used that are extracted each 16 ms over an analysis window of 32 ms. Twelve rasta-PLP features plus one log energy component and their deltas (the first order derivatives of the 13 RPLP features) are computed which gives a total of 26 features. The features are normalized to z-scores per utterance such that μ and σ of all features are 0 and 1 respectively for each utterance: $z = (x - \mu)/\sigma$ (this normalization procedure is also called z-scoring or z-normalization). We will refer to this method as 'Standard-GMM-rplp'.

II. GMM supervector based SVM using frame-level spectral features

More recently, the standard GMM approach has been extended with a Support Vector Machine (SVM) concept (Campbell et al. [33]), and improved results have been achieved with this method in the area of language and speaker recognition (e.g.,

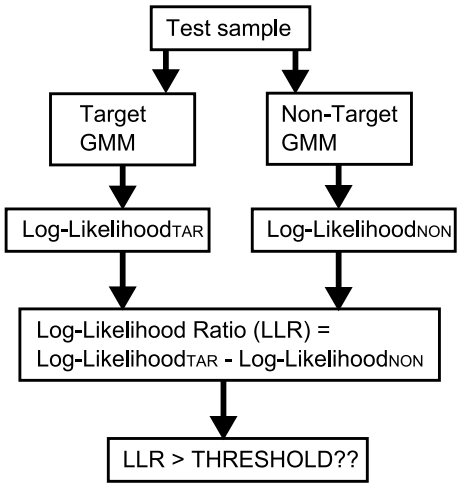


Figure 4.2: *Standard GMM method.*

Campbell et al. [33]), and it has also been successfully applied to emotion recognition (Hu et al. [81]). The combination of a generative learning algorithm (GMM) and a discriminative learning algorithm (SVM) appears to be very fruitful. The basic idea is to use the means of a GMM for SVM learning (see also Section 2.4.3). First, a Universal Background Model (UBM) is trained using the Expectation Maximization (EM) algorithm (five iterations) with all the emotions that are available in the emotion database. From this UBM, adapted GMMs are constructed (by MAP adaptation of the means of the UBM, see Reynolds et al. [149]) for each emotional utterance. The mean of each Gaussian component of each emotional utterance is stacked in a *supervector* and forms the input for SVM learning. Using the GMM KL divergence kernel according to Campbell et al. [33]

$$\begin{aligned} K(utt_a, utt_b) &= \sum_{i=1}^{N} w_i \mu_i^a \Sigma_i^{-1} \mu_i^b \\ &= \sum_{i=1}^{N} (\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^a)^t (\sqrt{w_i}\Sigma_i^{-\frac{1}{2}}\mu_i^b) \end{aligned} \tag{4.1}$$

where w_i, μ_i and Σ_i are the Gaussian parameters of the ith centroid, a score can be obtained by computing a single inner product between the target model and the GMM supervector. We used the same speech features as in the Standard-GMM-rplp approach (13 RPLP features and their derivatives, extracted each 16 ms over an analysis window of 32 ms). We refer to this method as the 'GMM-SV-SVM-rplp' method, i.e., a GMM supervector SVM based method trained with RPLP features. In Fig. 4.3, a graphical summary of this approach is given.

III. SVM using utterance-level prosodic features

GMMs seem to work best with frame-level spectral features in which the average log likelihood per frame gives a robust estimate of the likelihood of the data given the model. In Truong and van Leeuwen [187], we have used GMMs in combination

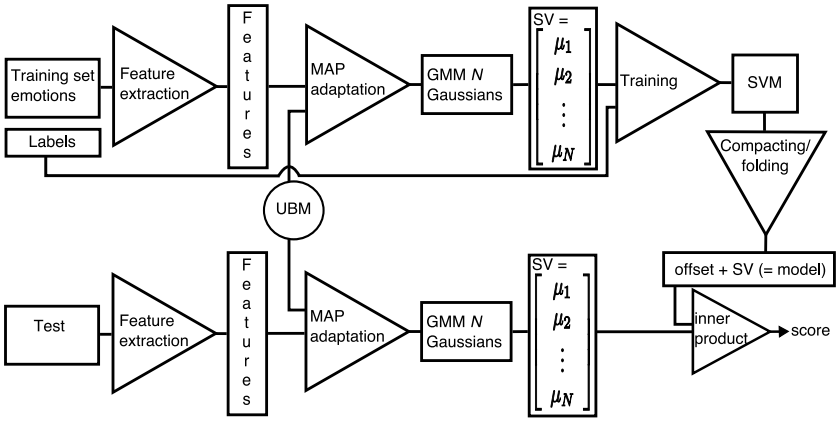


Figure 4.3: *GMM Supervector based SVM method.*

with local pitch and energy values. However, our results showed that these types of prosodic features were better modeled in a Support Vector Machine (SVM). Prosodic features usually have slow-varying acoustic characteristics and can be more informative when computed over a meaningful unit (in our case, a meaningful unit is a sentence). Therefore, we used utterance-level prosodic features in combination with SVM.

SVMs (see also Section 2.4.3) can be typically used for binary classification problems (multiclass classification for SVMs is also available). In our case, we trained seven SVMs, for each target emotion one SVM, for which the task is to discriminate the target emotion from the non-target emotion (the rest of the other emotions). We used *libsvm* [37] for our detection and classification experiments. Based on previous emotion recognition studies, see e.g., Schuller et al. [172], Vogt and André [203], Pittam et al. [135], we selected a number of acoustic features. All acoustic features were extracted with Praat (Boersma and Weenink [23]) and normalized to z-scores, with μ and σ calculated over a development set of utterances. The SVM RBF (radial basis function) kernel was used of which the cost and gamma parameters were optimized on the development set. Table 4.3 lists the features used in this method. The scores

Pitch-related	standard deviation, range (max − min), mean absolute slope
Intensity-related	standard deviation, range (max − min), mean absolute slope
Spectrum-related	slope LTAS, Hammarberg index (difference in max LTAS energy between 0–2 kHz and 2–5 kHz), center of gravity of spectrum, skewness of spectrum
Speech rate	articulation rate (#nuclei/total duration nuclei)

Table 4.3: *Acoustic features used in SVM-Praat method.*

that are produced with the SVM are used as soft decisions. We will refer to this method as 'SVM-Praat'.

4.4.2 Two fused systems

In order to achieve higher performance, two systems were fused: the GMM-SV-SVM-rplp system (II.) and the SVM-Praat system (III.) (since these two single systems appeared to perform best). Since each of the two systems is based on different techniques and uses different information available from the speech signal, it makes sense to combine the two systems and to see whether the performance can be improved. Firstly, the GMM supervector SVM based system uses both generative (GMM) and discriminative modeling techniques, while the SVM-Praat system is based on discriminative learning alone. Secondly, the GMM-SV-SVM system uses frame-level (spectral) features, while the SVM-Praat system uses statistics of prosodic features calculated over the whole sentence. Previous studies have already shown that using both spectral and prosodic features can improve classification performance (e.g., Barra et al. [14], Vlasenko et al. [202]). Fusion is performed on *decision-level*: this means that the *score output* of both systems are combined rather than the *features*. Fusion on *feature-level* is also an option but the advantage of decision-level fusion is that each detector can be optimized separately. In this Section, the decision-level fusions between systems II. and III. are presented: one fusion employs a relatively simple linear weighted sum-rule (system IV.), while the other fusion uses Linear Discriminant Analysis as a fuser (system V.).

IV. Fusion using a linear weighted sum-rule

We first employed a linear combination of the soft decision scores (see Fig. 4.4 and Eq. 4.2), sometimes also called the 'sum-rule':

$$S_{\text{fused}} = (1-\alpha) * S_{\text{systemA}} + \alpha * S_{\text{systemB}} \tag{4.2}$$

Prior to this fusion, we have to take into account the fact that the scores from the two systems do not necessarily have the same scale or range, which means that a form of normalization needs to be applied. So first, prior to any form of fusion, the scores were normalized to $\mu = 0$ and $\sigma = 1$ as we have done for our speech features ('z-norming').

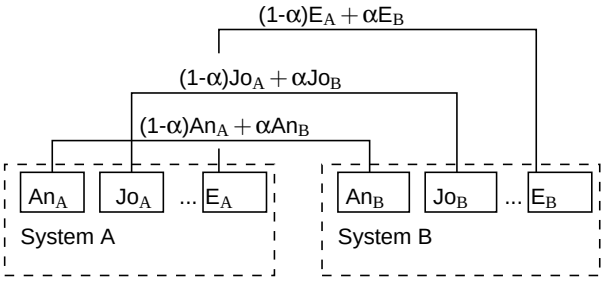


Figure 4.4: *Decision-level fusion using linear sum-rule.*

V. Fusion using Linear Discriminant Analysis

As an alternative to linear fusion, Linear Discriminant Analysis (LDA) was used as a fuser to find the best linear weighted combination of the $2N$ number of scores (see Fig. 4.5). Similar to linear fusion, the scores were first normalized to z-scores before fusion takes place.

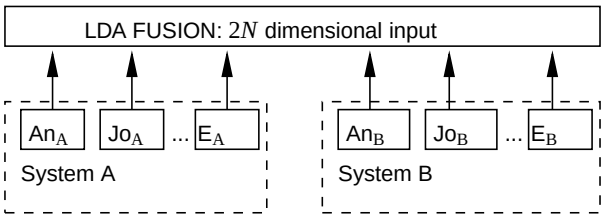


Figure 4.5: *Decision-level fusion with LDA as fusion mechanism.*

4.4.3 From detection to classification: a comparison

In detection, a binary classification is made between a target emotion and a non-target emotion. In this way, an N number of emotion detectors can be developed independently from each other and each detector gives an independent judgment about whether the sample belongs to the target or non-target emotion (in our case $N = 7$). However, the classical approach in emotion recognition is based on forced choice classification, i.e., the machine makes a forced choice between N (usually > 2) number of emotion classes.

In order to make comparison possible between our detection results and the more traditionally reported classification results, we will also perform classification with our binary detectors and report classification results. There are several ways to go from binary decision scores to multiclass classification.

The first method is relatively simple. For each binary detector, a score is obtained that indicates the amount of support for a certain emotion. At decision level, we compare these soft decision scores of the N separate emotion detectors to each other and maximum likelihood determines the final class. Prior to this comparison, the scores need to have the same scale and thus need a form of normalization. The most straightforward method is to normalize the scores (individually for each target emotion detector) such that $\mu = 0$ and $\sigma = 1$ ('z-norming'). Another normalization option is to use the optimal decision thresholds (that were found on an independent development set) belonging to each emotion detector and subtract these from the scores; we call this normalization option 'threshold-norming'.

A second method to go from binary soft decision scores to multiclass classification is to use Linear Discriminant Analysis (LDA) as an back-end that finds an optimal weighted linear combination of the soft decision scores. Then the emotion class with the highest posterior probability is chosen, at uniform prior distribution. Note that LDA here has been used for two different purposes: a) for the purpose of fusion, and b) for the purpose of going from decision scores to final multiclass classification.

4.5 Evaluation

As performance measures, we report Equal Error Rates (EER) and detection costs with equal priors and equal costs (C_{det}). For reference and comparison, we also report the traditional classification accuracy A and F_{emo}. In section 2.4.4, we have explained how EER, C_{det}, A, and F_{emo} are calculated. For the reader's convenience, the calculation of the evaluation metrics EER an C_{det} are shortly recapitulated here.

4.5.1 Detection performance measures

One of the goals of evaluation is to assess a system's performance and compare its performance to the performances of other systems. It is very convenient to be able to summarize and express the performance in a single figure. In traditional emotion classification studies, classification accuracy, defined as the number of correctly classified test samples divided by the total number of test samples, is usually given. One of the disadvantages of this measure is that it depends on the number of test samples in each emotion class, i.e., it is sensitive to evaluation priors (skewed class distributions), and the number of target classes. Therefore, we favor the use of detection evaluation measures which we will explain below (for a more thorough introduction in detection evaluation measures readers are referred to van Leeuwen and Brümmer [196]).

For each test sample that is tested against a detector, a score (e.g., log likelihood ratio, posterior probability) can be computed; the higher the score, the more support there is for the hypothesis that the test sample belongs to the target class (and vice versa). Deciding to which class the sample belongs, involves setting a well-chosen threshold, a process known as calibration. When a threshold is set, the two types of errors can be count and a false alarm rate (P_{fa}) and miss rate (P_{miss}) can be computed. Changing the threshold will also change the false alarm and miss rate: this tradeoff between the rates (as a consequence of changing the threshold) can be visualized in a Detection Error Tradeoff (DET) plot (see for example Fig. 4.10). A performance measure of discriminability that is widely used to summarize the DET curve is the Equal Error Rate (EER); the point where the false alarm rate is equal to the miss rate.

However, the EER is based on an optimal decision threshold which is set a posteriori. In a more realistic situation, a threshold has to be set a priori. In that case, we can measure the performance by the Detection Cost Function (DCF):

$$C_{\text{det}} = C_{\text{miss}} P_{\text{miss}} P_{\text{tar}} + C_{\text{fa}} P_{\text{fa}} (1 - P_{\text{tar}}) \tag{4.3}$$

C_{miss} and C_{fa} are the costs that one can attribute to a certain type of error. P_{tar} is the prior probability that the target emotion occurs. In our evaluation, we set the cost parameters $C_{\text{miss}} = C_{\text{fa}} = 1$ to have equal costs, and we set equal prior probability $P_{\text{tar}} = 0.5$. This specific case of C_{det} is also known as the Half Total Error Rate (HTER). For the calculation of C_{det}, the decision threshold was optimized at EER obtained with development data. Note that these are *error* rates: the lower EER and C_{det}, the better the performance.

4.5.2 Other performance measures

In many emotion recognition studies, the performance of an emotion classifier is usually assessed by showing a confusion matrix that gives insight in what types types of misclassifications (confusions) frequently occur. As a performance measure, the accuracy (A) is computed that is defined as the number of correct classifications (the diagonal in the confusion matrix) divided by the total number of test samples. In addition, an accuracy per emotion class can be given. In order to make comparison possible, we will also use a so-called F_{emo} measure, as used in e.g., Schuller et al. [173] which is computed as the uniformly weighted harmonic mean of RR and CL: $2 * CL * RR/(CL + RR)$, where RR is the accuracy A as defined above, and CL is the mean of the per-class-accuracy, see section 2.4.4. This is actually not the same F measure as we know from fields as information retrieval.

4.5.3 Cross-validation evaluation procedure

In order to carry out the evaluation of the emotion recognizers as sound as possible, we need three mutually exclusive independent datasets. For each speaker, the dataset is partitioned into three subsets: a training set, a development set and a test set. This was done a) to perform speaker-independent experiments, and b) to have independent development sets for finding optimized decision thresholds, for normalization of the decision scores and features, and for training an LDA (fuser). To achieve that with a relatively small emotional speech database (as opposed to speaker or language recognition where hundreds hours of speech data is available), we perform a double cross-validation experiment in which each training fold not one, but *two* speakers are left out for development and testing. In the inner-loop (nine 'inner-jacks'), the development takes place where classifiers are trained on a dataset excluding the development speaker *and* the test speaker. Development tests are carried out on the development speaker. The inner-loop rotates over the development speaker and collects all the scores for assessing the performance of the development set. In the outer-loop (ten 'outer-jacks'), the actual test takes place using the test speaker that has been left out of the training and development phase. The outer-loop rotates over the test speaker. Note that this scheme is only applied in cases where a development set is needed to find optimal parameters and optimal decision thresholds, to perform normalization or to train an LDA back-end or fuser. If this is not necessary, a single loop over the test speaker suffices, in each loop leaving out the test speaker from the training set in order to perform the actual test on the test speaker (speaker-independent-only experiment).

Note that this is the experimental setup for a closed-set detection experiment. Closed-set assumes a fixed set of classes and that there are no out-of-vocabulary classes. However, in real-life situations, especially in emotion recognition, chances are high that 'out-of-set' emotions do occur. Therefore, we present in Section 4.8 another cross-validation scheme that emulates an 'open-set' situation.

4.6 Results

To summarize, we have developed three basic systems, Standard-GMM-rplp, GMM-SV-SVM-rplp and SVM-Praat. The three basic systems can be optionally supplemented with an LDA backend for multiclass classification. Further, we have developed two fused systems that are a fusion between the GMM-SV-SVM-rplp and the SVM-Praat systems, based on either a linear or LDA fusion. We have defined several performance measures, C_{det}, A, and F_{emo} (for the sake of completeness and compatibility with other emotion recognition studies, we report all three of them). For readability, we will present the most important results in this section. For all results, including all performance figures, the reader is referred to the appendices.

One of the parameters that can be tuned in our systems is the numbers of Gaussian mixtures. We first look at how the number of Gaussian mixtures influences the detection performances of systems developed with GMMs. In Fig. 4.6, we can observe that the detection performances, expressed in C_{det}, increase with a larger number of Gaussians (the lower C_{det}, the better the detection performance). It seems that the ideal number of Gaussians used lies between 64 and 128. Therefore, in this section, we show results from systems that were developed with 128 Gaussian mixtures. Fig. 4.6 also shows that the fused systems perform best.

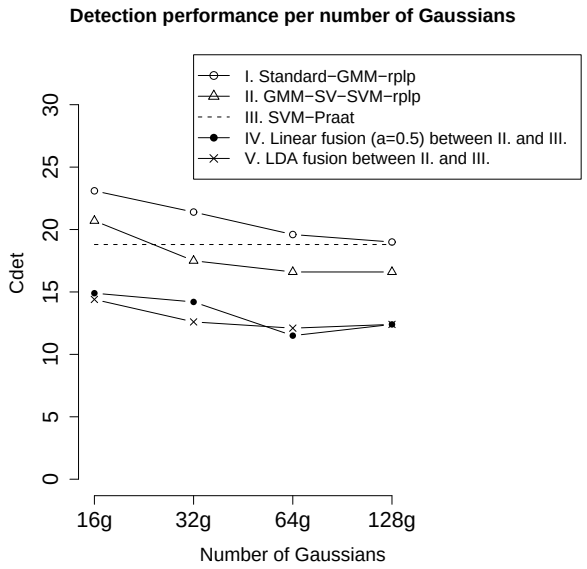


Figure 4.6: *Averaged detection performances as a function of number of Gaussians. The straight dotted line is the detection performance of the SVM-Praat system which is drawn for the reader's convenience, but is not influenced by the number of Gaussians.*

We also found that some emotions were better recognized with a specific system and its associated acoustic features. In Fig. 4.7, the detection performances of the five systems are shown per emotion. On the average, Sadness and Anger are the best detectable emotions in this database. Interestingly, the relatively simple SVM-Praat method appears to be very competitive with the GMM-SV-SVM-rplp system, except

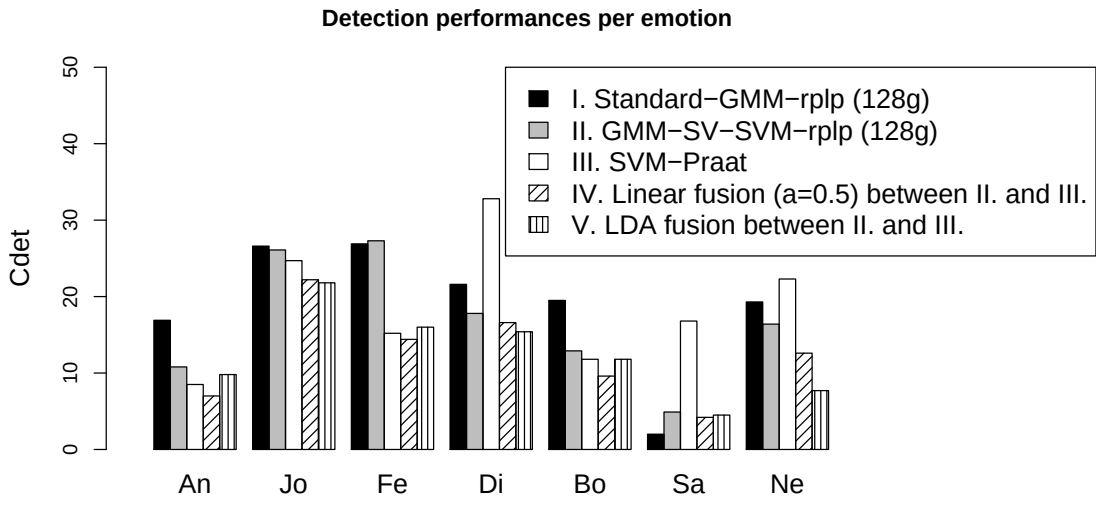


Figure 4.7: *Detection performances of five newly developed systems, shown per emotion (fusions were performed between the GMM-SV-SVM-rplp and SVM-Praat systems)*

System	C_{det}	A_{T}	$F_{\text{emo,T}}$
I. Standard GMM (128g)	19.0	65.6	65.5
II. GMM-SV-SVM-rplp (128g)	16.6	69.0	68.4
III. SVM-Praat	18.8	64.8	61.5
IV. Linear fusion of II. and III. ($\alpha = 0.5$)	12.4	75.1	74.1
V. LDA fusion of II. and III.	12.4	75.5	74.5

Table 4.4: *Results of newly developed speech-based emotion recognition systems: detection and classification, A_{T} and $F_{\text{emo,T}}$ are accuracy and F-measure respectively calculated using 'threshold-norming'*

for the emotions Disgust, Sadness and Neutral where SVM-Praat is performing worse than GMM-SV-SVM-rplp. On the other hand, SVM-Praat performed much better than GMM-SV-SVM-rplp in the case of Fear. It thus seems useful to treat each emotion separately and to develop separate single-emotion detectors. In all cases, a fusion of different systems has proven to be very effective: the stronger system can compensate for the weaker performing system, and combined together, the fused system outperforms the individual systems.

In Table 4.4, the performances of the five systems are shown in terms of C_{det}, accuracy A_{T} (multiclass classification accuracy based on 'threshold-norming' of the decision scores), and F_{emo}. It is clear from Table 4.4 and Fig. 4.7, that the Standard-GMM is much improved by GMM-SV-SVM-rplp, but the best performing systems are the ones that are fused: the averaged C_{det} obtained with SVM-Praat decreased from 18.8% to 12.4% when fused with GMM-SV-SVM-rplp. Note that we also tried different values for α in the linear fusion ($\alpha = 0.25$, $\alpha = 0.75$) but $\alpha = 0.5$ appeared to work best.

The confusion matrix of one of our best performing systems, system IV, is shown in Table 4.5 ($A_{\mathrm{T}} = 75.1\%$) in which we can observe what kind of misclassifications have occurred. There seem to be confusions between Joy and Anger, and between Boredom and Neutral. The confusion between Joy and Anger is a bit worrying since the two emotions lie on opposite sides of the valence scale (we also found this confusion in Truong and van Leeuwen [189]). From our studies in Truong and Raaijmakers [185], it became clear that the discrimination between spontaneous positive and negative emotions is mostly based on lexical content information. Separating positive from negative emotions based on acoustics only is still problematic.

	Predicted as						
	An	Jo	Fe	Di	Bo	Sa	Ne
An	106	8	8	4	0	0	1
Jo	18	36	4	6	0	0	0
Fe	0	4	39	4	0	2	6
Di	0	8	3	24	1	1	1
Bo	0	1	1	3	63	5	6
Sa	0	0	1	0	2	48	2
Ne	0	3	1	6	9	4	55

Table 4.5: *Confusion matrix of system IV, linear fusion between GMM-SV-SVM-rplp and SVM-Praat,* $\alpha = 0.5$*, 128 Gaussians.*

Comparing our classification results to related studies that have used the same database as well, see Table 4.1, we can observe that we achieve accuracies that are at least as well as the ones reported in previous studies. More specifically, Vlasenko et al. [202], Schuller et al. [173], Shami and Verhelst [174] used the exact same dataset and achieved 89.9%, 72.3% and 75.5% accuracy respectively, where we achieve 75.5% accuracy. We should point out that the 89.9% of Vlasenko et al. [202] is an 'upper bound' of the performance achieved (as the authors themselves have indicated) since speaker normalization was applied to the features using the whole speaker context: this is unrealistic since it would mean that each speaker has to utter a range of different emotions before normalization can take place. The study described in Schuller et al. [173] achieved an F_{emo} of 69.8%. Using the same dataset, we are able to obtain an F_{emo} of 74.5%, achieved with a fused system consisting of combinations of a) 'standard SVM' and a GMM supervector based SVM, and b) utterance-level prosodic features and frame-level spectral features.

Comparing these results to the performance of humans, we can conclude that humans are much better in recognizing emotions (the specific types of emotions that are under study): as shown in Table 4.1, humans are able to achieve accuracies between 85.4% and 94.4%.

4.7 Discrete emotions vs. emotion dimensions

Rather than adopting discrete emotion categories, a growing number of researchers is adopting a dimensional approach to emotion recognition. Main advantages of the

dimensional approach are that it allows a description of gradations of emotions, and it allows a language-independent description of emotion. Recently, speech-based emotion recognition systems have been developed that use regression techniques to predict emotions in terms of emotional dimensions such as Arousal, Valence and Dominance, e.g., Grimm et al. [68]. In Grimm et al. [69], Support Vector Regression is used to predict scalar values on Arousal, Valence, and Dominance scales. Detecting emotion in terms of emotion dimensions is a promising and attractive approach: it moves away from rigidly recognizing distinct emotion categories to predicting different grades of emotions on continuous Arousal/Valence/Dominance scales. An additional advantage is that each of these dimensions can be modeled separately if necessary. It has become apparent that Valence is difficult to model acoustically, while the expression of Arousal is typically voice-based (see e.g., Truong and Raaijmakers [185]). So for example, one could consider to use facial expression recognition technology for the Valence dimension and voice-based recognition technology for the Arousal dimension. In many ways, adopting the Arousal-Valence model offers much more flexibility.

Similar to Grimm et al. [69], we used Support Vector Regression (SVR) (Smola and Schölkopf [178]) to estimate emotion on continuous scales of Arousal and Valence, and applied this method to the BERLIN database. We focused on the Arousal dimension here because the Valence dimension is not well represented in the BERLIN database (there is only one positive emotion present namely Joy). In addition to regression, we also used ranking functions for emotion prediction. Ranking is based on ordered categories. Using the BERLIN speech data, a comparison is made between discrete emotion recognition techniques, regression and ranking techniques for speech-based emotion recognition.

Database As material, we used the exact same dataset from the BERLIN database (Burkhardt et al. [25]) as described in section 4.3, and Table 4.2. However, the discrete emotion labels need to be associated with an Arousal value and a rank order. We decided to use the locations of the Feeltrace (Cowie et al. [45]) landmarks as reference Arousal ratings and rankings, see Table 4.6.

Emotion	Arousal value	Rank order
Anger	0.75	1
Joy	0.52	2
Fear	0.13	4
Disgust	0.25	3
Boredom	−0.48	6
Sadness	−0.48	6
Neutral	0	5

Table 4.6: *Arousal value and ranking order adopted from the Feeltrace tool*

Method and features Support Vector Regression (SVR) (Smola and Schölkopf [178]) was used for the estimation/ prediction of Arousal values. We used the SVM-regression

function available from the *libsvm* toolbox (Chang and Lin [37]). As explained earlier, SVR is a regression method based on Support Vector Machine modeling, see section 2.4.3 and Smola and Schölkopf [178]. Rather than dealing with discrete outputs that are either $+1$ or -1, regression can deal with estimations of scalar outputs that for example lie in a range $[-1, +1]$. This has also been applied to emotion recognition, where there is a growing interest in predicting grades of emotions or grades of Arousal, Valence and Dominance (Grimm et al. [69, 68]). As speech features, the same set of features of the SVM-Praat method is employed. We refer to this method as 'SVM-Praat-regression'.

A disadvantage of having these real-valued outputs is that the estimation error is highly dependent on the ground truth value that can now be any real number; but, does Joy really have an exact Arousal value of zero point fifty-two? An alternative would be to assume ordered categories and focus on estimating the correct *ranking* of emotions on an Arousal scale. Ranking functions based on ordinal regression can predict categories on an ordinal scale based on their ranking information. Therefore, we also used ranking functions as implemented in *libsvm* (Chang and Lin [37]). Rather than using the scalar values for regression modeling, ordinal regression uses the ranking order of the scalar values. We used SVM-rank from *libsvm* to train a ranking function for predicting ranks on an arousal scale. The ranking model was trained with the same Praat features that were used in the SVM-Praat method. The ground truth rank ordering of the seven emotions is given in Table 4.6. We refer to this method as 'SVM-Praat-rank'.

Evaluation and performance metrics As performance measures, we used relatively simple error metrics. For SVM-regression, an error metric, E_{regr}, defined as

$$E_i = \left| x_i^{pred} - x_i^{true} \right| \tag{4.4}$$

was used (which is also employed in Grimm et al. [69]): it measures the distance between the predicted and the ground truth value. We report the E_{regr} averaged over all test samples.

Similarly, the SVM-rank classifier can be evaluated by calculating the distance between the predicted and ground truth rank using Eq. 4.4, E_{rank}, but now x is a rank number (an integer).

Results We performed speaker-independent experiments with the SVM-Praat-regression and SVM-Praat-rank methods. In Fig. 4.8, the output of the SVM-Praat-regression model is plotted: although we can see that there is a trend of good Arousal estimation, there is still a lot of spread in each emotion category. To assess the SVM-Praat-regression model, a 'regression' error E_{regr} as defined in Eq. 4.4 was used. In order to make comparison possible, we also computed E_{regr} for the other systems that assume discrete emotion categories, such as SVM-Praat. In the case of systems that assume discrete emotions, the ground truth and predicted discrete labels are substituted with the corresponding Arousal values as provided in Table 4.6. In Table 4.7 the results are shown in terms of E_{regr}. An averaged E_{regr} of 0.17 means that for each prediction on

the Arousal scale, an averaged error of 0.17 is made on a scale of $[-1, 1]$. Based on the results shown in Table 4.7, we can conclude that the SVM-Praat-regression method performs slightly worse than the discrete systems SVM-Praat and the linearly fused system in terms of the E_{regr} metric.

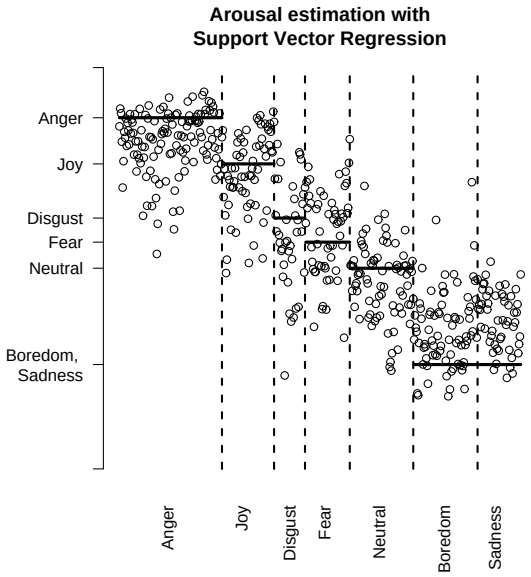


Figure 4.8: *The predicted Arousal scalar values (y-axis) of the SVM-Praat-regression system with the reference Arousal values displayed as horizontal lines.*

Approach	**System**	E_{regr}
Dimension	SVM-Praat-regression	0.17
Discrete	III. SVM-Praat	0.14
Discrete	IV. Linear fusion $\alpha = 0.5$, 128g	0.09

Table 4.7: *Averaged regression-errors E_{regr} of systems assuming emotion dimensions or discrete emotion categories.*

A similar experiment was performed for SVM-Praat-rank. Here, Arousal is predicted on an ordered scale that uses rank information rather than scalar output. In Table 4.8, a rank-error E_{rank} of 0.70 means that on the average, a prediction on the Arousal scale is made with a 0.70 shift in rank order on an Arousal scale with seven ranks. Similar to the SVM-Praat-regression method, we can observe in Table 4.8 that the SVM-Praat-rank method performs worse in terms of E_{rank} than the 'regular' systems that assume unordered, discrete emotion categories.

The results of both experiments indicate that the regression and ranking methods that assume scalar values and ordered categories produce errors that are slightly larger than methods that assume unordered, discrete categories. It suggests that, at least in this context, adopting emotion dimensions or ordered categories is not advantageous and does not have an additive value over the use of unordered categories. Further experiments with databases that are annotated on emotion dimensions should

Approach	**System**	E_{rank}
Ranking	SVM-Praat-rank	0.70
Discrete	III. SVM-Praat	0.66
Discrete	IV. Linear fusion $\alpha = 0.5$, 128g	0.39

Table 4.8: *Averaged rank-errors E_{rank} of different systems assuming ranked or unordered categories.*

be carried out to investigate in more detail how emotion prediction/estimation on emotion dimensions can be advantageous to discrete emotion recognition. In Chapter 6, we continue experimenting with Support Vector Regression and a spontaneous emotional speech database that is annotated on Arousal and Valence dimensions.

4.8 An 'open-set' detection evaluation methodology

As it is possible in automatic speech recognition to have 'out-of-vocabulary' words, it is also possible for an automatic emotion recognition system to encounter 'out-of-set' emotions. In other words, in real-life situations, the emotion recognition system can encounter an emotion that has never been 'heard' or 'learned' before by the system because it was not included in its training set. Especially in emotion recognition, it is not unlikely that this can occur, since it is difficult to obtain an abundance of emotion classes and emotion data. Most of the emotion databases available contain a relatively small number of rather 'arbitrarily chosen' emotion categories (e.g., 3–12) so it is not uncommon for an emotion detector that is trained on such small-sized databases to encounter 'new' emotions that were not included in the database. In current performance metrics and evaluation procedures, the possibility of encountering 'unheard' samples is disregarded, which implies that the performances reported do not always reflect the real 'application-readiness' of a system. For example, we do not know how an emotion recognition system that is trained to discriminate Anger from Joy reacts if it encounters Sadness. Therefore, in this Section, we address the notion of 'out-of-set' emotions and present an evaluation methodology in the detection framework based on a cross-validation scheme that simulates the existence of 'out-of-set' emotions. This performance evaluation is expected to produce performance figures that more closely reflect realistic situations, and hence, will increase the ecological validity of lab experiments. In van Leeuwen and Truong [195], a so-called 'open-set' detection evaluation methodology applied to language recognition was implemented to simulate the occurrence of 'surprise' languages. As the recognition task in language recognition is very similar to that of emotion recognition, the same 'open-set' detection evaluation methodology was applied here to emotion recognition[2].

Data In contrast to the dataset used in the previous experiments and described in section 4.3 and Table 4.2, the 'open-set' evaluation was performed on another subset

[2]The work described in this Section is based on our study that was previously published in Truong and van Leeuwen [188], van Leeuwen and Truong [195]

of the BERLIN database. This subset comprises of 535 utterances that fall under the same criteria of 80% recognition accuracy and 60% naturalness that we have applied previously. The difference is that this subset of the database has a number speech samples added to the filtered set of 494 samples in order to make the class distributions less skewed, see Table 4.9.

Emotion	N	Emotion	N
Anger (An)	127	Boredom (Bo)	81
Disgust (Di)	46	Fear (Fe)	69
Joy (Jo)	71	Sadness (Sa)	62
Neutral (Ne)	79		

Table 4.9: *Number of utterances used per emotion for the 'open-set' evaluation, total number of samples is 535.*

Method and features As speech features, we used Relative Spectral Transform - Perceptual Linear Prediction (RPLP) features [78, 3]. Each 16 ms, 12 RPLP coefficients plus 1 energy component were computed with a window width of 32 ms. In addition, delta coefficients were calculated by taking the first order derivatives of the 13 features over five consecutive frames. The features were normalized per utterance to obtain zero mean and unit standard deviation. We used GMMs (trained with five iterations of the Expectation-Maximization algorithm) as learning method to train the acoustic models. Four Gaussian components were used by a 'rule-of-thumb' (approximately 50 data points required per estimated parameter). These are the same GMMs as described in Fig. 4.2. In testing, a log likelihood ratio was obtained by subtracting the log likelihood given by the target GMM from the log likelihood given by the non-target GMM. This log likelihood ratio represents a degree of support for the target or non-target class. We trained 7 pairs of GMMs: for each target emotion one pair in a '1-vs-the-rest' set-up. Since our primary aim here was to implement the 'open-set' evaluation, we did not further optimize the performance of these classifiers.

'Open-set' detection evaluation methodology To emulate an 'open-set' situation, we implemented a cross-validation scheme that consists of several layers. In Fig. 4.9, the scheme is explained in pseudo-code. There are two layers: the outer layer ensures that the test is carried out speaker-independently by leaving out one speaker during training. In this layer, a target emotion model is trained with E_{TAR} samples that are *not* uttered by test speaker s, so this layer rotates over all speakers. In the inner layer, a non-target emotion model is trained on a subset of the rest of the emotional speech examples: excluded are the samples uttered by test speaker s in emotion TAR, *and* excluded are the samples uttered by all speakers in a certain non-target emotion e (i.e., this is the so-called 'surprise' emotion). The inner layer rotates over the set of non-target emotions (which always excludes the target emotion) so that each non-target emotion has served as a 'surprise' emotion for a specific target emotion detector. This 'double-layered' cross-validation scheme ensures that for each test performed, a target model is paired with a non-target model where both models do not have prior

information about the test samples that carry an 'unknown' emotion uttered by an 'unknown' speaker. This way, we obtained non-target scores for the 'new' emotion samples which can be pooled into a single non-target score distribution. The target scores cannot directly be pooled into a target score distribution; these need special attention. We obtain multiple, correlated target scores for the same target trial due to each iteration over the non-target emotions. We decided to take the minimum of these target scores obtained for the same target trial and discarded the rest to emulate the most pessimistic situation (an alternative would have been to average these scores). The performances were reported in EERs. In addition to the 'open-set' emotion detection experiments, single-loop, speaker-independent (SI) emotion detection experiments are performed too. In each fold of these 10-fold cross-validation experiments, the test speaker is left out of the training set so that no prior information about the test speaker is available during testing; this is a relatively straightforward evaluation procedure (also known as LOSO which we have applied before, 'leave-one-speaker-out').

$\{\text{speakers}\}$ = set of all speakers
$\{\text{emotions}\}$ = set of all emotions
$\{\text{nonemotions}\} = \{\text{emotions}\} \setminus \mathsf{E}_{\text{tar}}$
S = test speaker
E_{non} = the nontarget emotion
E_{tar} = the target emotion

foreach S of $\{\text{speakers}\}$
    train target model : $\neg\mathsf{S} \wedge \mathsf{E}_{\text{tar}}$
    foreach E_{non} of $\{\text{nonemotions}\}$
        train nontarget model : $\neg\mathsf{S} \wedge \neg\mathsf{E}_{\text{tar}} \wedge \neg\mathsf{E}_{\text{non}}$
        test : $\mathsf{S} \wedge (\mathsf{E}_{\text{tar}} \vee \mathsf{E}_{\text{non}})$
    end
end

and this can be repeated for each E_{tar}

Figure 4.9: *'Open-set' cross-validation scheme in pseudo code*

Results First, speaker-independent (single-loop) emotion detection experiments were performed. The performances of these SI emotion detection experiments can be seen in Fig. 4.10, where the DET curves of the target emotions are plotted.

We can observe in Fig. 4.10 that Sadness has the lowest EER, while Fear and Disgust have the highest EER. This is in concordance with the results, shown in Fig. 4.7, that we obtained previously with more advanced systems as described in section 4.6.

Subsequently, we applied the proposed 'open-set' detection evaluation methodology in which we test on a speech sample that is uttered by a 'surprise' speaker in a 'surprise' emotion, both of which have not been 'seen' before by the models. Table 4.10 presents the SI results in the left column and the 'open-set' results in the right column. Not surprisingly, the EERs have increased in the 'open-set' case: with

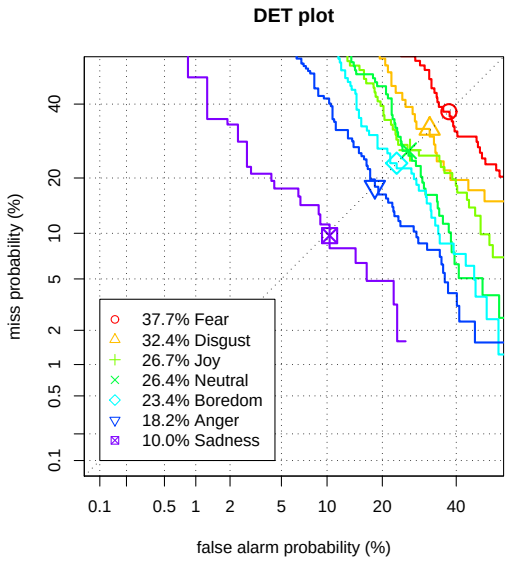


Figure 4.10: *Speaker-independent results (EERs) obtained with Berlin Emotional Speech database.*

an average EER of 37.5%, it appears that emotions are very difficult to detect if we do not have prior knowledge about the types of potential non-target emotions. As an exception, Sadness seems to be a very distinct emotion that is relatively easy to detect: even if there is no prior information about the non-target emotions, Sadness can be detected with an EER of 12.9% which is a small increase of +2.9% points in comparison with the SI experiment. Fear on the other hand, shows a very extreme EER in the 'open-set' case; it suggests that the alternative emotions are very poor representatives of the non-target emotion, and that Fear is not a very distinct emotion and can be confused easily with other emotions. This has also become clear in our 'acoustic map of emotions', see Section 4.9, in which emotions are located closer to each other when they are acoustically similar. For the visualization of confusions, we constructed an 'acoustic map of emotions' by performing pair-wise discrimination experiments. In this map, Fear lies 'in the middle', closely surrounded by other emotions which suggests poor discriminability, see Section 4.9.

To summarize, an 'open-set' detection evaluation methodology is proposed to evaluate emotion recognizers in a more 'realistic' way and produce performance figures that are less dependent of the emotions available in the database used. These 'open-set' experiments showed that Sadness (in this database) is a very distinct emotion that is relatively easy to detect.

4.9 Visualizing confusion in an acoustic map of emotions

In classification results, it is common to include a *confusion matrix* that provides information about the misclassifications made. Since we are following a single-emotion (i.e., detection of one target emotion) detection approach in a detection framework, this information is less visible in the results. As a way to compensate for this 'missing'

Emotion	Berlin		
	SI	'open-set'	Δ
Anger	18.2	25.2	+7.0
Boredom	23.4	36.0	+12.6
Disgust	32.4	46.1	+13.7
Fear	37.7	65.2	+27.5
Joy	26.7	35.5	+8.8
Neutral	26.4	41.7	+15.3
Sadness	10.0	12.9	+2.9
Mean	25.0	37.5	+12.5

Table 4.10: *EERs of speaker-independent and 'open-set' detection experiments on Berlin database*

information, we constructed an 'acoustic map of emotions'[3] and visualized potential confusions in this map by performing pair-wise discrimination experiments. The same set of data and method and features are used as described in the previous section, section 4.8. Since our primary aim here was to implement the 'acoustic map of emotions', we did not optimize the performance of the classifiers used.

For reference: multiclass classification For reference, we performed multiclass classification with the same basic models used in the 'open-set' case. GMMs were trained with 12 RPLP coefficients and 1 log energy component and four Gaussian components. Training and testing is performed through a leave-one-speaker-out cross-validation concept to ensure speaker-independency. Seven GMMs were trained, for each target emotion one GMM. During testing, the predicted emotion class was determined by maximum likelihood using the log-likelihood as a score. As a result, we obtained the following confusion matrix, see Table 4.11. According to this confusion matrix, Boredom and Neutral are often confused with each other, as well as Joy and Anger.

	Classified as						
	An	Bo	Di	Fe	Jo	Ne	Sa
An	63.8	3.1	7.9	7.1	13.4	4.7	0
Bo	1.2	45.7	3.7	6.2	1.2	35.8	6.2
Di	17.0	1.9	34.0	17.0	9.4	9.4	11.3
Fe	4.3	17.4	10.1	26.1	20.3	14.5	7.2
Jo	22.5	4.2	11.3	9.9	46.5	5.6	0
Ne	1.3	27.8	12.7	10.1	0	48.1	0
Sa	1.6	8.1	3.2	4.8	0	8.1	74.2

Table 4.11: *Confusion matrix obtained with multiclass classification (expressed in percentages of the 'true' class)*.

[3]The work described in this section is based on our study that was previously published in Truong and van Leeuwen [189]

From EERs to (acoustic) distances How can we assess what types of emotions are easily confused with each other in a detection evaluation framework? From the DET curves in Fig. 4.10, we can infer that Sadness can be detected relatively easily while Fear is very difficult to detect. But we do not know how the performances of these *single-emotion* detectors relate to each other mutually and what type of misclassifications were made. In that respect, a confusion matrix is a very helpful tool for assessing misclassifications. There are ways to go from single-emotion detection to multi-class emotion classification (see section 4.4.3) in order to obtain a confusion matrix, but a more elegant way to learn more about confusions in a detection concept would be to relate acoustic similarities between emotions to EERs. Luckily, one of the nice properties of the detection evaluation framework is that EERs can be related to (acoustic) distances. Subsequently, using Multidimensional Scaling, we can visualize these acoustic similarities in an acoustic map of emotions where geometric distance is related to acoustic similarity; the closer emotions lie to each other in that map, the more similar these emotions are acoustically.

The first step in obtaining this map is to perform $\binom{7}{2} = 21$ pair-wise emotion discrimination experiments, e.g., Anger vs. Boredom, Anger vs. Disgust, Anger vs. Fear etc. with the seven target GMMs. So in testing, only trials coming from the pair of emotions under discrimination were tested. The log-likelihood ratios obtained with the two GMMs were used as soft decision scores to determine EER. The EERs obtained with these pair-wise discrimination experiments represent the discrimination performance between the two emotions and can be interpreted as a similarity measure: the higher the EER, the more similar two emotions are. However, an EER does not have distance-like properties. Intuitively, a better distance representation of the EER is a quantity known from signal detection theory as d' ('d-prime'). Assuming equal variance of the target and non-target score distributions (the obtained log-likelihood ratios of the target and non-target trials form these distributions), d' is defined as the difference (distance) in mean between the distributions expressed in terms of the standard deviation, $d' = |\mu_{\text{tar}} - \mu_{\text{non}}|/\sigma$. Under the assumption of Gaussian score distributions DET curves are straight lines perpendicular to the equal-probability diagonal. The EER P_{EER} and d' are related through the inverse cumulative normal distribution, or probit function (see also van Leeuwen and Brümmer [196]):

$$d' = -2\,\text{probit}(P_{\text{EER}}) = -2\sqrt{2}\,\text{erf}^{-1}(2P_{\text{EER}} - 1) \tag{4.5}$$

This probit function, expressed here in terms of the inverse error function, is just the warping function of the DET axes, so that d' varies *linearly* in the DET plot from 0 in the upper-right corner to about 6 in the lower-left corner. So through Eq. 4.5 and pair-wise emotion discrimination experiments, we can calculate acoustic distances between the 21 pairs of emotions. Table 4.12 shows these distances d' computed on the basis of P_{EER}: the larger d', the less acoustic similar the two emotions are.

Although d' can be interpreted as a distance measure, the d's given in Table 4.12 do not satisfy the triangle inequality theorem that is valid in a Euclidean geometric space. Hence, we refrain from stating that d' is a pure distance metric. Note that the EERs are based on relatively small sets of target and non-target trials, 49–127 trials (see Table 4.9), which may have rendered the EERs, and consequently the d's, less

	An	Bo	Di	Fe	Jo	Ne	Sa
An							
Bo	3.07						
Di	2.07	2.22					
Fe	1.88	1.29	1.02				
Jo	0.77	2.43	1.58	1.35			
Ne	2.29	0.74	1.74	0.48	2.28		
Sa	4.29	1.87	2.44	2.12	4.33	3.05	

Table 4.12: *D-primes computed at EER.*

accurate.

Visualization via Multidimensional Scaling We can use the d's calculated to draw a 2-dimensional plot, our so-called 'acoustic map of emotions', via Multidimensional Scaling. Multidimensional Scaling (MDS, Venables and Ripley [198]) is a statistical technique that can visualize distance-like data in a low-dimensional geometric picture (for readability, the number of dimensions is usually 2 or 3). In non-metric MDS, the goal is to minimize the differences between the reproduced distances d_{ij} in the map and a monotonic transformation of the input distance data $f(d'_{ij})$ (see Eq. 4.6). The method uses the relative orderings of the given distances d'_{ij} (hence the 'non-metric') to construct the metric structure of the input data.

$$S^2 = \frac{\sum_{i \neq j}[f(d'_{ij}) - d_{ij}]^2}{\sum_{i \neq j} d_{ij}^2}, \tag{4.6}$$

where S is the so-called stress measure. This stress measure is an indication of how good the fit is of the MDS analysis on the data. The lower the stress, the better the fit. A scree plot of the MDS analysis applied to our data is given in Fig. 4.11.

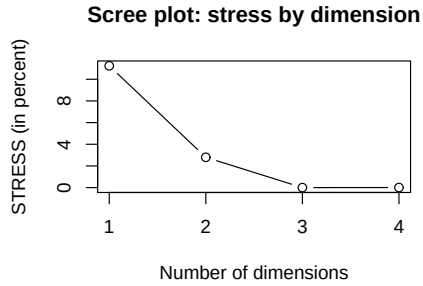


Figure 4.11: *Stress values plotted against number of dimensions of our non-metric MDS analysis.*

In Fig. 4.12, we can observe the 2-dimensional plot as a result of the non-metric MDS analysis applied to the distance data in Table 4.12. In this 'acoustic map of emotions', emotions that are difficult to discriminate from each other (i.e., high EER)

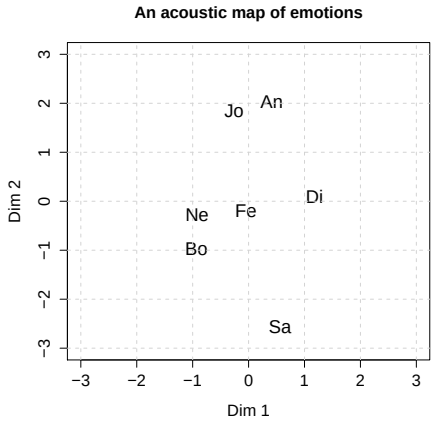


Figure 4.12: *Visualization of acoustic differences between emotions based on d' and scaled by MDS (number of dim=2), the arbitrary dimensions have been rotated to fit the Feeltrace representation for the Arousal dimension.*

lie next to each other while emotions that are easy to distinguish from each other (i.e., low EER) lie far away from each other. This plot in fact nicely summarizes both Fig. 4.10 and Table 4.12. We can now actually *see* what types of confusions between emotions are made and how the EERs are related to these confusions. It appears that Fear is difficult to detect (see Fig. 4.10) because it is surrounded by relatively close neighbors in Fig. 4.12: Neutral, Boredom and Disgust appear to be acoustically similar to Fear. Previously, we found that Sadness is a very distinct emotion (see Table 4.10, Fig. 4.10); this is also reflected in our map (Fig. 4.12) where Sadness does not have any close neighbors.

Note that Joy and Anger lie close to each other which implies that they are acoustically similar, although semantically, Joy and Anger are two opposites on the Valence dimension. Clearly, acoustic discrimination on the Valence dimension is still problematic as was shown in earlier studies, e.g., Banse and Scherer [12], Schröder et al. [169], Yildirim et al. [215]. Furthermore, in Truong and Raaijmakers [185], we found that in spontaneous speech, Valence information is better captured in the lexical than acoustic content. The acoustic map confirms that acoustic discrimination is easier on the Arousal dimension: Sadness and Boredom (low Arousal) are very distant from Anger and Joy (high Arousal).

4.10 Discussion and conclusions

The main goal of this Chapter was to present existing and accepted evaluation methodologies from similar recognition technologies and apply these to the field of emotion recognition where shared evaluation is still an underexposed topic. We hope to have raised increasing awareness on evaluation techniques specifically attuned to emotion recognition. Furthermore, we have presented a way to improve the ecological validity of lab classification experiments by implementing an 'open-set' detection evaluation methodology. Also, we have shown what the current performances of emotion de-

tectors are when clean and acted emotional speech and state-of-the-art recognition techniques are used. Finally, by relating EERs to 'acoustic distances', we were able to visualize acoustic similarities between emotions.

We developed three different emotion recognition systems and combined the two best performing systems using a decision-level fusion strategy. The best performing system was developed with a GMM supervector SVM based method and spectral frame-level RPLP features. The second best performing system was based on supra-segmental prosody-related features trained with an SVM. Combining these two systems on decision-level yielded the best results with C_{det} between 4% and 25%. Sadness was relatively easy to detect, while Fear, Joy and Disgust were much more difficult to detect. Clearly, combining different systems and different types of features on decision-level is advantageous; we suspect that the main contribution lies in the supra-segmental information that is added to the frame-level spectral based system since both systems share the similar SVM learning concept but differ in the types of features used.

These speech-based emotion detectors were evaluated in a detection evaluation framework which is very applicable to the field of emotion recognition, but not very often used in this field. In this framework, the emotion recognition problem can be much more approached as a 'true' emotion recognition task rather than an emotion *discrimination* task. In addition, the detection evaluation framework is widely accepted in fields like language recognition, that has a similar task definition as emotion recognition, and offers common (international) benchmark style evaluation protocols. A relatively young research field like emotion recognition, in which results and performances are reported fragmentally (mainly due to lack of standardization) can take profit of these existing tools to advance towards more standardization and a more structural (incremental) development of emotion recognition technology. In this framework, we have implemented a cross-validation scheme that emulates an 'open-set' situation and that addresses one specific aspect of the ecological validity of traditional lab emotion classification experiments. In these lab emotion classification experiments, a closed set of emotion classes is assumed which is not a realistic situation, especially in the case of emotion recognition since it is difficult to obtain an abundance of emotion classes and emotion data. In order to obtain performance figures and results that are more 'realistic' and less dependent on the number and types of emotion classes available in the database, we implemented a cross-validation procedure in which the detector is tested on 'surprise' non-target emotions uttered by 'surprise' speakers. Although the design is rather unpractical, it ensures that during testing, the detector has no prior information about the potential non-target emotions and speakers. Applying this procedure to the BERLIN database, we found (and confirmed) that Sadness is a very distinct emotion in this database; even when there is no prior information about the possible non-target emotions, Sadness can be detected with an EER of 12.9%.

In the traditional multiclass classification paradigm, the confusion matrix is obtained as a by-product of the evaluation which gives information about erroneous confusions made by the classifier. One disadvantage of the detection evaluation framework is that it does not directly provide insight into the characteristics of the

errors made by the detector. To compensate for this, we performed pair-wise emotion detection experiments between all pairs of emotions and related the EERs obtained to a similarity (distance) measure d'. Multidimensional Scaling was applied to these distances in order to obtain 'an acoustic map of emotions' (see Fig. 4.12). The closer the emotions are in this map, the more similar they are acoustically. In this map, Fear lies 'in the middle' and is closely surrounded by Neutral, Boredom, and Disgust. This suggests that Fear is difficult to detect (because it is acoustically similar to its close neighbors Neutral, Boredom, and Disgust) which is in accordance with our detection results shown in Table 4.10.

Finally, we have also compared a discrete emotion detection approach to a continuous dimension emotion recognition approach. In the latter approach, a dimensional model of emotion is adopted and regression techniques are used to estimate scalar values on scales of Arousal, Valence or Dominance. We applied a dimensional approach to the Berlin database and used SVM regression and ranking to estimate values on the Arousal scale. The labels of the discrete emotion categories in the BERLIN database were replaced with the Feeltrace landmarks. A relatively simple error metric, that measures the absolute difference between the predicted and reference value, was applied to both the discrete and continuous dimension approach. The error was higher for the continuous dimension approach than for the discrete emotion approach. These preliminary results suggest, that at least in this case, it is not advantageous to adopt a dimensional approach. Note that the database was not annotated in a dimensional way. In Chapter 6, we describe our experiments performed with regression techniques to estimate scalar values on Arousal and Valence scales. In these experiments, a spontaneous emotional speech database was used that is annotated specifically on Arousal and Valence dimensions.

The work presented in this Chapter is based on a database that is freely available for research which is of great value for evaluation purposes. However, one major disadvantage of this database is that it contains *acted* emotional speech. Obviously, it would be more sound and ecologically valid to use a database that contains real, naturalistic affective speech. Therefore, in the following Chapters, we present emotion detection experiments carried out with spontaneous affective speech data.

Chapter 5

Recognition of spontaneous affective behavior in meetings

To increase the ecological validity of our studies, we advance towards the use of real affective speech data and focus on the recognition of spontaneous affective conversational behavior. Nonverbal vocal sounds are often the most informative and communicative cues in conversational behavior (e.g., Campbell [29]). Cries, yawns, sighs, coughs, etc. are examples of **paralinguistic events** that implicitly convey affective information about the speaker's affective state. **Laughter** is another well-known example of a paralinguistic event (also called an 'affect burst' by Scherer [161], Schröder [167]). Since laughter is a (relatively) distinct affective event that occurs relatively frequently and that is often annotated in speech databases, we decided to investigate the automatic detection of laughter in the context of meetings.

In the past few years, several large meeting corpora have been recorded and enriched with different types of meta-information with the goal to support multidisciplinary research. These corpora form an important data source for natural language and speech processing research. Partly motivated by the wealth of natural meeting speech data available for research, and partly motivated by the demand for means to browse through these speech data, we decided to investigate, in addition to laughter detection, the recognition of **subjective content**, i.e., **sentiments** and **opinions**, in the context of meetings. The assumption is that when people express their opinions, they are more aroused and involved than when they express facts. Increased involvement may indicate so-called 'hot spots' in meetings, e.g., discussions with heated arguments, points of excitement, which could be interesting for browsing and summarization purposes.

In this Chapter, we describe detection experiments in which we compared several types of features and classifiers for the detection of a) laughter[1], and b) subjectivity[2] in the context of meetings. Note that laughter and subjectivity are both **implicit** carriers of affective information, which seems to be characteristic of naturalistic affective behavior.

[1]The work about laughter detection described in this Chapter is based on earlier work that was published in Truong and van Leeuwen [186, 190, 187]

[2]The work about subjectivity detection described in this Chapter is based on earlier work that was published in Raaijmakers, Truong and Wilson [143]

This Chapter is structured as follows. In section 5.1, we describe what types of affect can be found in meeting environments. Subsequently, we first describe the laughter detection experiments. Section 5.2.1 summarizes the work of other studies on laughter (detection). In Section 5.2.2, we define the laughter detection and segmentation tasks. Section 5.2.3 and Section 5.2.4 give descriptions of the material and method and features used to develop the laughter detector. The results of the laughter detection experiments are given in Section 5.2.5. The laughter segmentation experiment is described in Section 5.2.6. In Section 5.2.7, an interactive laughter application that was developed by TNO and their projectpartners, using the method developed here, is described. Finally, the laughter detection study is summarized in Section 5.2.8. The second focus of this Chapter is on subjectivity detection. Related work on subjectivity research is summarized in Section 5.3.1. The subjectivity detection task is defined in Section 5.3.2. The material, method and features used for subjectivity recognition are described in Section 5.3.3 and Section 5.3.4 respectively. The results and a summary of the subjectivity detection study are given in section 5.3.5 and Section 5.3.6. Finally, we conclude with a discussion and conclusions in Section 5.4.

5.1 What is happening in meetings?

For natural language and speech processing research, recorded meetings form a welcome source of data. Not only core technologies such as (far field) automatic speech recognition, speaker detection and speaker segmentation, speech activity detection etc. can use these data for development and evaluation, meeting data is nowadays also frequently used in studies analyzing higher-level meeting structures. Dialogue act analysis and detection (e.g., Shriberg et al. [176], Ang and Shriberg [7], Zimmermann et al. [220]), hot spot detection (Wrede and Shriberg [211]), disfluency detection (Baron et al. [13]), detection of agreement and disagreement (Galley et al. [65]), analysis of overlaps (Cetin and Shriberg [36]) and meeting summarization (Murray et al. [121, 122]) are some of the topics investigated in meeting data.

Attempts have also been made to model and recognize emotions in meetings. This has appeared to be difficult due to the naturalistic nature of the data. One of the first difficulties encountered is that of description and annotation of the speech data: how and what types of emotions should be annotated in the context of meetings? For example, Laskowski and Burger [103] proposed an emotion annotation scheme for the ISL Meeting Corpus (Burger et al. [24]) that describes more closely how humans *behave* rather than how they *feel*. Emotional valence was annotated in 3 classes (Negative, Neutral and Positive) and subsequently, classifiers were trained to detect emotional valence in meetings (see Neiberg et al. [124]). Reidsma et al. [146] and Heylen et al. [79] initially used Feeltrace (Cowie et al. [45]) for the annotation of emotions in the AMI Meeting corpus (Carletta [35]). However, the inter-rater agreement obtained with this method was low. One of the reasons why this did not work as well as expected, was that the annotators felt that most of the changes in the mental states of the participants could not be described in terms of Arousal and Valence dimensions. Therefore, another annotation procedure was proposed in which anno-

tators were asked to segment and label when there was a clear change in the mental state type and a clear change in intensity of the mental state. The labels describe a 'mental state' which is defined by 'a feeling' in a broad sense, including not only the typical emotion categories such as 'irritated' or 'amused' but also so-called meta-cognitive states and processes such as 'paying attention' or 'interest' (Reidsma et al. [146], Heylen et al. [79]). So-called 'hot spots' in meetings can also be considered 'affective' events: a 'hot spot' is defined in Wrede and Shriberg [211] as a 'region in a meeting where there is high involvement on the part of two or more participants', e.g., discussions with heated arguments, points of excitement etc. Finally, the topics under investigation, laughter and subjectivity, are also frequently encountered phenomena in meetings which will be discussed in subsequent sections.

It is clear from the studies mentioned above, that the type of 'emotions' we can encounter in meetings are not so much the typical 'basic' ones such as Anger or Sadness, but rather are different types of mental and cognitive states that fall under the broader definition of 'emotion' and that are more related to the way people behave and interact in multiparty conversation.

5.2 Automatic detection of laughter in meetings

In this Section, we focus on the development of automatic laughter detectors for meeting speech data. The detection task is kept clear and simple, and hence, we do not take contextual factors into account, and we do not interpret the laughter, i.e., we do not attach a *meaning* to the laughter. The task is purely based on the acoustic characteristics of laughter. We defined two tasks, see Section 5.2.2: 1) laughter vs. speech discrimination, and 2) laughter segmentation. As speech material, we used the ICSI Meeting corpus (Section 5.2.3). Classifiers developed with several types of features were trained and tested in the detection framework, see Section 5.2.4. The results of these detection experiments are presented in Section 5.2.5. Furthermore, we show an example application of real-time laughter detection in Section 5.2.7. But first, in Section 5.2.1, we give a summary of related work on the acoustics of laughter and automatic laughter detection.

5.2.1 Related work

Earlier studies on laughter involved the acoustics of laughter which were compared to that of speech, e.g., Mowrer et al. [120], Bachorowski et al. [11], Trouvain [184], Bickley and Hunnicutt [21], Rothganger et al. [152], Nwokah et al. [126], Campbell et al. [30] . One of the largest studies on the acoustics of laughter is the one described in Bachorowski et al. [11]. In Bachorowski et al. [11], 1024 naturally produced laugh bouts from 97 young adults were recorded as they watched funny video clips. They noted that cues of individual identity can be conveyed in laughter acoustics and that laughter can thus be an aid to automatic speaker recognition (Knox and Mirghafori [94], Knox et al. [95]). The most important conclusion from the study by Bachorowski et al. [11] is that laughter is a highly complex and variable vocal signal, rather than a stereotyped vocal signal. The high complexity and variability of laughter is also reflected in the mixed results that were obtained by several studies on

the acoustics of laughter. For example, Bachorowski et al. [11] and Rothganger et al. [152] both reported a higher F_0 for laughter than (modal) speech, and they concluded that speech is rather monotonic, lacking a strongly varying melodic contour that is present in laughter. On the contrary, Bickley and Hunnicutt [21] reported that mean F_0 and amplitude measures of laughter are rather speech-like. However, Bickley and Hunnicutt [21] also reported an important difference between laughter and speech in the durations of the voiced portions: a typical laugh reveals an alternating voice-unvoiced pattern in which the ratio of the durations of unvoiced to voiced portions is greater for laughter than for speech.

The complexity and variability of laughter also reveals itself in contextual and cultural dependency. For example, Smoski and Bachorowski [179] found that laughter varies with social factors such as the gender of and familiarity with one's social partner. Although Rothganger et al. [152] did not find any significant differences in the acoustics of laughter sounds from Italian and German students, other studies did find differences in the expressive behavior of laughter. For instance, Ekman [56] found in his experiment that the Japanese subjects, more than the American subjects, masked their negative expressions with smile.

In Schröder [167], the perceived emotional content of so-called 'affect bursts', including laughter, was investigated. Affect bursts are defined as 'very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events'. It was concluded that these affect bursts can convey a clearly identifiable emotional meaning, although laughter could not be related to a prototypical emotion.

The main view that arises from these studies, is that laughter is a very variable and complex acoustic signal that carries affective information and that is affected by factors like context, culture, and personality. Due to its complexity, in most automatic laughter detection studies, these factors are not taken into account. In the majority of laughter detection studies, the aim is to detect laughter without having to specify or to interpret the laughter, e.g., Kennedy and Ellis [90], Laskowski and Schultz [104], Cai et al. [27], Lockerd and Mueller [110], Campbell et al. [30], Laskowski and Schultz [104], Reuderink et al. [147], Knox and Mirghafori [94], Knox et al. [95], Petridis and Pantic [130, 131, 132], Truong and van Leeuwen [186, 190, 187], see Table 5.1.

In Table 5.1, where short summaries of laughter detection studies are given, it can be observed that laughter detection has often been investigated in the context of meetings. The study by Kennedy and Ellis [90] was one of the first that investigated laughter detection in meetings: an SVM was trained with MFCCs, their deltas, spatial cues and modulation spectra coefficients, and a correct accept rate and false alarm rate of 87% and 13% respectively were achieved. Knox and Mirghafori [94], Knox et al. [95] used neural networks and HMMs in combination with MFCCs and prosodic features for laughter segmentation. Laskowski and Schultz [104] used a multiparticipant 3-state vocal activity recognition module to detect so-called *laughter-in-interaction*. Recently, audiovisual approaches to laughter detection have been undertaken by Reuderink et al. [147] and Petridis and Pantic [130, 131, 132]: according to their work, fusion between the visual and auditory modalities helps, but it remains unclear how this fusion between visual and auditory information should work.

Note that in automatic speech recognition, laughter is considered non-speech,

Study	Focus	Data	Method & Features	Performance
Cai et al. [27], 2003	laughter segmentation	102 hours tv shows (laughter, applause, cheer)	HMM & short-term energy, zero- crossing rate, sub-band energies, MFCCs	Recall: 93%, Precision: 87%
Lockerd and Mueller [110], 2002	segmentation and laughter vs. speech discrimination	40 laughter and 210 speech segments, single user	HMM & spectral coefficients	Correct: 88%
Kennedy and Ellis [90], 2004	localization of simultaneous laughter (tabletop recordings)	ICSI Meeting Corpus	SVM & MFCCs+deltas, spatial cues, modulation spectrum	Correct Accept rate: 87%, False Alarm rate: 13%
Campbell et al. [30], 2005	classification of different types of laughter	Japanese ESP corpus (Campbell [28]), 3000 laughs	HMM & MFCCs (?)	Correct: 75%
Truong and van Leeuwen [186], 2005	laughter vs. speech discrimination	ICSI Meeting corpus, CGN (Dutch)	GMM & PLP, pitch, energy, voicing features, modulation spectrum	EER: 7.1–15.6%
Knox and Mirghafori [94], 2007	laughter segmentation	ICSI Meeting corpus	Neural networks & MFCCs pitch, energy	EER: appr. 8%
Truong and van Leeuwen [190], 2007	laughter segmentation	ICSI Meeting corpus	GMM, Viterbi, LDA & PLP, prosodic features	EER: appr. 10%
Truong and van Leeuwen [187], 2007	laughter vs. speech discrimination	ICSI Meeting corpus, CGN (Dutch)	GMM, SVM, MLP, fusion & PLP, pitch, energy, voicing features, modulation spectrum	EER: appr. 3%
Knox et al. [95], 2008	laughter segmentation	ICSI Meeting corpus	hybrid MLP/HMM, Viterbi & MFCC, pitch, energy, prosodics, modulation filtered spectrogram	Precision: 79%, Recall: 85%
Laskowski and Schultz [104], 2008	laughter segmentation	ICSI Meeting corpus	HMM & MFCCs	Precision: 25%, recall: 55%
Reuderink et al. [147], 2008	audiovisual laughter vs. non-laughter discrimination	AMI corpus	GMM, HMM, SVM, fusion & Video: 20 2-d facial points, tracking, PCA, Audio: RLP	EER: 14%
Petridis and Pantic [130], 2008	audiovisual laughter vs. speech discrimination	AMI corpus	AdaBoost, neural networks, fusion & Video: 20 2-d facial points, tracking, PCA, Audio: PLP	Precision: 77%, recall: 87%

Table 5.1: *Overview of (audiovisual) laughter detection studies.*

and hence, it is considered not interesting. Possibly, because of the main interests of the speech research community for the automatic recognition of speech, rather than non-verbal speech elements, laughter detection has been a somewhat underexposed area. The work described here (Truong and van Leeuwen [186, 187, 190]) were among the first laughter detection studies that extensively compared several recognition techniques and acoustic features, and that acknowledged the importance of laughter detection with respect to automated higher-level understanding of meetings and automatic affect recognition.

5.2.2 Defining the discrimination and segmentation tasks

As noted earlier, laughter is a nonverbal vocalization with various social, communicative functions (e.g., Campbell et al. [30], Bachorowski and Owren [10]). As an effective means to express emotion (Schröder [167]), laughter is often associated with pleasant feelings and zygomatic activity (i.e., smiling, Russell et al. [155]). However, the relation between laughter and emotion is too complex to state that laughter is *always* associated with happiness. In fact, happiness is neither necessary nor sufficient for smiling. Laughs can also be produced out of anger and anxiety feelings (see Darwin [47]). Moreover, context plays an important role in producing different types of laughs; laughter varies with social factors such as the gender of and familiarity with one's social partner (Smoski and Bachorowski [179]). Although we are well aware of the fact that laughter can have several meanings and interpretations, the current tasks in this laughter detection study do not involve any interpretation of the laughter. Hence, the laughter detection tasks are defined as follows:

Task I. Discrimination between pre-segmented laughter and speech segments

The first task for the detector is to discriminate between laughter and speech, i.e., to classify a given pre-segmented acoustic signal as either laughter or speech. We decided to keep the discrimination problem clear and simple. Firstly, we used *pre-segmented* laughter and speech segments whose segment boundaries are determined by human transcribers. Automatically specifying the onset (i.e., the beginning) and offset (i.e., the ending) of a laughter event is thus not part of this task. Secondly, we only use (homogeneous) signals containing solely audible laughter or solely speech; signals in which laughter co-occurs with speech are not used. Consequently, 'smiled speech' is not investigated in this study. And thirdly, we use close-talk recordings from head-mounted microphones rather than far-field recordings from desktop/table top microphones.

Task II. Laughter segmentation

The second task for the classifier is to localize (i.e., to segment) laughter in a given acoustic signal. Thus, in contrast with Task I., the detector also has to position the start and end time of a laughter event.

5.2.3 Laughter and speech material: ICSI *Meeting Corpus and* CGN *corpus*

As speech and laughter material, we used the ICSI Meeting Recorder Corpus (Janin et al. [85]) and the Dutch CGN corpus (Oostdijk [127]). In this Section, we describe the amount of data used in our laughter detection experiments.

ICSI *Meeting Recorder corpus*

We used the ICSI Meeting Recorder Corpus (Janin et al. [85]) that includes manually transcribed annotations of spontaneous laughter, see Section 2.4.1 for a brief description of the database. There are simultaneous recordings available of up to 10 close-talking microphones of varying types and four high quality desktop microphones. We only used the close-talk recordings in our detection experiments. The data was divided in training and test sets: the first 26 ICSI 'Bmr' ('Bmr' stands for the type of meeting, in this case 'Berkeley's Meeting Recorder' weekly meeting) subset recordings were used for training and the last three ICSI 'Bmr' subsect recordings were used for testing. The 'Bmr' training and test sets contain speech from sixteen (fourteen male and two female) and ten (eight male and two female) speakers respectively. Because the three ICSI 'Bmr' test sets contained speech from speakers who were also present in the 26 ICSI 'Bmr' training sets, another test set was added to perform speaker-independent detection experiments. Four ICSI 'Bed' ('Berkeley's Even Deeper Understanding' weekly meeting) sets with eight (six male and two female) unique speakers that were not present in the 'Bmr' training were selected as speaker-independent test material. Laughter segments were in the first place determined from laughter annotations in the human-made transcriptions of the corpus. The laughter annotations were not carried out in fine detail, it is comparable to word-level annotation. After closer examination of some of these annotated laughter segments in the corpus, it appeared that not all of them were suitable for our classification experiments: for instance, some of the annotated laughs co-occurred with speech and sometimes the laugh was not even audible. Therefore, we decided to listen to all annotated laughter segments and made a quick and rough selection of laughter segments that do not contain speech or inaudible laughter. Furthermore, although we are aware of the fact that there exist different types of laughter (see e.g., Campbell et al. [30]), e.g., voiced, unvoiced, 'snort-like', we decided not to make distinctions between these types of laughter. Speech segments were also determined from the transcriptions: segments that contained only vocalized sounds (excluding laughter) were labeled as speech. In total, from the ICSI corpus, we used 3264 speech segments with a total duration of 110 minutes (mean duration $\mu = 2.20$ s and standard deviation $\sigma = 1.87$ s) and 3574 laughter segments with a total duration of 108 minutes (mean duration $\mu = 1.80$ s and standard deviation $\sigma = 1.25$ s), see Table 5.2.

Spoken Dutch Corpus (Corpus Gesproken Nederlands CGN*)*

In addition to the ICSI meeting recorder corpus, the Spoken Dutch Corpus (Oostdijk [127]), Corpus Gesproken Nederlands, CGN) was used as a language and speaker-independent test set, see Section 2.4.1 for a brief description. The Spoken Dutch Corpus contains speech recorded in the Netherlands and Flanders and comprises a variety

	Training	**Test**		
	26 ICSI 'Bmr' meetings	**3 ICSI 'Bmr' meetings**	**4 ICSI 'Bed' meetings**	**14 CGN conversations**
	dur/N	dur/N	dur/N	$dur/\ N$
Speech segments	81 min/2422	10 min/300	15 min/378	4 min/164
Selected laughter segments	83 min/2680	10 min/279	11 min/444	4 min/171

Table 5.2: *Amount (duration in minutes) of laughter and speech data used in our laughter detection research.*

of speech types such as spontaneous conversations, interviews, broadcast recordings, lectures and read speech. We used speech data from the spontaneous conversations ('face-to-face') recordings and selected laughter segments that were annotated as non-speech sounds. Note that the CGN recordings originate from table top microphones, while from the ICSI corpus, we used the close-talk recordings. The amount of laughter and speech data used from the CGN corpus is displayed in Table 5.2.

5.2.4 Method and Features

In the following Section, we discuss several types of features that were investigated in combination with several types of modeling techniques. The features are divided into *frame-level* and *utterance-level* features. Frame-level features consisted of PLP, pitch, and energy features. Utterance-level features consisted of pitch and voicing features, and modulation spectrum features. The main modeling techniques used were GMM and SVM. For decision-level fusion, a multi-layer perceptron and a sum rule fusion were used.

Features

Frame-level spectral features (PLP) Spectral or cepstral features, such as MFCC and PLP (see Section 2.4.2 for a brief description), are often successfully used in speech and speaker recognition to represent the speech signal. We chose PLP features (mainly for practical reasons, but MFCCs would also have been good candidates) to model the spectral properties of laughter and speech. Each 16 ms, twelve PLP coefficients and one energy feature were computed over an analysis window of 32 ms. In addition, delta features were computed by calculating the deltas of the PLP coefficients (by linear regression over five consecutive frames) which resulted in a total of 26 PLP features. Furthermore, the features were normalized by performing z-normalization per utterance. This means that after normalization, for each utterance, the features have a mean of 0 and a standard deviation of 1 ($\hat{x}_{\text{frame}} = (x_{\text{frame}} - \mu_{\text{utterance}})/\sigma_{\text{utterance}}$).

Frame-level Pitch & Energy features (P&E) Several studies, e.g., Williams and Stevens [207], Banse and Scherer [12] have shown that with increased Arousal, for instance, when one is laughing, the speech measurements show an increased F_0 variability or range, with more source energy and friction accompanying increased intensity of effort. Furthermore, Bachorowski et al. [11] found that the mean pitch in both male and female laughter was considerably higher than in modal speech. Therefore, pitch and energy features would be good candidates for laughter vs. speech discrimination. Hence, each 10 ms, pitch and Root-Mean-Square (RMS) energy were measured over a window of 40 ms using Praat (Boersma and Weenink [23]). In Praat, we set the pitch floor and ceiling in the pitch algorithm at 75 Hz and 2000 Hz respectively. Note that we changed the default value of the pitch ceiling of 600 Hz, which is appropriate for speech analysis, to 2000 Hz since studies have reported pitch measurements of over 1000 Hz in laughter. If Praat could not measure pitch for a particular frame (for example if the frame is unvoiced), we set the pitch value at zero to ensure parallel pitch feature streams and energy feature streams. The deltas of pitch and energy were calculated and z-normalization was applied as well which resulted in a total of four P&E features.

Utterance-level Pitch & Voicing features (P&V) In addition to pitch measurements per frame, we also measured some more global, higher-level pitch features to capture better the fluctuations and variability of pitch in the course of time: we employed the mean and standard deviation of pitch, pitch excursion (maximum pitch–minimum pitch) and the mean absolute slope of pitch (the averaged local variability in pitch) since they all carry (implicit) information on the behavior of pitch over a period of time. Furthermore, Bickley and Hunnicutt [21] found that the ratio of unvoiced to voiced frames is greater in laughter than in speech and suggest this as a method to separate laughter from speech: "*. . . A possible method for separating laughter from speech, a laugh detector, could be a scan for the ratio of unvoiced to voiced durations . . .*". Therefore, we also calculated the fraction of locally unvoiced frames (number of unvoiced frames divided by the number of total frames) and the degree of voice breaks (the total duration of the breaks between the voiced parts of the signal divided by the total duration of the analyzed part of the signal). A total of six global P&V features per utterance were calculated using Praat (Boersma and Weenink [23]). The features were normalized by an 'utterance-based' z-normalization. In contrast with the z-normalization applied on the frame-level features where μ and σ were calculated over one utterance, μ and σ are now calculated over the whole training set (since we have one fixed-length feature vector per utterance): $(\hat{x}_{\text{utterance}} = (x_{\text{utterance}} - \mu_{\text{training}})/\sigma_{\text{training}})$

Utterance-level Modulation spectrum features (ModSpec) We tried to capture the rhythm and the repetitive syllable sounds of laughter, which may differ from speech: Bickley and Hunnicutt [21] and Bachorowski et al. [11] report syllable rates of 4.7 syllables/s and 4.37 syllables/s respectively in laughter, while in normal speech, the modulation spectrum exhibits a peak at around 3–4,Hz, reflecting the average syllable rate in speech (Drullman et al. [55]). Thus, according to these studies, it seems

that the rate of syllable production is higher in laughter than in conversational speech. Modulation spectrum features for laughter detection were also previously investigated by Kennedy and Ellis [90] who found that the modulation spectrum features they used did not provide much discriminative power. We calculated our modulation spectra of speech and laughter by first obtaining the amplitude envelope via a Hilbert transformation. The envelope was further low-pass filtered and downsampled. The power spectrum of the envelope was then calculated and the first 16 spectral coefficients (modulation spectrum range up to 25.6 Hz) were used as input features. The 16 ModSpec features were also normalized by an 'utterance-based' z-normalization.

Modeling techniques

Gaussian Mixture Modeling (GMM) We trained 'laughter' GMMs and 'speech' GMMs with the four different sets of frame-level and utterance-level features. The GMMs were trained using five iterations of the Expectation Maximization (EM) algorithm and with varying numbers of Gaussian mixtures (varying from 2 to 1024 Gaussian mixtures for different feature sets) depending on the number of extracted features. In testing, a maximum likelihood criterion was used. Similar to the 'Standard GMM' method summarized in Fig. 4.2, a 'soft detector' score is obtained by determining the likelihood ratio of the data given the 'laughter' and 'speech' GMMs respectively. For a brief description on Gaussian Mixture Modeling, the reader is referred to section 2.4.3.

In addition, as is frequently done in speaker recognition, we trained a Universal Background Model (UBM, see Reynolds et al. [149]) that represents all alternative classes (or speakers in the case of speaker recognition). We pooled together all laughter and speech data to train a UBM and derived a laughter model by adapting the parameters of the UBM to laughter. The tighter coupling between the laughter model and the background model could possibly improve performance.

Support Vector Machine (SVM) Support Vector Machines (Vapnik [197]) have become popular among many different types of classification problems, for instance face identification, bioinformatics and speaker recognition (for a brief description on SVMs see section 2.4.3). For the laughter detection experiments, we used the toolkit SVMTorch II, developed by the IDIAP Research Institute [42]. Three different kernels were employed: a linear, Gaussian, and a Generalized Linear Discriminant Sequence (GLDS) kernel (Campbell [31]). The latter kernel was used to transform the variable-length frame-level feature vectors (PLP and P&E) to static-length feature vectors. The GLDS kernel expanded the feature vectors explicitly into a higher-dimensional feature space. Subsequently, these expanded feature vectors were trained in SVM using a linear or Gaussian kernel.

Multi Layer Perceptron (MLP) For fusion of the classifiers, a Multi Layer Perceptron (MLP) was used (which has successfully been applied before, e.g., El Hannani and Petrovska-Delcretaz [60], Campbell et al. [32]). This popular type of feedforward neural network consists of an input layer (the input features), possibly several hidden layers of neurons and an output layer. The neurons calculate the weighted

sum of their input and compare it to a threshold to decide if they should "fire". As MLP implementation, the LNKnet Pattern Classification software package was used, developed at MIT Lincoln Laboratory (Lippmann et al. [107]). Note that the MLP was used as a fuser; the scores obtained with the individual GMM and SVM classifiers were used as input for the MLP.

Fusion techniques

We applied two different fusion techniques on *decision-level*. This means that the output (e.g., log-likelihood ratios, posterior probabilities) of several classifiers were combined to obtain a final decision score. The first fusion method applied is the relatively simple 'sum-rule':

$$S_{\text{fuse}} = \alpha S_{\text{A}} + (1 - \alpha) S_{\text{B}} \tag{5.1}$$

where α is a weight that can be optimized on a development set. In our case, we did not optimize α, instead we set a fixed $\alpha = 0.5$ so that each classifier, A and B, is equally important. Prior to this fusion, we normalized the scores by applying an adjusted form of T(est)- normalization (Auckenthaler et al. [9], Campbell et al. [32]). This was done by using a fixed set of non-target scores as a basis (in our case the 'Bmr' test set) from which μ and σ were calculated; these were used to normalize the target and non-target scores of the other two test sets ('Bed' and CGN) by subtracting μ from the score and dividing by σ.

Secondly, as an alternative fusion method, a second-level classifier was applied to 'learn' the fusion between the scores. As fusers, SVM and MLP were used. The training of these fusers was performed on the scores obtained with the 'Bmr' test set. Fig. 5.1 gives an overview of the fusion combinations that were carried out.

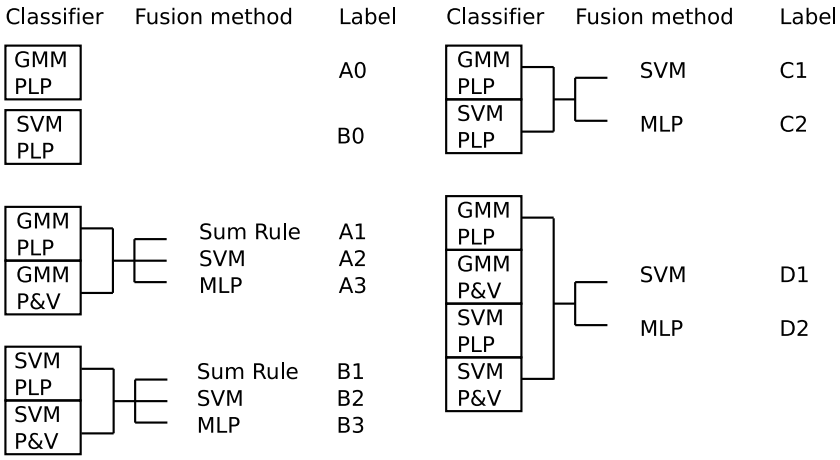


Figure 5.1: *Combinations of fusions of feature sets and classifiers for laughter vs. speech discrimination.*

5.2.5 Evaluation and Results

In this Section, the evaluation metrics used for the laughter detection experiments are described and the results are presented. First, the results of the *individual* detectors are presented. Subsequently, we present the results of the *fused* detectors.

Evaluation metrics

As evaluation metrics, we report EER (false alarm rate is equal to miss rate) and C_{det}. For the calculation of C_{det} we assume equal prior probabilities ($= 0.5$) and equal costs ($= 1$). Furthermore, the decision threshold for the calculation of C_{det} is determined at EER. Brief descriptions of EER and C_{det} are given in section 2.4.4.

Results of individual classifiers

To recapitulate, we have extracted four different sets of features PLP, P&E, P&V, and ModSpec, and we have employed two classification techniques namely GMM and SVM. For the GMMs, we varied the number of Gaussians from 2–1024. The results obtained with the GMMs are presented in Table 5.3 where the numbers of mixtures that produced the lowest EERs for that feature set are reported.

	Frame-level features		**Utterance-level features**	
	GMM PLP	**GMM P&E**	**GMM P&V**	**GMM ModSpec**
	1024 Gauss.	64 Gauss.	4 Gauss.	2 Gauss.
Bmr	6.4	14.3	20.0	37.7
Bed	6.3	20.4	20.9	38.7
CGN	17.6	32.2	28.1	44.5
Mean	10.1	22.3	23.0	40.3

Table 5.3: *Equal Error Rates (in %) of GMM classifiers trained with frame-level or utterance-level features and with different numbers of Gaussians*

The results in Table 5.3 show that GMM PLP outperforms all other GMM classifiers, whereas GMM ModSpec performs worst. As expected, the EERs increase as the degree of mismatch between training and test set increases. We also trained a Universal Background Model, as is often done in speaker-recognition, and build adapted GMMs but the performance did not improve (and hence were not reported) which was probably due to the small number of non-target classes: our UBM was trained with only twice as much data compared to the class-specific GMMs.

The results obtained with the SVMs are presented in Table 5.4. For the PLP and P&E features, a GLDS kernel was used to expand and transform the variable-length feature vectors into static-length feature vectors. Subsequently, the SVMs were trained with a linear or Gaussian kernel. Table 5.4 reports results of SVMs trained with a Gaussian kernel since that appeared to perform best.

We can observe in Table 5.4 that the SVM GLDS PLP outperforms the other SVM classifiers. Note that the second-best performing feature set for SVM is the P&V feature set. Taking into consideration the number of features, 26 PLP features per frame

	Frame-level features		Utterance-level features	
	SVM GLDS PLP	**SVM GLDS P&E**	**SVM P&V**	**SVM Mod-Spec**
Bmr	2.6	14.0	9.0	28.7
Bed	7.2	18.0	11.4	32.9
CGN	19.4	29.3	23.3	29.3
Mean	9.7	20.4	14.6	30.3

Table 5.4: *Equal Error Rates (in %) of SVM (Gaussian kernel) classifiers trained with frame-level or utterance-level features.*

as opposed to 6 P&V features per utterance, it is quite impressive for the SVM P&V method to achieve these results.

On the overall, when we compare the GMM and SVM methods to each other, it appears that SVMs perform slightly better. This can probably be mainly attributed to the GLDS kernel and the fact that our utterance-level features work much better in combination with SVMs than GMMs.

Results of fused classifiers

Since PLP and P&V are the two best performing feature sets (according to the results presented in Table 5.3 and 5.4), we performed fusion with these two feature sets and discarded the P&E and ModSpec features. Fusions were carried out on decision-level using a) the sum rule, or b) a SVM or MLP classifier. We first performed fusions *within* classifier-type and *between* feature sets, see Fig. 5.1, A1–A3 and B1–B3. The results are presented in Table 5.5 (for the SVM and MLP fusion methods we do not have results for the Bmr test set since this set was used for training of the SVM and MLP). The A0 and B0 classifiers represent the best performing individual (baseline) classifiers GMM PLP and SVM GLDS PLP respectively. The significance of improvement was assessed through a McNemar test (Gillick and Cox [66]) with a significance level of 0.05. Table 5.5 shows that the addition of a P&V classifier to a PLP classifier in many cases significantly increases performance, especially in the case of the SVM classifiers. The method of fusion does not seem to have much influence, although the linear fusion 'sum-rule' method seems to perform slightly worse (but note that the weight α was not optimized).

Subsequently, we performed 'cross' fusions *between* classifier-type and *between* feature sets, see Fig.5.1 fusions C1–C2 and D1–D2. Since the 'sum-rule' performed worse than the second-level classifiers, we only tested a second-level SVM and MLP as fusers. The results are presented in Table 5.6 in which we can observe that these 'cross' fusions are very powerful, achieving the best performances for the 'Bed' and CGN test sets.

Instead of placing thresholds 'virtually' a posteriori to report EER, we also placed thresholds a priori to know the *actual* performance of a classifier. In this case, the threshold was determined by choosing the score threshold where the probabilities of error are equal (EER); this threshold was then used to classify new samples resulting

Classifier		Features	Fusion method	EERs (%) Bmr	Bed	CGN	Compare to
A0	GMM×	**PLP**	**none**	**6.4**	**6.3**	**17.6**	-
A1		(PLP+P&V)	sum rule	8.6	11.7*	22.7	A0
A2		(PLP+P&V)	SVM	-	5.8	13.4*	A0
A3		(PLP+P&V)	MLP	-	6.1	12.8*	A0
B0	SVM ×	**PLP**	**none**	**2.6**	**7.2**	**19.4**	-
B1		(PLP+P&V)	sum rule	2.6	5.6	12.2*	B0
B2		(PLP+P&V)	SVM	-	5.2*	12.2*	B0
B3		(PLP+P&V)	MLP	-	4.7*	11.6*	B0

Table 5.5: *EERs of fused classifiers* within *classifier-type and* between *feature-sets (∗ indicates whether the difference in performance is significant with respect to the single classifier, A0, B0, displayed in the last column, × and + indicate fusion combinations).*

Classifiers		Features	Fusion method	EERs (%) Bmr	Bed	CGN	Compare to
C1	(GMM+SVM) ×	PLP	SVM	-	3.2*	11.6*	A0, B0
C2			MLP	-	3.7*	11.0*	A0, B0
D1	(GMM+SVM) ×	(PLP+P&V)	SVM	-	3.2	8.7*	C1
D2			MLP	-	2.9	7.5*	C2

Table 5.6: *EERs of fused classifiers* between *classifier-type* and between *feature-sets (∗ indicates whether the difference in performance is significant with respect to another classifier displayed in the last column, × and + indicate fusion combinations).*

in an evaluation of the actual performance of the system. This threshold was calibrated using the scores of the 'Bmr' test set. This was done for the best performing classifier D2, and as we can observe in Table 5.7, the *actual* performances are worse than the EERs reported, especially in the case of the CGN test set. It shows the difficulty of determining a threshold based on one data set and subsequently applying this threshold to another data set (threshold calibration). In addition, the unequal error rates, especially in the case of the CGN test set, are also an indication of mistuned thresholds.

Classifier	Test set	EER	Minimum C_{det}	Actual C_{det}	Actual Miss rate	Actual False Alarm rate
Fusion D2 (Table 5.6)	Bed	2.9%	0.028	0.045	1.8%	7.1%
	CGN	7.5%	0.075	0.173	31.6%	3.0%

Table 5.7: *Actual decision performances of fused classifier obtained by Fusion D2, see Table 5.6.*

5.2.6 Laughter segmentation

The second task concerned laughter *segmentation* rather than *laughter vs. speech discrimination*. Knox and Mirghafori [94], Knox et al. [95] performed laughter segmentation on the ICSI corpus using neural networks and a combination of spectral and prosodic features. They achieved relatively low frame-based EERs. However, note that they only used annotated vocalized segments; silence and other sounds were thus discarded. Laskowski and Schultz [104] segmented laughter based on three distinct states, namely laughter, speech and non-vocalizations. Their system is not only based on the acoustic characteristics of laughter, but also makes use of the vocal activity of multiple participants by constraining the number of simultaneous speakers and the number of simultaneous laughter.

For this laughter segmentation task, one could also use an automatic speech recognizer that can segment laughter as a by-product. However, since the aim of an automatic speech recognizer is to recognize speech, it is not specifically tuned for detection of non-verbal speech elements as laughter. Further, a speech recognition system employing a full-blown transcription may be a bit computationally inefficient for the detection of laughter events. Therefore, we rather built a relatively simple detector based on a small number of acoustic models.

In this Section, we discuss our work on automatic laughter segmentation (see Truong and van Leeuwen [190]). We characterized meetings with three distinct states, namely laughter, speech and silence. The task of the detector is to localize laughter, i.e., to specify beginning and ending of a laughter event, in an audio stream.

Data

As training material, we used the same data as presented in Table 5.2, i.e., the laughter and speech segments from the 26 'Bmr' meetings were used to train the laughter and speech models. Approximately an equal amount of silence was extracted from these meetings to model silence. For testing, the 3 ICSI 'Bmr' meetings as described in Table 5.2 were used.

Features and Method

The laughter, speech and silence GMMs were trained in a similar way as was done for the laughter vs. speech task. In order to determine the segmentation of the acoustic signal into segments representing the N defined classes (in our case $N = 3$), we used a relatively simple Viterbi decoder (Rabiner and Juang [144]). In an N-state parallel topology the decoder finds the maximum likelihood state sequence. The state sequence was used as the segmentation result. The number of state transitions, or the segment boundaries, were controlled by using a small state transition probability. The state transition probability a_{ij} from state i to state $j \neq i$ was estimated on the basis of the average duration of the segments i and the number of segments j following i in the training data. The self probabilities a_{ii} were chosen such that $\sum_j a_{ij} = 1$. After the segmentation into segments $\{s_i\}$, $i = 1, \ldots N_s$, we calculated the average log-likelihoods L_{im} over each segment i for each of the models m. We defined a

log-likelihood-ratio as $L_{laugh} - \max(L_{speech}, L_{silence})$. These log-likelihood-ratios determined final class-membership.

Evaluation and Results

One of the reasons to define log-likelihood ratios for the segments found by the detector, is to be able to compare the current results based on segmentation to other results that were obtained with given pre-segmented segments and that were evaluated with a trial-based DET analysis (Detection Error Tradeoff, Martin et al. [112]). In this analysis we could analyze a detector in terms of DET plots and post-evaluation measures such as Equal Error Rate and minimum decision costs. In order to make comparison possible we extended the concept of the trial-based DET analysis to a (frame-based) time-weighted DET analysis for two-class decoding (van Leeuwen and Huijbregts [194]). The basic idea is (see Fig. 5.2) that each segment in the hypothesis segmentation may have sub-segments that are either

- correctly classified (hits and correct rejects)
- missed, i.e., classified as speech (or other), while the reference says laughter
- false alarm, i.e., classified as laughter, while the reference says speech (or other)

We can now form tuples (λ_i, T_i^e) where T_i^e is the duration of the sub-segment of segment i and e is the evaluation over that sub-segment, either 'correct', 'missed' or 'false alarm'. These tuples can now be used in an analysis very similar to the DET analysis. Define θ as the threshold determining the operating point in the DET plot. Then the false alarm probability is estimated from the set $\mathcal{T}_\theta$ of all tuples for which $\lambda_i > \theta$

$$p_{\mathrm{FA}} = \frac{1}{T_{\mathrm{non}}} \sum_{i \in \mathcal{T}_\theta} T_i^{\mathrm{FA}} \tag{5.2}$$

and similarly the miss probability can be estimated as

$$p_{\mathrm{miss}} = \frac{1}{T_{\mathrm{tar}}} \sum_{i \notin \mathcal{T}_\theta} T_i^{\mathrm{miss}} \tag{5.3}$$

Here T_{tar} and T_{non} indicate the total time of target class (laughter) and non-target class (e.g., speech or silence) in the reference segmentation. This is basically a frame-based (or time-weighted) DET- analysis.

The laughter segmentation experiments were carried out on a total of 18 full-length individual channels of the close-talk recordings taken from the three ICSI 'Bmr' test meetings. The scores (i.e., the log-likelihood ratios) from these separate audio channels were pooled together to obtain EERs. In order to enable better comparison between the laughter vs. speech discrimination and the laughter segmentation results, we have also performed a segmentation experiment on a chain of laughter and speech segments that consisted of the *pre-segmented* laughter and speech segments concatenated to each other randomly. The results are presented in Table 5.8 and Fig. 5.3.

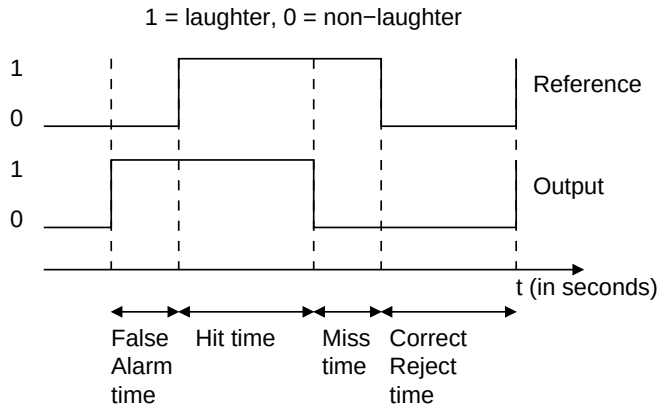


Figure 5.2: *Definitions of correct classifications and erroneous classifications in time.*

Laughter segmentation	
Concatenated laughter/speech	**Whole meetings**
GMM PLP	**GMM PLP**
8.2	10.9

Table 5.8: *EERs of laughter segmentation performed on concatenated laughter and speech segments, or on whole meeting audio streams (tested on 3* ICSI *Bmr meetings).*

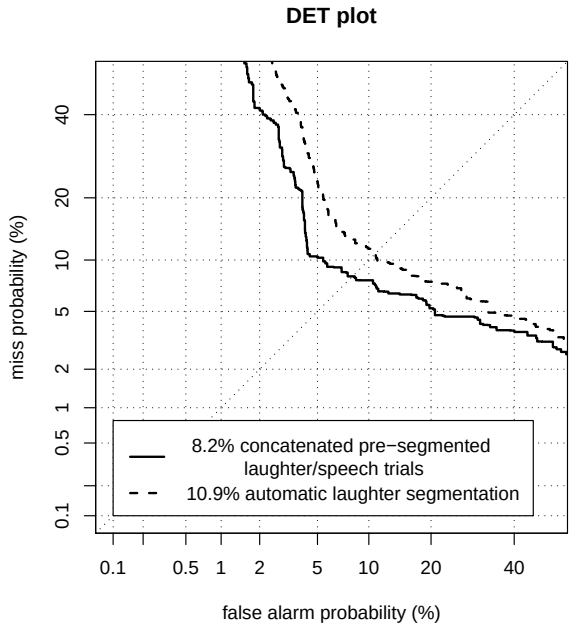


Figure 5.3: *Time-weighted DET curves of laughter segmentation, tested on 3* ICSI *Bmr meetings.*

The 'time-weighted' EER was 10.9% for the 'real' segmentation task and 8.2% for the concatenated laughter-speech segments. The difference in performance is mainly caused by the presence of other non-vocalized sounds, e.g., silence, in meetings. Note that this time-weighted DET curve does not take into account the absolute number of times there was an error, and that it is sensitive to priors since it is time-based.

5.2.7 Example of applied laughter recognition: Affective Mirror

Automatic voice-based laughter recognition can be useful for a range of various applications. For instance, for automatic summarization, data mining, or meta-data generation purposes, laughter can be an important event that signals expressive human behavior. For interactive human-machine communication systems, the social function of laughter can be utilized to create better mutual understanding or more in general, better human-machine communication. One example of a real-time laughter recognition application is that of LAFCam, Leveraging Affective Feedback Camcorder (Lockerd and Mueller [110]) which detects laughter from the camera operator with the goal to enhance the user interface (and experience) during video editing.

In the framework of the Dutch BSIK project MultimediaN, so-called Golden Demos were developed and build that use state-of-the-art multimedia technology to create unique user experiences. In collaboration with our project partners Waag Society[3], V2: Institute for the Unstable Media[4], and VicarVision[5], TNO developed the 'Affective Mirror' (build by Willem Melder, see Melder et al. [117]). The 'Affective Mirror' uses recognition technology and models developed in the work described above to sense laughter in real-time. The goal of this virtual carnival mirror is to deliver a fun and positive experience to the user by sensing and eliciting laughter. Its uniqueness lies in the fact that the 'Affective Mirror' is able to influence the user's state by first sensing the user's emotional state and subsequently, generate appropriate feedback that affects the user, who in turn, will react to that feedback. This way, an interactive loop between user and machine is established. Currently, the system is based on a visual and vocal subsystem that can detect facial expressions and vocal laughter. The mirror detects and reacts to the user's laughter, and then provides visual feedback by distorting the user's face in the virtual mirror, just like a traditional carnival mirror would do. The more one laughs, the further one proceeds in different levels of distortions. The distortions are driven by the amount of laughter, detected by the facial or vocal subsystem. One possible 'Affective Mirror' scenario is presented in Table 5.9. After each session with the mirror, which lasts for a couple of minutes, the user receives a score card with his/her laughter statistics and a photo (see Fig. 5.4).

Several different groups of people have undergone the 'Affective Mirror' experience: playful children, curious parents, and serious scientists have all sat in the front of the mirror. Some visitors started to laugh very quickly and others were more sensitive to the way their behavior influenced the mirror behavior. This resulted into user-mirror cooperative behavior to produce funny distorted faces and reciprocal user-mirror action-reaction cycles in which the user is expressing weird facial and vocal

[3] http://www.waag.org
[4] http://www.v2.nl
[5] http://www.vicarvision.nl

Entering ...	
User	walks to the mirror, is curious and does not know what to expect
Mirror	detects the user's presence and tries to attract the user
User	is intrigued by the mirror and comes closer to see what is going on
Mirror	detects the user's face and highlights the face
Interacting ...	
User	is surprised by the adaptive mirror and waits for the mirror to change
Mirror	classifies the facial expression as 'surprised' and starts a visual effect 'blow up eyes'
User	sees his mirror image being distorted and starts to look happy
Mirror	classifies the facial expression as 'happy' and starts a visual effect 'raise mouth corners'
User	notices that the mirror plays with his facial expression and starts to laugh
Mirror	detects the vocal laughter and starts a visual effect 'swirl'
User	is amazed by the interaction and tries to look disgusted
Mirror	classifies the facial expression as 'disgust' and starts an audio effect 'wobble'
Leaving ...	
User	looses interest and walks away
Mirror	looses track of the face and fades away

Table 5.9: *A possible single user interaction scenario, adopted from Melder et al. [117]*

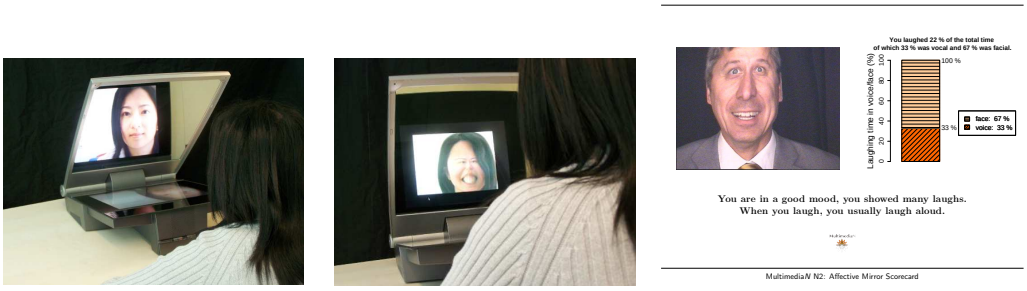


Figure 5.4: *Affective Mirror: an interactive virtual carnival mirror. Left: the mirror, middle: manipulated mirror image, right: the score card.*

behavior. Most visitors found it a positive and fun experience, especially children were very enthusiastic about the mirror. Although the facial and vocal subsystems might not work perfectly under 'real-world' conditions, the mirror has proven to be robust enough to deliver these fun experiences. Moreover, in future research, the 'Affective Mirror' will be used a research tool to investigate, for instance, the effect of other people's presence on the expression of laughter.

5.2.8 Conclusions

In this first half of this Chapter, we investigated laughter vs. speech discrimination and laughter segmentation. For laughter vs. speech discrimination, we experimented with potential discriminative features and modeling techniques. We employed GMMs

and SVMs in combination with PLP, P&E, P&V, and ModSpec features. The two best performing feature sets were PLP and P&V. After several careful fusion combinations, it appeared that a fusion between different classifier and feature-types performs best (i.e., cross-fusions between GMM and SVM methods, and PLP and P&V features): EERs obtained with these combinations ranged between 3% and 9% for speaker-independent test sets. It became clear from these results that the use of different feature types, e.g., spectral and prosodic, in combination with different classifiers boosted performance. For laughter segmentation, we used GMMs trained with PLPs and a relatively simple Viterbi decoder: the task was to localize laughter events in (whole) meetings. In order to be able to express our segmentation results in terms of EER, we evaluated our laughter segmentation with 'time-weighted' (i.e., frame-based) DET curves: the time-weighted EERs obtained ranged between 8% and 11%. Many of the errors made in laughter segmentation were introduced by breath sounds, cough sounds, background noises or crosstalk (softer speech from other participants). Finally, the implementation of our laughter recognition technology in a real-time application namely the 'Affective Mirror' has shown that laughter detection can lead to a fun and positive interactive user experience. The fact that laughter is a relatively frequently occurring affective event that can be detected with an acceptable real-time performance opens up many research and application-oriented opportunities in the field of affective computing. The growing interest for laughter detection in the context of affect recognition, following our initial work on laughter vs. speech discrimination, proofs that laughter is not *just* a non-speech event.

5.3 Multimodal subjectivity analysis in meetings

In this Section, we report on our investigations on feature and classifier combinations for the recognition of subjective content in meetings. Opinions, sentiments and other types of subjective content can play an important role in meetings. Meeting participants express their pros and cons about ideas and they may agree or disagree with opinions. This type of higher-level information can be of value for meeting summarization purposes and can enhance the functionality of meeting browsers. On textual level, a substantial amount of research has been carried out on automatic subjectivity and sentiment recognition in all kinds of (on-line) media, such as blogs, news and reviews. On acoustic level however, little or no work has been carried out on subjectivity recognition. However, terms like 'subjectivity' and 'sentiment' entail phenomena like agreement/disagreement, involvement/hot spots and affect, which *have* gained much interests on acoustic level. A logical step would be to combine these two modalities, acoustic and textual, to aim for better performance in subjectivity recognition. Here, we will focus on two tasks: 1) the recognition of subjective utterances, and 2) the discrimination between positive subjective utterances and negative subjective utterances. For these tasks, we used both acoustic and textual information sources. This Section is structured as follows. In Section 5.3.1, we describe related work on subjectivity and sentiment analysis. The goals and tasks of the subjectivity classification experiments are defined in Section 5.3.2. The material used in the classification experiments is described in Section 5.3.3. In Section 5.3.4, descriptions of the lex-

ical and acoustic features, and the modeling techniques used in the experiments is given. We present our results in Section 5.3.5, and we summarize the conclusions in Section 5.3.6.

5.3.1 Related work

On textual level, there has been a significant amount of research on subjectivity and sentiment recognition, ranging from work at the phrase level to work on classifying sentences and documents. Works on sentence-level subjectivity classification by e.g., Riloff and Wiebe [150], Yu and Hatzivassiloglou [218], and works on sentiment classification by e.g., Yu and Hatzivassiloglou [218], Kim and Hovy [92], Hu and Liu [82], Popescu and Etzioni [139], are most related to our work on subjectivity and polarity classification. Raaijmakers and Kraaij [142] compared wordspanning character n-grams to word-internal character n-grams for subjectivity classification in news data. They found that character n-grams spanning words perform the best. In the context of meetings or multiparty conversation, subjectivity research include works by Somasundaran et al. [180] who recognized sentiments and arguing in meetings. Somasundaran et al. [180] used lexical and discourse features to recognize sentences and turns where meeting participants express sentiments or arguing. They also used the AMI corpus but different annotations and task definitions. Wilson and Raaijmakers [209] compared the use of word n-grams, character n-grams, and phoneme n-grams in the AMI corpus for recognizing subjective utterance in multiparty conversation, and showed that character n-grams from a manual reference transcription performed best.

In acoustics, subjectivity as a topic has not been investigated frequently, however, sentiments and emotions (related to subjectivity) in meetings have. Neiberg et al. [124] used spectral features (MFCCs) and pitch features and lexical n-grams for recognizing emotions in the ISL Meeting Corpus (Burger et al. [24]). Agreement and disagreement recognition (using both lexical and prosodic cues), and hotspot detection in meetings were investigated by e.g., Hillard et al. [80], Galley et al. [65], Hahn et al. [72], and Wrede and Shriberg [211] respectively. Hotspots are events in meetings where the participants are highly involved in a discussion. Although high involvement does no necessarily equate subjective content, in practice, we expect more sentiments, opinions, and arguments to be expressed during heated discussions.

In our work, we follow-up the study carried out by Wilson and Raaijmakers [209] and extend it with new research questions. Wilson and Raaijmakers [209] showed that very shallow character and phoneme representations yield promising results for subjectivity detection. We extended this work and added another information source to the textual information source, namely prosody. In our study we made all possible combinations of these multimodal information sources and fused these sources in several different ways. In addition to subjectivity recognition, we also performed polarity classification, i.e., we classified whether the subjective sentence is positive or negative.

5.3.2 Defining the tasks and goals

We analyzed subjectivity in the context of multiparty conversation in meetings, but how exactly is **subjectivity** defined in this context? For this work, we used the AMI Meeting Corpus (Carletta [35], see section 5.3.3 for a brief description of the data). The AMI Meeting Corpus has been annotated for subjective content using an AMIDA annotation scheme described in Wilson [208]. There are three main categories of annotations, namely *Subjective Utterances*, *Subjective Questions*, and *Objective Polar Utterances*, see Table 5.10.

Subjective Utterances	positive subjective, negative subjective, uncertainty, other subjective, positive and negative subjective
Subjective Questions	positive subjective question, negative subjective question, general subjective question
Objective Polar Utterances	positive objective, negative objective

Table 5.10: *AMIDA subjectivity annotation types.*

A *Subjective Utterance* is a span of words where a *private state* is being expressed either through a choice of words or through prosody in the voice. A *private state* (Quirk et al. [140]) is an internal mental or emotional state, including opinions, beliefs, sentiments, emotions, evaluations, uncertainties, and speculations, among others. Examples of Subjective Utterances are given in (1) and (2) (Wilson and Raaijmakers [209], Wilson [208]):

> (1) so I believe the the advanced functions should maybe be hidden in a drawer, or something like that from the bottom of it
> (2) people uh additionally arent arent liking the appearance of their products

Positive Subjective Utterances include agreements, positive sentiments (emotions, evaluations and judgments), positive suggestions, arguing for something or beliefs from which positive sentiments can be inferred. Negative Subjective Utterances are typically comprised of disagreements, negative sentiments, negative suggestions, etc. Example (3) contains two Positive Subjective Utterances and one Negative Subjective Utterance (indicated by a pair of angle brackets):

> (3) Um <**POS-SUBJ** it's very easy to use>. Um <**NEG-SUBJ** but unfortunately it does lack the advanced functions> <**POS-SUBJ** which I I quite like having on the controls>.

The Positive And Negative Subjective category is for marking cases of positive and negative subjectivity that are so closely interconnected that it is difficult or impossible to separate the two. For example, some subjective words or phrases inherently evoke both a positive and negative sentiment. An example of such a word is *bittersweet*.

The category Uncertainty includes utterances from individuals who express their uncertainty, or utterances that indicate undecided things. An example of Uncertainty is given in (4):

(4) Um I'm not entirely sure what the corporate color is.

Subjective Questions are questions in which the speaker is eliciting the private state of someone else. In other words, the speaker is asking about what someone else thinks, feels, wants, likes etc., and the speaker is expecting a response in which the other person expresses what he/she thinks, feels, wants, or likes etc. If the person is specifically trying to elicit a positive or negative private state of someone else, this is annotated as a Positive or Negative Subjective Question. Subjective Questions that do not specifically ask for a positive or negative state of someone else are General Subjective Questions. For example, (5) and (6) are General Subjective Questions:

(5) Do you like the large buttons?
(6) What do you think about the large buttons?

Objective Polar Utterances are utterances that describe positive or negative factual information about something without conveying a private state. Examples are given in (7) and (8). Generally, breaking something the first time it is used is not good, so this is marked as Negative Objective. (8) is an example of Positive Objective.

(7) The camera broke the first time I used it.
(8) The camera lasted for several years past its warranty.

Wilson [208] performed an agreement study and measured agreement for each class separately at the level of dialogue act segments. Table 5.11 gives the Kappa (Cohen [41]) and % agreement for Subjective Utterances, Positive Subjective Utterances, Negative Subjective Utterances, and Subjective Questions.

	Kappa	% Agree
Subjective Utterances	0.56	79
Positive Subjective Utterances	0.58	84
Negative Subjective Utterances	0.62	92
Subjective Questions	0.56	95

Table 5.11: *Inter-annotator agreement for the AMIDA subjectivity annotations.*

We defined two main binary decision tasks in this study. The first task is to discriminate between *Subjective* and *Non-Subjective* utterances. An utterance is considered Subjective if it falls in the category Subjective Utterances or Subjective Questions. The second task is to discriminate between *Positive Subjective Utterances* and *Negative Subjective Utterances*. For this task, the utterances that are both Positive and Negative Subjective are excluded.

Using multimodal information sources, lexical and acoustic, we investigated subjectivity recognition in the context of meetings and we focused on the following questions:

- Given multiple acoustic and lexical features, which of these sources are particularly valuable for subjectivity analysis in multiparty conversation?

- Does the combination of these sources/features lead to further improvement?
- What are the optimal representations of these sources/features from a machine learning point of view?

5.3.3 *Material:* AMI *Meeting Corpus*

We used 13 meetings[6] from the AMI Meeting Corpus (Carletta [35]). Each meeting has four meeting participants and is approximately 30 minutes long. The participants play specific roles (e.g., Project Manager, Marketing Expert) and together function as a design team. Within the set of 13 meetings, there are 20 participants, with each participant taking part in two or three meetings as part of the same design team. For a brief description of the AMI corpus, the reader is referred to Section 2.4.1. Table 5.12 lists the number of utterances (and their mean durations) used in the classification experiments.

		N	mean and standard deviation duration
Task 1	Subjective	6226	$\mu = 1.9\,\text{s}$, $\sigma = 2.0\,\text{s}$
	Non-Subjective	8707	
Task 2	Positive Subjective	3157	$\mu = 2.6\,\text{s}$, $\sigma = 2.3\,\text{s}$
	Negative Subjective	1052	

Table 5.12: *Material used in classification experiments.*

5.3.4 *Method and Features*

For the subjectivity recognition tasks, we used prosodic features and three different text representations: word, character and phoneme-level transcriptions. These features were used in combination with the boosting algorithm AdaBoost.

Lexical (textual) features

We employed three types of textual features which are all based on a manual transcription of the speech on different levels: word-level (WORDS), character-level (CHARS), and phoneme-level (PHONES). The PHONES were produced through dictionary lookup on the words in the reference transcription. Both CHARS and PHONES representations include word boundaries as informative tokens. The textual features for a given utterance are simply all the WORDS, CHARS or PHONES in that utterance. Selection of n-grams is performed by the learning algorithm. Examples of these representations are given in Table 5.13.

Acoustic features

Based on earlier research on the acoustics of emotion (e.g., Banse and Scherer [12]) and 'hot spots' (e.g., Wrede and Shriberg [211]), we extracted prosodic features

[6] ES2002b, ES2002c, ES2002d, ES2008b, ES2008c, ES2009b, ES2009c, ES2009d, IS1003c, IS1003d, TS3005b, TS3005c, TS3005d

WORDS	yeah i like the idea
CHARS	WB y e a h WB i WB l i k e WB t h e WB i d e a WB
PHONES	y eh ax sil ay sil l ay k sil dh ax sil ay d iy ax sil

Table 5.13: *Examples of different feature representations.*

(PROS) that are mainly based on pitch, energy, and the distribution of the energy in the Long-Term Averaged Spectrum (LTAS), see Table 5.14. These features were extracted at word-level and aggregated to the dialogue-act level by taking the average over the words per dialogue act. We then normalized the features per speaker per meeting by converting the raw feature values to z-scores ($x_{\mathrm{z}} = (x_{\mathrm{raw}} - \mu)/\sigma$). The program Praat (Boersma and Weenink [23]) was used to extract the acoustic features.

Pitch	mean, standard deviation, minimum, maximum, range, mean absolute slope
Intensity (energy)	mean, standard deviation, minimum, maximum, range, RMS energy
Distribution energy in LTAS	slope, Hammarberg index, center of gravity, skewness

Table 5.14: *Acoustic features PROS.*

Learning method: AdaBoost

The AdaBoost algorithm was used as classification method. As described in Section 2.4.3, AdaBoost is an iterative meta-algorithm that combines many simple weak learners or rules into one single, strong classifier. For our classification experiments, we used BoosTexter (Shapire and Singer [175]), an implementation of the AdaBoost algorithm that is specifically attuned to text categorization tasks. In the case of BoosTexter, these weak rules have the same basic form as that of a one-level decision tree. The test at the root of this tree is a simple check for the presence or absence of a term in the given text. In case of continuous values, the test checks if a value is above or below a certain threshold. We chose to use BoosTexter because, in addition to it having a proven track record for working well for many NLP tasks, the tool's parameters allow for easy trial of many different n-gram configurations. An additional advantage of BoosTexter is that it can deal with both continuous-valued input (e.g., age) and textual input (e.g., a text string) at the same time.

5.3.5 Evaluation and results

In this Section, we discuss the experimental setup and present our results. We discuss how we trained and tested the single-source classifiers, and how these classifiers were combined via weighted linear combinations to investigate what combination of features and classifiers are most valuable for subjectivity recognition.

Experimental setup

The experiments were performed using 13-fold cross validation. Each meeting constitutes a separate fold for testing, e.g., all the segments from meeting 1 make up the test set for fold 1. Then, for a given fold, the segments from the remaining 12 meetings were used for training and parameter tuning, with roughly a 85%, 7%, and 8% split between training, tuning, and testing sets for each fold. The assignment to training versus tuning set was random, with the only constraint being that a segment could only be in the tuning set for one fold of the data. As main performance measure, we report C_{det} with equal prior probabilities ($= 0.5$) and equal costs ($= 1$) (see also Section 2.4.4). In addition, we report the False Rejection Rate (FRR), False Alarm Rate (FAR), F_1 (harmonic mean between precision and recall), and accuracy. Note that threshold calibration was not performed. For significance tests, we used Wilcoxon Signed Rank test, two-sided, $p < 0.05$.

The classification experiments performed involved two steps. First, we trained and optimized a classifier for each type of feature separately using BoosTexter; we call these the 'single-source' classifiers. Then, we investigated the performance of all possible combinations of features using linear combinations of the individual feature classifiers.

Single-Source Classifiers Four single-source classifiers were trained using BoosTexter, one for each type of feature. For the WORDS, CHARS, and PHONES, we optimized the classifier by performing a grid search over the parameter space, varying the number of rounds of boosting (100, 500, 1000, 2000, 5000), the length of the n-gram (1, 2, 3, 4, 5), and the type of n-gram. BoosTexter can be run with three different n-gram configurations: n-gram, s-gram, and f-gram. For the default configuration (n-gram), BoosTexter searches for n-grams up to length n. For example, if $n = 3$, BoosTexter will consider 1-grams, 2-grams, and 3-grams. For the s-gram configuration, BoosTexter will in addition consider sparse n-grams (i.e., n-grams containing wildcards), such as *the * idea*. For the f-gram configuration, BoosTexter will only consider n-grams of a maximum fixed length, e.g., if $n = 3$ BoosTexter will only consider 3-grams. For the PROS classifier, only the number of rounds of boosting was varied. The parameters were selected for each fold separately; the parameter set that produced the lowest error rate (we used C_{det}) on the tuning (development) set is used to train the final classifier for that fold.

Classifier combination After the single-source classifiers have been trained, they were combined into an aggregate classifier. To this end, we decided to apply a simple linear interpolation strategy. Linear interpolation of models is the weighted combination of simple models to form complex models, and has its roots in generative language models (Jelinek and Mercer [86]). Raaijmakers [141] has demonstrated its use for discriminative machine learning. In the present binary class setting, BoosTexter produces two decision values, one for each class. For each individual single-source classifier (i.e., PROS, WORDS, CHARS and PHONES), separate weights were estimated that were applied to the decision values for the two classes produced by these classifiers. These weights express the relative importance of the single-source classi-

fiers. The prediction of an aggregate classifier for a class c is then simply the sum of all weights for all participating single-source classifiers applied to the decision values these classifiers produce for this class. The class with the maximum score wins, just as in the simple non-aggregate case.

Formally, then, this linear interpolation strategy finds for n single-source classifiers n interpolation weights $\lambda_1, \ldots \lambda_n$ that minimize the empirical loss (measured by a loss function $\mathcal{L}$), with λ_j the weight of classifier j ($\lambda \in [0, 1]$), and $C_c^j(x_i)$ the decision value of class c produced by classifier j for datum x_i (a feature vector). The two classes are denoted with $0, 1$. The true class for datum x_i is denoted with $\hat{x}_i$. The loss function is in our case based on C_{det}, measured on heldout development training and test data.

The aggregate prediction $\tilde{x_i}$ for datum x_i on the basis of n single-source classifiers then becomes

$$\tilde{x_i} = \arg\max_c (\sum_{j=1}^{n} \lambda_j \cdot C_{c=0}^j(x_i), \sum_{j=1}^{n} \lambda_j \cdot C_{c=1}^j(x_i)) \tag{5.4}$$

and the lambdas are defined as

$$\lambda_j^n = \arg\min_{\lambda_j^n \subset [0,1]} \sum_{i}^{k} \mathcal{L}(\hat{x}_i, \tilde{x}_i; \lambda_j, \ldots, \lambda_n) \tag{5.5}$$

The search process for these weights can easily be implemented with a simple grid search over admissible ranges. In the experiments described below, we investigated all possible combinations of the four different sets of features (PROS, WORDS, CHARS, and PHONES) to determine which combination yields the best performance for subjectivity and subjective polarity recognition.

Results

Results for the two tasks are given in Table 5.15. We report the results of two baseline classifiers: one that randomly chooses a class based on class priors (BASE-RAND) and one that always chooses the *target*-class (BASE-SUBJ for subjectivity recognition and BASE-POSS for Positive Subjectivity detection). The bullets in a given row indicate the features that were evaluated for a given experiment. All values in Table 5.15 are averages over 13 folds.

We can observe in Table 5.15 that the combination of different sources of information is beneficial, and in general, the more information sources are used, the better the performance. The best results for Task 1 were obtained with all four information sources, achieving a C_{det} of 26.7. For Task 2, the best results were also obtained with the four information sources, except PROS (including PROS did not significantly improve performance), achieving a C_{det} of 26.6.

The effects of adding more information to the single-source classifiers was also measured. From Table 5.16, it seems that, of the various feature types, prosody PROS seems to be least informative for both subjectivity recognition and polarity classification. In all cases, except for PROS, adding an extra information source yields significantly better performance. In addition, all other single-source classifiers significantly outperform the single-source classifier based on PROS, see Table 5.17 and 5.18. Fur-

	TASK 1: SUBJECTIVE VS. NON-SUBJECTIVE								
	PROS	WORDS	CHARS	PHONES	C_{det}	FRR	FAR	F_1	A
UNI	•				38.2	45.6	30.8	54.6	63.0
		•			31.5	45.7	17.3	60.4	71.0
			•		31.0	42.9	19.0	61.7	71.2
				•	31.8	44.4	19.2	60.5	70.5
BI	•	•			28.9	43.1	14.7	63.9	73.6
	•		•		28.0	40.0	16.1	65.5	74.1
	•			•	28.4	41.8	14.9	64.8	74.0
		•	•		27.3	40.5	14.2	66.1	75.0
		•		•	27.9	41.5	14.2	65.4	74.6
			•	•	27.4	40.0	15.2	66.3	74.8
TRI	•	•	•		27.2	40.3	14.1	66.3	75.2
	•	•		•	27.7	41.2	14.1	65.6	74.8
	•		•	•	27.1	39.3	15.0	66.6	75.0
		•	•	•	26.8	39.5	14.2	66.9	75.4
QUADRI	•	•	•	•	**26.7**	**39.3**	**14.0**	**67.1**	**75.6**
BASE-SUBJ					50.0	0	100	60.3	43.4
BASE-RAND					48.9	56.9	40.9	43.5	52.0

	TASK 2: POSITIVE SUBJ. VS. NEGATIVE SUBJ.								
	PROS	WORDS	CHARS	PHONES	C_{det}	FRR	FAR	F_1	A
UNI	•				49.7	8.9	90.1	82.0	70.5
		•			35.5	10.3	60.8	85.2	77.0
			•		34.7	12.5	57.0	84.5	76.3
				•	35.6	10.7	60.5	85.0	76.8
BI	•	•			29.1	7.7	50.6	88.2	81.5
	•		•		28.6	9.1	48.1	87.8	81.0
	•			•	29.3	7.9	50.7	88.1	81.3
		•	•		27.3	7.0	47.6	89.0	82.8
		•		•	28.0	6.8	49.2	88.9	82.5
			•	•	27.5	7.8	47.3	88.6	82.3
TRI	•	•	•		27.1	6.8	47.5	89.1	83.0
	•	•		•	27.9	6.4	49.4	89.0	82.8
	•		•	•	27.5	7.5	47.4	88.7	82.4
		•	•	•	26.8	6.7	46.8	89.2	83.2
QUADRI	•	•	•	•	**26.6**	**6.5**	**46.7**	**89.4**	**83.4**
BASE-POSS					50.0	0	100	85.6	75.0
BASE-RAND					50.1	26.4	73.8	74.1	61.7

Table 5.15: *Results Task 1 and Task 2. Reported are C_{det}, FRR (False Rejection Rate aka Miss Rate), FAR (False Alarm Rate), F_1 (harmonic mean between recall and precision), A (accuracy=number of correct classifications/total number of utterances).*

thermore, the single-source classifier PROS achieves a better performance for Task 1 than Task 2; PROS is thus more useful for subjectivity recognition than for polarity classification. The best performing feature types are both CHARS and WORDS: from the single-source classifier results in Table 5.17 and 5.18, we can observe that CHARS are not significantly better than WORDS.

A possible question that remains is what the effect is of classifier interpolation on the results. To answer this question, we performed two additional classification exper-

	+PROS		+ WORDS		+CHARS		+PHONES	
	1	2	1	2	1	2	1	2
PROS			+	+	+	+	+	+
WORDS	+	+			+	+	+	+
CHARS	-	+	+	+			+	+
PHONES	+	+	+	+	+	+		
PROS+WORDS					+	+	+	+
PROS+CHARS			+	+			+	+
PROS+PHONES			+	+	+	+		
WORDS+CHARS	+	-					+	+
WORDS+PHONES	+	-			+	+		
CHARS+PHONES	+	-	+	+				
PROS+WORDS+CHARS							+	+
PROS+WORDS+PHONES					+	+		
PROS+CHARS+PHONES			+	+				
WORDS+CHARS+PHONES	+	-						

Table 5.16: *Addition of features separately (for Task 1 and 2). '+' for a row- column pair (r, c) means that the addition of column feature c to the row features r significantly improved r's F . '-' indicates no significant improvement.*

	WORDS	CHARS	PHONES
PROS	$<$	$<$	$<$
WORDS		$=$	$=$
CHARS			$>$

Table 5.17: *Task 1: significance single-source classifiers.*

	WORDS	CHARS	PHONES
PROS	$<$	$<$	$<$
WORDS		$=$	$=$
CHARS			$=$

Table 5.18: *Task 2: significance single-source classifiers.*

iments for both tasks. First, we investigated the performance of an **uninterpolated** combination of the four single- source classifiers, that is, $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$. In essence, this combines the separate feature spaces without explicitly weighting them. Second, we investigated the results of training a single BoosTexter model using all the features, essentially merging all feature spaces into one agglomerate feature space at once (**feature-level** fusion). The results in Table 5.19 and 5.20 show that interpolation significantly outperforms the uninterpolated model for both tasks. The interpolated model outperforms the feature-level fusion significantly only in Task 2. The uninterpolated model appears to be performing similar to the feature-level fusion.

Task	Combination	C_{det}	FRR	FAR	F_1	A
	interpolated	**26.7**	**39.3**	**14.0**	**67.1**	**75.6**
1	uninterpolated	28.7	41.7	15.8	64.4	73.6
	feature-level	27.7	38.8	16.7	66.0	74.2
	interpolated	**26.6**	**6.5**	**46.7**	**89.4**	**83.4**
2	uninterpolated	32.7	4.2	6.1	8.5	8.1
	feature-level	30.8	9.2	52.5	86.9	79.6

Table 5.19: *Results interpolated, uninterpolated and feature-level fusion.*

		uninterpolated	feature-level
Task 1	interpolated	$>$	$=$
	uninterpolated		$=$
Task 2	interpolated	$>$	$>$
	uninterpolated		$=$

Table 5.20: *Comparing performances between interpolated, uninterpolated and feature-level fusion models ('$<$' means significantly worse, '$>$' means significantly better, $p < 0.05$).*

5.3.6 Conclusions

We compared the use of prosodic features, word n-grams, character n- grams, and phoneme n-grams for subjectivity recognition and polarity classification. The classification experiments showed that prosody is outperformed by textual features in both subjectivity and polarity detection. Prosody performed substantially worse in polarity classification than in subjectivity recognition: as a single-source classifier PROS achieves a C_{det} of 49.7 as close as classification by chance. As an additional feature, PROS does not always significantly improve performance in polarity classification (see Table 5.16). WORDS, CHARS and PHONES on the other hand, appear to be very competitive feature sets: word n-grams, character n-grams and phoneme n-grams all achieve similar performances, with a small advantage for character n-grams since these, in the single-source classifier case, significantly outperformed phoneme n-grams in Task 1 subjectivity recognition (see Table 5.17). Combining all information sources available yielded the best performances, except for Task 2 where the addition of PROS to WORDS, CHARS and PHONES did not significantly improve performance. We have also shown that interpolation outperforms the unweighted and the feature-level fusion combination significantly, at least for Task 2 (see Table 5.20).

To conclude, we can answer the questions posed in Section 5.3.2 as follows: a) textual representations in the form of words and characters are relatively valuable information sources for subjectivity recognition and polarity classification, b) the combination of these sources carried out by an interpolation strategy yields significantly better performances, and c) the optimal representation of features for subjectivity recognition includes CHARS, WORDS, PHON, and PROS features, while for polarity classification, PROS can be excluded.

5.4 Discussion and conclusions

This Chapter dealt with the recognition of spontaneous emotionally colored behavioral phenomena in a meeting context. The first topic focused on the detection and segmentation of **laughter**. The second topic focused on **subjectivity** analysis and **polarity** classification of subjective utterances. We discuss ideas for further research and elaborate on possible explanations for the results found.

For laughter vs. speech discrimination, a combination of spectral and prosodic features and different learning algorithms yielded the best performances. The errors made by the classifier suggest that it might be a good idea to define different types of

laughter and develop separate laughter-type-dependent classifiers. A first division in laughter types could be that of voiced laughter vs. unvoiced laughter since they not only differ significantly from each other acoustically, they also elicit different emotions (Bachorowski and Owren [10]). Furthermore, a study by Campbell et al. [30] indeed showed that laughters *can* be classified into several types of classes.

Our laughter research has focused on the acoustics only, but when one laughs, one obviously also uses facial muscles. Audio-visual laughter detection has recently gained much attention by the works of e.g., Ito et al. [84], Reuderink et al. [147], and Petridis and Pantic [130, 131, 132]. Both Reuderink et al. [147] and Petridis and Pantic [130, 131, 132] use AMI Meeting data and perform decision-level fusion which yielded the best performance. Future research can for example focus more closely on the inter-relations between the vocal and visual act of laughing in the temporal domain.

Furthermore, it has been suggested that laughter can reveal someone's identity and can enhance speaker recognition systems. Research has shown that people can use laughter as a cue to someone's identity (e.g., Bachorowski et al. [11]). Knox and Mirghafori [94], Knox et al. [95] have developed laughter detection systems with the aim to enhance speaker recognition systems, but no speaker recognition performance results have been reported yet. Hence, future work can focus on the integration of automatic laughter detection in speaker recognition systems with the aim to improve performance.

Finally, as an expression of emotion or paralinguistic event or 'affect burst' as Schröder [167] calls it, laughter carries important cues for someone's emotional state. Note that in our work, we have only detected laughter without giving an interpretation to it. We foresee that a laughter detector can serve as one of the input modules for meta-analyzers that can give an interpretation of the laughter or a prediction of someone's emotional state based on separate modules/detectors, e.g., raised voice detector, harsh voice detector, cough detector etc. (see Fig. 5.5).

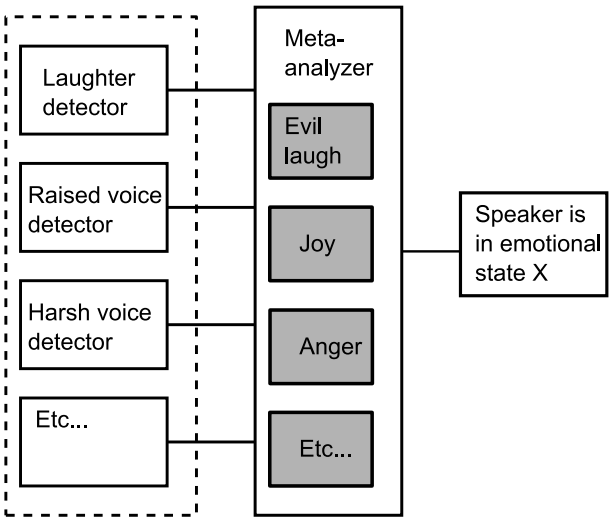


Figure 5.5: *Emotion detection based on low-level event detectors.*

In the second part of this Chapter, we investigated the use of shallow linguis-

tic representations such as character n-grams and phoneme n-grams for subjectivity recognition and polarity classification. It appeared that these representations work at least as good as (or even better than) word n-grams. Phoneme n-grams are interesting because now that we know that they perform similar to word n-grams, this opens up interesting possibilities for the development of real-time subjectivity recognition systems. Automatic phoneme recognition has a lower latency than ASR which is currently much slower than real-time. A system based on phoneme n-grams can make a halt at phoneme level, which speeds up subjectivity classification since phoneme recognition comprises one of the first levels in ASR. However, the reason why phoneme n-grams and character n-grams perform so well remains a bit unclear. Character n-grams have previously been used successfully in named-entity recognition (Klein et al. [93]) and subjective sentence recognition (Raaijmakers and Kraaij [142]). We suspect that character n-grams are flexible enough to capture several types of linguistic information useful for subjectivity recognition. For example, part-of-speech information that can be correlated with subjective language use can be captured in the character 4-gram `ould` which covers modal verbs like *would, could, should* or in the character 3-gram `ly#` which covers the set of adverbs ending in *-ly*.

In addition to textual features, we also used prosodic features for subjectivity recognition and polarity classification. Subjectivity has not been investigated yet (to the best of our knowledge) in the context of prosody. However, phenomena that are closely related to subjectivity such as involvement, (dis-) agreement and 'hot spots', have been studied in the context of prosody and have shown that prosody can be predictive of the aforementioned phenomena. Our experiments have shown that prosody can also be predictive for subjectivity, although its predictive power is smaller than textual features. Note that we used a relatively small set of prosodic features and that this set can be extended to the use of e.g., spectral features. Also, the boosting algorithm used here might not be the best learning algorithm for acoustic features. Hence, the performance of the acoustic subjectivity recognizer could be improved by the use of extra features or other learning algorithms (which fell outside the scope of this work). For polarity classification, we expected that prosody would be outperformed by textual features. It remains a challenge to find acoustic features that are predicative of positiveness and negativeness in speech. More (fundamental) research is needed to uncover the acoustic characteristics of polarity classification (Positive vs. Negative emotions) in speech.

The results obtained in this Chapter prove that a combination of several different information sources and algorithms can boost performance significantly. For laughter and subjectivity recognition, most of the time it is better to combine several information sources on decision-level and to weight separately developed classifiers or to use an algorithm that can learn and attribute weights to the decision values of the separate classifiers. The fusion strategies we used in this Chapter are relatively straightforward. When more multimodal information streams, e.g., facial expression recognition, gesture recognition, posture recognition etc., become available, more complex fusion strategies will be needed to cope with the several streams of data and all its difficulties: what to do if one or two of the information streams are missing, how to define a common unit of analysis and how to synchronize these units, how to

deal with incongruent decisions etc. We have not tackled these complexities in this current work, but it certainly needs to be tackled in the future, especially since the expression of emotion involves multimodal processes which may behave asynchronously and incongruently.

Finally, it has become clear that affect in naturalistic data, in this case meetings, can express itself in subtle ways that is more related to conversational behavior rather than primary basic emotions. Even in natural speech data, the performance was relatively good for laughter detection (EER 3%–9%) which can probably be attributed to the fact that laughter is a relatively distinct event. For subjectivity and subjective polarity recognition, error rates were much higher with EERs around 27%. From an affective perspective, subjectivity is a less distinct concept that is only indirectly linked to affect; we *assume* that people are more affectively expressive when people express their personal opinions rather than factual statements, but it is not a prerequisite. Subjectivity is perhaps too broadly defined and it is perhaps expressed in very subtle ways that is not 'detectable' enough by our current affect recognition technology. It seems that relatively distinct non-verbal vocal events such as laughter are interesting events to detect for affect recognition since 1) recognition technology appears to be fit for this detection task, and 2) they carry low-level affective information. The logical next step would be to analyze this low-level information, using the conversational context, to grasp a higher-level understanding of meetings or conversations (see e.g., Fig. 5.5).

Chapter 6

Arousal and Valence prediction: felt versus perceived

Assigning labels to natural affective signals is a complex and time-consuming process that is susceptible to subjectiveness. The main difficulty in natural emotion annotation is the absence of a 'ground truth'; what does one consider as an appropriate emotion label of a specific signal? One obvious possible consideration is to ask persons who have undergone the emotion to describe what they **felt**. An alternative is to ask observers to describe what emotional expressions they **perceive** from the person who is undergoing the emotion; this procedure is currently the most common one in emotional speech research.

In this Chapter, the challenge is taken up to develop a speech-based emotion recognizer that can detect **felt** emotions. A speech-based emotion recognizer that can detect **perceived** emotions has been developed in parallel so that these two recognizers can be compared. The underlying assumption that we make, is that emotion annotations made by the persons who have undergone the emotions themselves are 'more true', and more closely approximate 'ground truth' than the annotations made by naive observers 'More true' in the sense that these labeled expressions better reflect the emotional meaning that the sender intended to send. In order to develop these recognizers, spontaneous affective audiovisual data was collected with subjects who are playing a videogame. All data was annotated by the subjects themselves (felt), and a subset was also annotated by a group of naive observers (perceived/observed). With these data, two experiments were performed. The first experiment involved a relatively small perception experiment that aimed at investigating how observers agree on emotion when unimodal or multimodal information is provided, and how annotations made by the gamers themselves differ from annotations made by observers. In the second experiment, we compared the use of the annotations made by the gamers themselves and the annotations made by the observers for the development of speech-based emotion recognizers. Furthermore, these machine performances were compared to human performance. One of the key elements of the emotion recognizers developed is that these are developed to predict Arousal and Valence scalar values rather than to detect emotion categories.

This Chapter is structured as follows. The development and annotation proce-

dures of the audiovisual emotion database of gamers are described in Chapter 6.2. Chapter 6.3 describes the experimental setup and the results of the first experiment. In Chapter 6.4, we describe the second experiment and we present an analysis on the use of 'felt'-annotations vs. 'perceived'-annotations for the development of speech-based affect recognizers. Finally, we discuss important findings of both studies in Chapter 6.5.

6.1 Emotion labeling: felt vs. perceived emotions

In emotion recognition research, 'ground truth' labels to be used for the development of emotion recognizers, are difficult to acquire and are to a certain extent subjective. There is (usually) no discussion about who is speaking or what language he or she is speaking, but people do not always agree on what emotional state a person is in. Hence, the labeling (annotation) of spontaneous expressive corpora remains a major topic in emotion research. A frequently adopted approach to acquire 'ground truth' labels for expressive signals, is to have several (naive) humans to label these signals. When these annotators (or a majority of annotators) agree with each other, we can consider their judgments as 'ground truth'. We can distinguish two types of annotators: one that is a naive observer who annotates the observed or perceived emotion, and one who labels his/her own **felt** ('self') emotions that he/she has just undergone. From an emotion recognition perspective, it is important to know how the emotion signals were labeled and by whom. If the labels are annotated by observers who label perceived emotion, the machine will learn to recognize perceived emotion. For some people, the ultimate goal is to develop a machine that can recognize a person's felt emotions. So far, we have not seen results of speech-based emotion recognizers (to the best of our knowledge) that are developed with felt emotion labels to detect felt emotions, although the literature does report some studies on the use of 'self' labeling for emotion corpora.

The majority of emotion corpora contain emotion annotations that are made by (naive) observers. Only a small number of studies has investigated the use of annotations that are made by the subject who has undergone the emotion him/herself for expressive corpora. Aubergé et al. [8] proposed to use 'auto-annotation', annotation performed by the subject him/herself, as an alternative method to label expressive corpora. The subjects were asked to label what they *felt* rather than what they *expressed*. There were no conclusive results: they concluded that 'felt'-annotations or 'expressed'-annotations both have their strengths and weaknesses. In Busso and Narayanan [26], the expression and perception of emotions were studied and 'self'-assessments of emotion were compared to assessments made by observers. In that study, they found a mismatch between felt and perceived emotions. The 'self'-raters appeared to assign their own emotions to more specific emotion categories which led to more extreme values in the Arousal-Valence space.

In the current study, we analyze differences between felt and perceived emotion annotations and investigate what the consequences of these observations are for the development of automatic speech-based affect recognizers. To that end, we first recorded an audiovisual emotion database of subjects playing videogames.

6.2 *The* TNO-GAMING *corpus: a corpus of gamers' vocal and facial expressions*

In this Section, it is described how we elicited and recorded spontaneous audiovisual emotion data with subjects playing a videogame. Furthermore, we describe results obtained with the 'self'-annotation, performed by the gamers themselves.

6.2.1 Participants

Seventeen males and eleven females with an average age of 22.1 years (2.8 standard deviation) participated in the gaming experiment. Participants played a videogame against each other in teams of two against two. We asked each participant to bring along a friend as team mate. A compensation was paid to all participants. Fifteen participants were relatively experienced gamers, while thirteen participants hardly ever or never played videogames (see Table 6.1).

How often do you play videogames?	
	number of participants
each day	5
1–3 times a week	10
hardly ever	8
never	5

Table 6.1: *Gaming experience of participants.*

6.2.2 Recordings

Speech recordings were made with high quality close-talk microphones that were attached near the mouth to minimize the effect of crosstalk (speech from other speakers) and other background noise. Recordings of facial expressions were made with high quality webcams (Logitech Quickcam Sphere). The webcams were placed at approximate eye-level on top of the monitor such that a frontal view of the face was captured under an angle that was acceptable for reliable automatic facial recognition. Further, lighting and background conditions were controlled by adjusting the light when needed and by placing evenly colored dark curtains behind the participants to avoid clutter and noise in the background. Noldus' FaceReader (by VicarVision [4], an automatic face recognition software application) was used to test the quality of the video recordings under these environmental settings and conditions. The game content itself was also stored by capturing the frames (per 1 second) of the video stream during game play.

6.2.3 Procedure

At the beginning of the gaming experiment, the participants received general instructions and training sessions to get acquainted with the game and the annotation task. Each participant played the game twice (2×20 minutes). Each gaming session was

followed by a break and the annotation tasks. Between the first and the second game session the participants had a long break.

6.2.4 The game

The participants played a multiplayer first-person shooter videogame called *Unreal Tournament 2004*, developed by Epic Games. The gamemode 'Capture the flag' was selected in which the goal was to capture each other's flag as many times as possible.

6.2.5 Eliciting emotions

The goal was to evoke a broad range of different emotions, including emotions like Frustration, Joy, Amazement and Malicious Delight. We employed several strategies to evoke these emotions and to stimulate vocal and facial expressive behavior:

1. each participant had to bring a friend as team mate.
2. bonuses were granted to the winning team, and the team with 'best collaboration'.
3. surprising events were generated in the game, for example, sudden deaths, sudden appearances of monsters, and hampering keyboard or mouse controls, were inserted in the game (at an approximate rate of one event per minute).

6.2.6 Annotation procedure

After each game session, the participants watched their own videos recorded and judged their own emotions in two different ways: one based on emotion categories and the other one based on emotion dimensions. In addition next to the video recorded, the video stream of the game itself was also provided as context information. The participants annotated the running video and could not pause or rewind the video. Prior to the annotation task, the participants had received a training of 20 minutes long.

Categories: event/category-based

Participants were asked to select and de-select emotion labels whenever they *felt* the emotion that they experienced at that moment in the game: in other words, they had to click to select an emotion label and to mark the beginning of the corresponding emotion and click again on the same label to de-select and to mark the ending of that emotional event. The twelve emotion labels from which the participants could choose are based on the 'Big Six' (universal basic) emotions and are supplemented with typical game-related emotions as described in Lazarro [105]. We expected that these labels, shown in Table 6.2, would cover most of the emotions that could occur during gaming. The selection of multiple emotion labels at the same time was allowed, which made it possible to have 'mixed' emotions. The participants also had the option to come up with their own emotion label that was not listed in the alternatives, but it appeared that the participants had not used this option.

Happiness (*Blijdschap*)	Fear (*Angst*)
Boredom (*Verveling*)	Anger (*Boosheid*)
Amusement (*Amusering*)	Relief (*Opluchting*)
Surprise (*Verbazing*)	Frustration (*Frustratie*)
Malicious Delight (*Leedvermaak*)	Wonderment (*Verwondering*)
Excitement (*Opgewondenheid*)	Disgust (*Walging*)

Table 6.2: *The emotion categories used in the category-based annotation task (with the Dutch labels that were offered to the participants in brackets)*.

Continuous emotion dimensions: continuity/dimension-based

The participants were asked to rate their emotions *felt* on two emotion scales namely the Arousal scale (Active vs. Passive) and the Valence scale (Negative vs. Positive). The third scale Dominance was not used in this study. As opposed to the category-based approach where the participants had to mark the beginning and ending of an emotional event, the participants now had to give ratings on emotion scales running from 0 to 100 (with 50 being neutral) each 10 seconds *separately* (thus not *simultaneously* as is done with some annotation tools such as Feeltrace (Cowie et al. [45]). Each 10 seconds, an arrow appeared on the screen to signal the participants to give an Arousal and Valence rating, see Fig. 6.1.

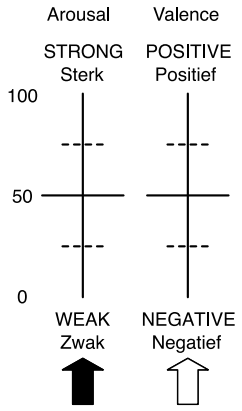


Figure 6.1: *The emotion scales offered to the participants in the dimension-based annotation task.*

6.2.7 Analyses of the 'felt' emotion annotations

The emotion data collected were not (immediately) ready to use for analysis, and required some post-processing as explained here.

Post-processing the emotion annotations

After the annotation tasks were completed by the gamers themselves, we needed to post-process the annotated emotion data which was performed in several steps:

1. first, speech activity was detected and segmented with a relatively simple energy-based silence detection algorithm (available in Praat [23])

2. the speech segments obtained with the silence detection algorithm were transcribed manually at the word level by the current author

3. since delays are possible between the moment one decides to click and the actual emotion event, and since we are only interested in the speech segments, the speech segments needed to be synchronized with the category-based and dimension-based emotion annotations

The silence detection algorithm determined the *units of analysis*, i.e., the *segments* that will be used for further analysis and experiments. The synchronization process between the emotion annotations and the speech segments is explained below.

In the category-based approach, participants marked the beginning and ending of an emotional event. We assumed that the marker of the beginning is more reliable than the ending marker. One of the reasons is that we noticed that some of the emotional events were extremely long; we suspect that participants might have forgotten to de-select the emotion label to mark the ending. Also, we allow for a delay between the real occurrence of an emotional event and the moment that an emotion label was selected. Fig. 6.2 shows how we associated speech segments with emotional events in the category-based annotation task: check for a maximum number of N segments (we chose $N = 5$) prior to the moment that an emotion label was selected whether 1) the segment ends within a margin of T seconds (we chose $T = 3$) before the label was selected, and 2) the segment is labeled as non-silence by the silence detection algorithm.

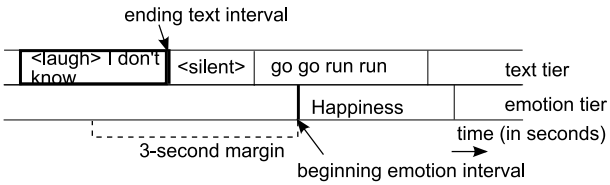


Figure 6.2: *Procedure for finding speech segments that can be associated with an emotional event.*

In the continuous dimension-based approach, a similar synchronization procedure was applied. Each 10 seconds, an arrow appeared to signal the participants to give an Arousal and Valence rating (see Fig. 6.1). We allow for a delay between the moment that the arrow appeared and the moment that participants gave their ratings: for a maximum of N segments (we chose $N = 5$), check whether 1) the segments starts within a margin of T seconds (we chose $T = 3$) from the moment that the arrow appeared, and 2) the segment is labeled as non-silence by the silence detection algorithm.

Results after post-processing

The procedure as described above resulted in a set of speech segments that are labeled with an emotion category label and/or an Arousal and Valence label. In Fig. 6.3, we can observe the frequency of emotion category labels as used by the gamers themselves. It seems that Frustration, Excitement, Happiness, Amusement and Surprise are frequently occurring emotions, while Boredom, Fear and Disgust are hardly experienced by the gamers. Also note that a lot of emotional events co-occurred with silent segments (the white areas in the bars).

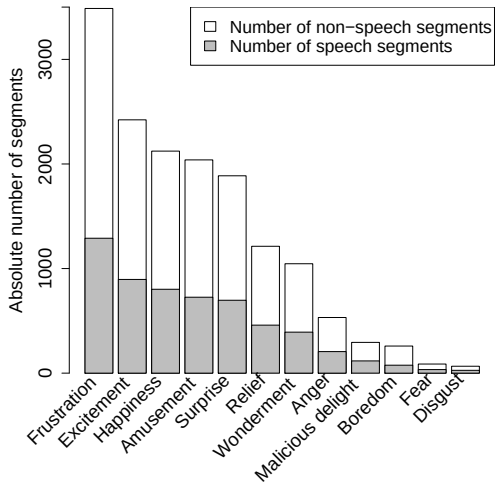


Figure 6.3: *Numbers of speech (and non-speech) segments that could be associated with a category emotional event (total number of category emotion speech segments is 2830).*

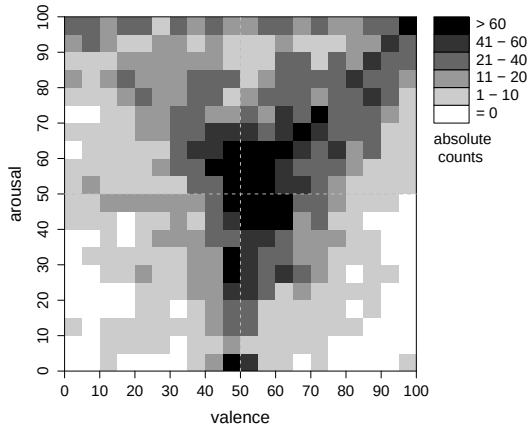


Figure 6.4: *2D Histogram plot: the gamers' Arousal and Valence ratings that could be associated with speech segments,* $N = 7473$.

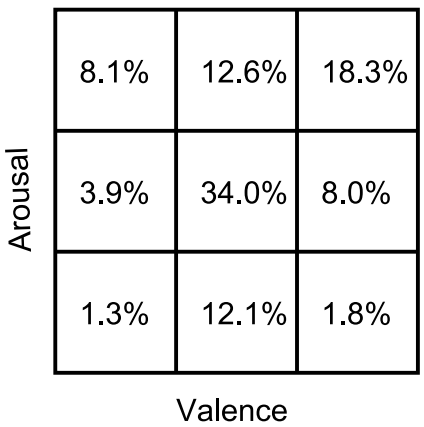


Figure 6.5: *The gamers' Arousal and Valence ratings that could be associated with speech segments, expressed in percentages.*

The results of the dimension-based annotation task are presented in Fig. 6.4 and 6.5. The figures show that the majority of speech segments is annotated as Neutral. The Positive-Active area is also relatively well-filled with speech segments, followed by the Negative-Active area in the Arousal-Valence space. There are apparent blank spots in the Positive-Passive and Negative-Passive areas. It appears that the participants did not often feel very Positive or Negative in a Passive way which is imaginable. This relation between Valence and Arousal has also been encountered in previous emotion rating studies, Lang [102], Hanjalic and Xu [74], where a similar 'boomerang'-shape was found in the Arousal-Valence space when subjects were asked to rate certain stimuli along Arousal and Valence scales. Finally, the participants mentioned that they sometimes had trouble interpreting the Arousal scale: they had some trouble rating something as Passive or Neutral.

	Number of speech segments	Total length of speech segments in minutes (mean and standard deviation in seconds)	Number of unique words
Category-based	2830	78.6 m (1.67 s, 1.26 s)	1322
Dimension-based	7473	186.2 m (1.50 s, 1.12 s)	1963

Table 6.3: *Amount of emotionally labeled speech data according to the gamers' emotion labeling.*

In summary, this gaming experiment resulted in a substantial amount of labeled speech data (see Table 6.3) that can be used for the training and development of automatic emotion recognizers. Due to the spontaneous character of this gaming experiment, we have obtained emotional speech data that do not always contain extreme emotions, and we do not have an equally well-balanced dataset in the sense that not all areas in the Arousal-Valence space are uniformly covered with speech segments. One important novelty of the data collected in this gaming experiment, is the fact that all data is annotated by the gamers themselves. We will refer to these annotations as SELF-annotations (or SELF-ratings). The participants (i.e., the gamers) who have labeled their own *felt* emotions after playing the videogame are referred to as the SELF-raters. In subsequent experiments, we investigated the relation between these SELF-annotations and annotations made by other observers. We also used the data to train and test emotion recognizers. The database collected and described here will be referred to as the TNO-GAMING corpus. In Fig. 6.6 and Table 6.4 some examples of emotional expressions are shown that were captured while the subjects were playing the videogame.

6.3 Experiment I: 'felt' and 'observed' emotions in unimodal and multimodal conditions

In this Section, we describe our first experiment carried out with TNO-GAMING data. In this perception experiment[1], movie clips from the TNO-GAMING database were used

[1]Part of the work described in this Section was previously published in Truong et al. [191].

Figure 6.6: *Some video stills of emotional expressions observed in the* TNO-GAMING *corpus.*

	Val	Aro	Transcription
NA	0	99	*urgh wat zijn dat voor monsters?* 'urgh what are those type of monsters?'
	1	100	*nee jaa klote klote! no! rennen! no no!* 'no yes shit shit! no! run! no! no!'
	3	80	*zoo irritant* 'soo irritating'
NP	10	31	*ja ik ik probeer daar heen te gaan* 'yes I I try to go there'
	13	5	*ik wil die klotewapens niet hebben* 'I don't what those shit weapons'
	13	33	*na ik zie helemaal niks* 'na I don't see anything'
PA	97	99	*oh dat ben jij sorry [lach]* 'oh that's you sorry [laughter]'
	81	98	*rennen rennen rennen ja goed zo* 'run run run yes good job'
	97	83	*maak een punt maak een punt [lach]* 'score a point score a point [laughter]'
PP	71	8	*oke dan gaan we nu een punt scoren* 'OK now we are going to score a point'
	77	18	*we gaan voor de twintig he ik heb de blauwe vlag* 'we are going for the twenty right I have the blue flag'
	74	29	*ik heb ze ik maak ze dood loop maar* 'I have them I kill them just walk'

Table 6.4: *Examples of gamers' word transcriptions and emotion annotations (with an English translation), Val=Valence, Aro=Arousal, NA=Negative Active, NP=Negative Passive, PA=Positive Active, PP=Positive Passive*

and provided to naive observers in several different unimodal and multimodal listening conditions. The task of the observers was to rate these movie clips on Arousal and Valence scales, similar to what the gamers themselves had done. In this way, we could compare the SELF-ratings to the observers' ratings. Additionally, we analyzed the emotion ratings obtained from the several unimodal and multimodal conditions to see what type of information leads to higher inter-rater observer agreement.

6.3.1 Related work

In areas of research where human labelers are used for data annotation, the quality of the annotation can be assessed through an agreement analysis (aka reliability anal-

ysis). The usual procedure involves a number of annotators (raters) who annotate overlapping pieces of data such that inter-rater agreement can be assessed: the more raters mutually agree with each other, the higher the quality (and reliability) of the annotation. The annotation of natural emotion is known to be a complex and difficult process, e.g., Laskowski and Burger [103], Reidsma et al. [146], Douglas-Cowie et al. [54]. Laskowski and Burger [103] proposed an annotation scheme that describes how people are *behaving* rather than how they are *feeling*. They reported inter-rater κ agreements (Cohen [41], this is Cohen's Kappa which ranges from 0 meaning no agreement to 1 meaning perfect agreement) between 0.15 and 0.67 for three annotators who annotated Valence in meeting speech data. Similarly, Reidsma et al. [146] proposed an annotation procedure that is more attuned to a behavioral description of emotion. In a first trial that was performed with Feeltrace (Cowie et al. [45]), carried out on spontaneous AMI audiovisual meeting data (Carletta [35]), they obtained relatively low averaged pair-wise agreement figures between 0.07 and 0.18. Douglas-Cowie et al. [54] performed an agreement analysis on a natural multimodal emotion database, the EmoTV database (Douglas-Cowie et al. [54]), and report inter-rater agreement figures κ between 0.37 and 0.54, achieved with a category-based annotation. More precisely, κ was 0.37, 0.43, and 0.54 when audiovisual, video only and audio only information respectively was provided. Surprisingly, agreement was lowest in the audiovisual condition. Busso and Narayanan [26] compared emotion assessments of 'self' versus 'other'. They found that in a category-based labeling approach, annotations made by the subjects themselves judging their own emotions ('self') differed from the ones made by observers ('other'). In a continuous-based labeling approach (labeling on Valence, Activation and Dominance scales), they found no differences between 'self' and 'other'.

In the current study, we assessed human emotion judgments under audio-only, video-only, audiovisual, and audiovisual plus context information conditions. In addition, we compared 'self' vs. 'other' emotion assessments (Experiment I) and evaluated their usefulness for automatic affect recognition (Experiment II).

6.3.2 Defining the goals of Experiment I

The goal of an agreement (reliability) analysis is to assess how reliable the labels given by the annotators are. 'Reliable' can be defined in terms of level of agreement: if many people agree upon a label, this label can be considered 'reliable'. In our study, we defined two groups of annotators and their corresponding annotations: we compared SELF-annotations, i.e., 'felt' emotion ratings of the gamers themselves, to OTHER-annotations coming from 'other' people, i.e., perceived emotion ratings from external (naive) observers.

With the data collected, we performed a perception experiment. In Experiment I, we are interested in two aspects: we aim to assess

- how well observers agree on the perception of spontaneous emotions when audio only, video only, audiovisual or audiovisual plus context information is provided
- the reliability of SELF-ratings of emotion in comparison with OTHER-ratings of

emotion.

6.3.3 Participants: observers

Twelve female and six male participants with an average age of 21.9 years were asked to participate in a small perception experiment. These 18 people had not participated in the previous gaming experiment. We will refer to this group of 18 participants as the observers who produced the OTHER-ratings. The gamers who played the video game and who rated their own emotions are referred to as SELF-raters.

6.3.4 Experimental setup

For the selection of the stimuli, movie clips from six gamers were selected by a number of criteria: the move clips had to contain a sufficient amount of vocal and facial expressions, and the aim was to have movie clips originating from different regions in the Arousal-Valence space. These regions are the four well- known quadrants: Positive-Active (PA), Negative-Active (NA), Positive-Passive (PP), and Negative-Passive (NP). In addition, we selected movie clips that have a large emotion change in Arousal (CA) and a large change in Valence (CV). This makes a total of 6×6 movie clips that were presented to each observer. However, it appeared to be difficult to satisfy all of these criteria. As a result, not all emotion quadrants were equally well represented in the set of stimuli offered to the observers, see Fig. 6.7. Each movie clip has a length of 55 seconds and 4 rating moments. At each rating moment, an arrow appeared to signal the observer to give an Arousal and Valence rating, similar to the rating task that was performed by the gamers themselves.

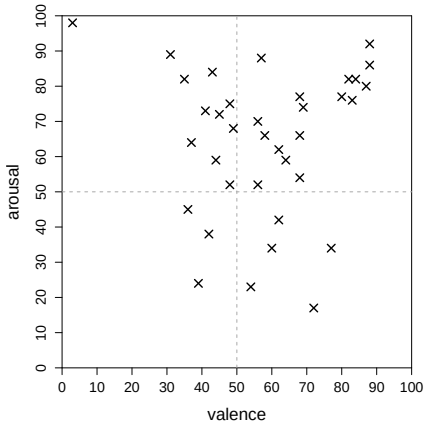


Figure 6.7: *The averaged locations of the 36 movie clips that were offered to observers in Experiment I.*

The movie clips were presented to the observers in six different conditions: audio only (A), video only (V), audiovisual (AV), audio+context (AC), visual+ context (VC), and audiovisual+context (AVC). With 'context', we mean the game content video stream that was recorded during game play. The AVC condition is best comparable to the gamers' rating task in which the SELF-annotations were collected (the

gamers too had the audiovisual and the context information available during annotation).

In a within-subject design, the 36 movie clips were distributed over 36 cells in a 6 (conditions) by 6 ('emotion regions') matrix and presented to the observers in a balanced design such that each movie clip of a specific gamer with a specific 'emotion region' was rated in each condition by at least two observers (see Table 6.5 for one example design of one observer).

	'emotion region'					
cond	PA	NA	CA	CV	PP	NP
A	gamer1	gamer2	gamer3	gamer4	gamer5	gamer6
V	gamer1	gamer5	gamer4	gamer2	gamer3	gamer6
AV	gamer4	gamer2	gamer3	gamer1	gamer5	gamer6
AC	gamer2	gamer6	gamer3	gamer5	gamer1	gamer4
VC	gamer1	gamer6	gamer2	gamer3	gamer5	gamer4
AVC	gamer3	gamer6	gamer5	gamer2	gamer4	gamer1

Table 6.5: *Example of distribution of movie clips over conditions and 'emotion regions' for one observer.*

6.3.5 Agreement computations: Krippendorff's α

We used Krippendorff's α as agreement measure. Although Cohen's κ is frequently used by researchers, it is not as flexible and generic as α as explained below. For example, κ can only be calculated between 2 raters (for more than 2 raters, Fleiss' κ can be used), and cannot deal with missing data values.

Krippendorff's Alpha

For all agreement computations, Krippendorff's α (Krippendorff [98, 97], Hayes and Krippendorff [76]) was used. It was proposed in Hayes and Krippendorff [76] as the standard reliability measure. According to Hayes and Krippendorff [76], an index of reliability should have the following properties:

1. It should assess the agreement between two or more observers who describe each of the units of analysis separately from each other. For more than two observers, this measure should be a) independent of the number of observers employed, and b) invariant to the permutation and selective participation of observers. Under these two conditions, agreement would not be biased by the individual identities and number of observers who happen to generate the data.

2. The index should not be confounded by the number of categories or scale points made available for coding.

3. The index should constitute a numerical scale between at least two points with sensible reliability interpretations. By convention, perfect agreement is set to 1.00. The absence of agreement is typically set at 0.00 and should represent a

situation in which the units of analysis bear no statistical relation to how they end up being identified, coded, or described.

4. It should be appropriate to the level of measurement of the data.

5. Its sampling behavior should be known or at least computable.

Krippendorff's α satisfies all of the conditions discussed above, and hence, is the measure preferred. α counts pairs of categories or scale points that observers have assigned to individual units. It is defined on a scale from -1 to 1 where 1 means perfect reliability, 0 means absence of reliability, and a negative α means disagreement. α can measure agreement for nominal, ordinal, interval and ratio data. Furthermore, it can deal with data that contain missing values.

Krippendorff's α (see Krippendorff [98] for a step-by-step description of its computation) is generally computed as:

$$\alpha = 1 - \frac{\mathrm{D_0}}{\mathrm{D_e}} = 1 - \frac{\text{observed disagreement}}{\text{expected disagreement}} \tag{6.1}$$

The general form of the observed disagreement D_o is:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck \ metric} \delta^2_{ck} \tag{6.2}$$

The disagreement that one would expect when the coding of units is attributable to chance, D_e, can be computed as:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c \times n_{k \ metric} \delta^2_{ck} \tag{6.3}$$

o is a so-called *coincidence* matrix that can be computed from a *reliability* matrix. The reliability data matrix is an $m \times r$ matrix, filled with the judgments from m observers for r units. From this matrix, a coincidence matrix o can be constructed, see Fig. 6.8. n_c and n_k are the numbers of $c-k$ pairs for c and k respectively.

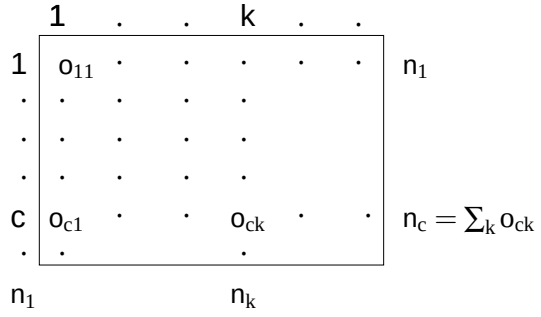


Figure 6.8: *A coincidence matrix.*

In the coincidence matrix in Fig. 6.8, $\mathrm{o_{ck}}$ is computed as $\mathrm{o_{ck}} = \sum_{\mathrm{u}} \frac{\text{Number of c - k pairs in unit u}}{\mathrm{m_u - 1}}$, where $\mathrm{m_u}$ is the number of judgments given for unit u. δ^2_{ck} is a difference function

that depends on the metric of the data, i.e., nominal, ordinal, interval or ratio. Our data is ordinally scaled so we can use the ordinal difference function that is defined as:

$$_{ordinal}\delta^2_{ck} = \frac{n_c}{2} + \sum_{g>c}^{g<k} n_g + \frac{n_k}{2}, \tag{6.4}$$

where $c < k$, and g is the rank running between c and k.
Finally, Krippendorff's α can be computed as follows:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\frac{\sum_c \sum_{k>c} o_{ck}\delta^2_{ck}}{\sum_c \sum_{k>c} n_c \times n_k \delta^2_{ck}} \tag{6.5}$$

Procedure

We used Krippendorff's Alpha α [98] on an ordinal scale to assess the agreement between multiple ($>= 2$) raters. For each emotion dimension, there are 144 ratings (36 movie clips with each 4 ratings). We chose to have a within-subjects design that is balanced but incomplete in the sense that not each movie clip is rated by all observers. Each movie clip is rated by at least two observers. In assessing the reliability of content data where multiple raters are used to annotate the data, it is not uncommon that raters code different subsamples of the data. Krippendorff's α is flexible enough that it can deal with N raters $>= 2$, and it can accommodate for 'missing values'. Prior to calculating α, all ratings were discretized into 5 classes with boundaries on 20, 40, 60, and 80. These 'raw' ratings will be referred to as RAW-ratings.

Furthermore, we also computed so-called 'delta' ratings (referred to as DELTA-ratings), i.e., *changes* between subsequent emotion ratings, to evaluate whether people judge emotion better in a *relative* manner than an *absolute* manner. These DELTA ratings were computed by subtracting the previous rating from the current rating in each movie clip, see Fig. 6.9.

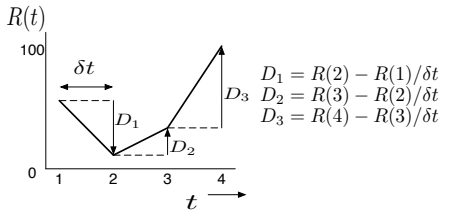


Figure 6.9: *Computation of* DELTA*-ratings for each movie clip, in this case δt is 1 because each rating $R(t)$ is given at a fixed interval δt ($R(t)$ is an emotion rating given at moment t.).*

Finally, to adjust for personal differences between observers and personal differences of the gamers, e.g., some observers or gamers tend to use the whole scale while others only use a small part of the scale, we linearly re-scaled all the Arousal and Valence ratings such that each person has a minimum and maximum of 0 and 1 respectively. These scaled ratings will be referred to as SCALED-ratings.

6.3.6 Results: inter-observer agreement in unimodal and multimodal conditions

The movie clips were presented to the observers under various unimodal and multimodal conditions; we report results obtained in the A (audio-only), V (video-only), AV (audiovisual), and AVC (audiovisual+context) conditions. The results are presented in Fig. 6.10 and Fig. 6.11. The inter-observer agreement figures based on RAW ratings range from 0.12 in the audio only condition to 0.48 in the audiovisual condition, see Fig. 6.10. For both Arousal and Valence, the highest αs are obtained in the AV condition when the ratings are RAW. Apparently, observers do benefit from the multimodal information that is made available to them, although the addition of context does not seem to help, at least not in the RAW case. The visual channel seems to provide more information than the acoustic channel. Furthermore, the inter-observer agreement on the Arousal scale is systematically worse than on the Valence scale in the RAW case. However, in the DELTA case, α has increased considerably for Arousal, but not for Valence: this suggests that people are better able to judge *changes* in Arousal rather than absolute Arousal.

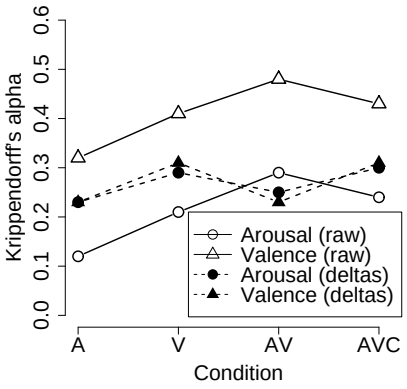


Figure 6.10: *Krippendorff's α inter-observer agreement:* RAW*-ratings and* DELTA*-ratings.*

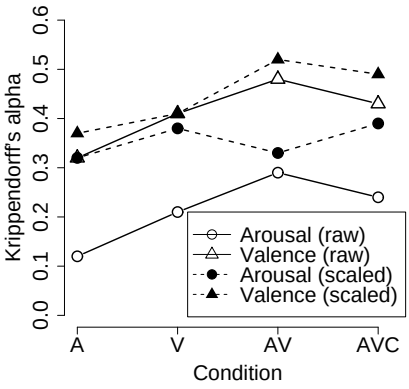


Figure 6.11: *Krippendorff's α inter-observer agreement:* RAW*-ratings and* SCALED*-ratings.*

In Fig 6.11, we can observe that when we linearly scale all the ratings to $[0, 1]$, we obtain a substantial improvement of α for Arousal. For Valence, this improvement is small. The αs for the scaled ratings range from 0.32 to 0.52. Finally, similar to the raw case, Arousal is less agreed upon than Valence, multimodal information is (usually) beneficial, and visual information is stronger than acoustic information.

6.3.7 Results: agreement between SELF-ratings and OTHER-ratings

One way to assess how SELF-ratings compare to OTHER-ratings, is to add the SELF-rater's ratings to the group of the OTHER-raters' ratings, calculate inter-rater agreement, and compare this outcome with the α computed *without* the SELF-rater. If α does not decrease, it indicates that the SELF-rater did not influence the inter-rater agreement negatively. If this is the case, then this could imply that the observers

agree equally well with the SELF-rater, and that observers have the ability to perceive 'felt' emotion. However, in Table 6.6, we can observe that the addition of the SELF-rater to the OTHER-raters' ratings affects α negatively. This is an indication that there is a discrepancy between SELF-ratings and OTHER-ratings.

	RAW-ratings		DELTA-ratings		SCALED-ratings	
		+SELF		+SELF		+SELF
Arousal	0.24	0.23	0.30	0.21	0.39	0.36
Valence	0.43	0.37	0.31	0.29	0.49	0.40

Table 6.6: *Krippendorff's α inter-rater agreement between 3 (=OTHER) or 4 annotators: either without or with the SELF-rater (+SELF), for the AVC condition.*

Another way to assess the reliability of SELF-ratings is to compute pair-wise agreements between an observer and a SELF-rater. The averaged, minimum and maximum agreement α between an observer and a SELF-rater are shown in Table 6.7. In the raw case, the averaged pair-wise α is 0.16 and -0.09 for Arousal and Valence respectively which indicates very low agreement between the SELF-raters and the OTHER-raters. In the SCALED case, the averaged pair-wise agreement is improved. The large differences between minimum and maximum pair-wise agreements indicate that there are (large) differences between the observers: some observers do agree with the SELF-raters while others disagree.

	Arousal			Valence		
	mean	min	max	mean	min	max
RAW	0.16	-0.27	0.51	-0.09	-0.45	0.34
DELTA	0.13	-0.37	0.48	0.24	-0.29	0.69
SCALED	0.34	-0.07	0.70	0.30	-0.21	0.62

Table 6.7: *Krippendorff's α for pair-wise agreement between an observer and the SELF-raters in the AVC condition.*

6.3.8 Conclusions

With this perception experiment, we have investigated to which degree observers agree on the assessment of spontaneous emotions that were shown in audio-only, video-only, audiovisual and audiovisual+context conditions, and we have compared the reliability of SELF-judgments of emotions to those of observers. With αs ranging from 0.32 to 0.52 (after scaling the ratings), we achieved agreement figures that are in line with results from other studies, see Laskowski and Burger [103], Reidsma et al. [146], Douglas-Cowie et al. [54]. We found that agreement is consistently higher on the Valence scale than on the Arousal scale. Improvements on the Arousal scale can be achieved when the *relative changes* are taken into account instead of the absolute values. This does not seem to apply for the Valence scale. In general, agreement was higher in the multimodal conditions AV and AVC, than in the unimodal conditions A and V. These results are somewhat different from what was found in Douglas-Cowie

et al. [54] who found in their study that agreement was lowest in the multimodal audiovisual condition. However, it should be noted that their study was based on TV clips and a category-based annotation method. Furthermore, visual-only information was usually stronger than audio-only information. Finally, adding context information to audiovisual information did not always result in higher agreement.

The reliability of the SELF-ratings were assessed by computing agreement when the SELF-rater was added to the OTHER-ratings, and by computing pair-wise agreements between the SELF-ratings and individual observers' ratings. We found indications that SELF-ratings and OTHER-ratings differ, sometimes substantially, from each other disadvantageously. We conclude this based on our observations that the inter-rater agreement was lowered when the SELF-rater was added to the OTHER-ratings. In addition, the pair-wise agreement between the SELF-raters and the OTHER-raters were relatively low.

These findings indicate that the assessment of spontaneous emotion involves a complex multimodal interpretation process that is not quite well described yet given the mixed findings of several studies. Furthermore, emotion annotation and labeling are not straightforward processes, and the eventual goal, namely to develop an affect recognition system, depends much on how the labeled emotion data is structured and annotated, and by *whom* the annotations have been carried out. An affect recognition system developed with SELF-labeled data most likely will learn to recognize 'felt' emotions, while an affect recognition system that is developed with OTHER-labeled data will learn how to recognize 'expressed' or 'perceived' emotions. In the following experiments, we compared the use of SELF-rated and OTHER-rated data for the development of an automatic affect recognition systems that aim to predict Arousal and Valence values in the Arousal-Valence space.

6.4 Experiment II: speech-based emotion prediction in the Arousal-Valence space

In this Section, we describe Experiment II carried out with the TNO-GAMING corpus. We describe two different speech-based affect recognition systems: one that was trained to detect 'felt' emotions, and one that was trained to detect 'perceived' emotions. The main characteristics of these systems are that they are trained to predict scalar values on Arousal and Valence scales, rather than to classify emotions in categories. We discuss the results of the detection experiments and its implications. Subsequently, the performances of the machines are compared to human performance. Furthermore, we show what type of features are most predictive of the Arousal and Valence scale.

6.4.1 Related work

The majority of speech-based emotion recognition systems reported in the emotion literature are trained to classify emotion categories, while few have adopted the Arousal-Valence space to predict scalar values on Arousal and Valence scales. Recently, Grimm et al. [69, 68] have presented methods to predict scalar values on

Valence, Activation (Arousal), and Dominance scales. In Grimm et al. [69], the VAM corpus was used as speech material, which contains data from a German TV talk-show in which several guests talk about personal issues. The emotion annotation of this database was based on the Self-Assessment-Manikin method (see Lang [101] and Section 3.1). The regression method Support Vector Regression (SVR) was used to train the continuity-based emotion prediction model. This method was compared to Fuzzy k-Nearest Neighbor and Rule-based Fuzzy Logic. The SVR performed best with an average error (which was defined as the absolute difference between reference and prediction) of 0.13, 0.15, and 0.14 for Valence, Activation, and Dominance respectively. In Grimm et al. [68], more extensive estimation experiments were performed with Rule-based Fuzzy Logic. The averaged errors obtained varied from 0.17 to 0.28 for the VAM corpus.

Our current work differs from theirs in that we not only use observers' perceived annotations (as is commonly done in the majority of studies), but we also use the gamers' 'felt' emotion annotations as reference to predict 'felt' emotion. Further, in addition to acoustic features, we also used lexical features to model affect. Finally, machine performance was compared to human performance in terms of agreement: to what degree do machines agree with human annotators?

6.4.2 Defining the goals of Experiment II

In Experiment I, we found that there are differences between SELF-annotations made by the gamers themselves, and the OTHER-annotations made by observers. In Experiment II, both types of annotations for the development of emotion recognizers were used. Rather than detecting categories of emotions, the task of the emotion recognizers is to predict scalar values on Arousal and Valence scales which run from $[-1, 1]$. Rather than referring to 'classification' or 'detection' experiments and 'classifiers' or 'detectors' we prefer to use terms as 'prediction experiments' or 'emotion predictors' to emphasize the fact that we are not working with emotion categories here. For the development of our emotion predictors, we used acoustic and lexical information. The expectation is that Arousal is better modeled with acoustic information and that Valence is better modeled with lexical information in spontaneous emotional speech. Also, since we are most interested in the use of acoustic information, we investigate what types of acoustic features are most predictive of which emotion dimension in spontaneous emotional speech. Finally, the results obtained with the machines were compared to human performance. The research questions of Experiment II can be summarized as follows:

- What are the effects of the use of 'felt' emotion annotations versus 'perceived' emotion annotations in the development of automatic emotion predictors?
- What types of features can best be used to model what emotion dimension?
- How does machine performance compare to human performance in predicting Arousal and Valence in spontaneous emotional speech?

6.4.3 Material

Here, we describe the speech material used in the emotion prediction experiments. From the TNO-GAMING corpus, we selected speech material to be re-annotated by a group of observers (which was not the same group of observers that participated in Experiment I) such that a part of the corpus was annotated with both 'felt' and 'perceived' affect.

SELF-*annotation*

As explained above, part of the TNO-GAMING corpus was annotated by both the gamers themselves *and* naive observers. Because the number of segments of the whole corpus is relatively large, we decided to make a selection of 2400 segments, out of the original set of 7473 segments, that was offered to a group of naive observers. The selection procedure of these 2400 movie clips that were offered to the observers was partly randomized, partly restricted by our criterion to roughly maintain the same proportions of the segments in the Arousal-Valence space of the original set (see Fig. 6.4 and 6.5), and partly driven by the need for a larger number of segments in the lower Arousal area to adjust for this strongly imbalanced distribution on the Arousal scale. The distribution of the segments selected for re-annotation in the Arousal-Valence space is displayed in Fig. 6.12 and 6.18. The total length of the whole set of 2400 segments is approximately 76 minutes. The mean duration and standard deviation of a segment is 1.9 and 1.2 seconds respectively. The scales of the Arousal and Valence dimensions are linearly re-scaled from [0,100] to a range of [-1,1] which allows for comparison with previous studies (the linear re-scaling will not affect the analyses or results), e.g., Grimm et al. [69].

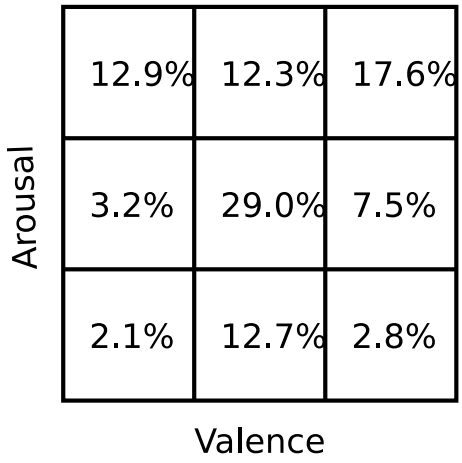


Figure 6.12: *The distribution of the 2400 selected speech segments in the Arousal-Valence space based on the* SELF-*annotations, expressed in percentages.*

Histograms of the Arousal and Valence ratings selected for re-annotation (based on the SELF-annotations) are displayed in Fig. 6.13 and 6.14. We can observe that on the Arousal dimension, there is a relative scarcity of low Aroused speech segments, which could be a consequence of our emotion elicitation method.

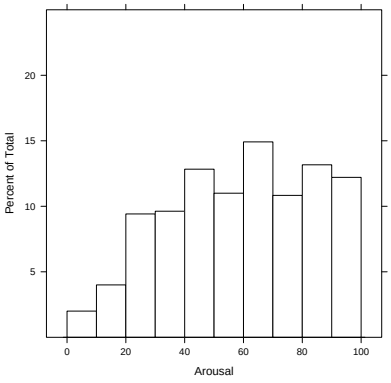


Figure 6.13: *Histogram of the Arousal* SELF*-ratings of the 2400 selected speech segments.*

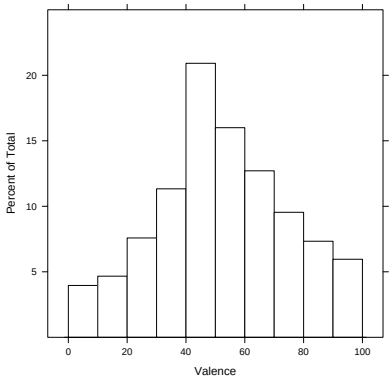


Figure 6.14: *Histogram of the Valence* SELF*-ratings of the 2400 selected speech segments.*

Re-annotation by a group of observers

The set of 2400 emotion segments were (audiovisually) presented to six annotators who were not involved in Experiment I. The six annotators have an average age of 25.4 years. Similar to Experiment I and the SELF-rating procedure, the annotators were asked to rate each audiovisual segment on the Arousal and Valence scale that run from 0 to 100, with 50 being Neutral (afterwards we re-scaled to [-1,1]). The differences with the SELF-annotation procedure are that 1) the audiovisual segments are already segmented, 2) the annotators now can re-play the segment if they like, and 3) no context information was given. We will refer to the emotion annotations of the six annotators as OTHER.3 ('3' because each segment is annotated by 3 different annotators, this will be explained below), so note that these OTHER.3-ratings are different from the OTHER-ratings used in Experiment I.

Each observer (TH, PI, CO, RA, FR, and AT) annotated different parts (A, B, C, and D) of the dataset that overlapped with parts that were annotated by other observers (see Fig. 6.15). This ensured that each segment was annotated by three different annotators (in order to obtain more reliable reference annotations that can be used for the emotion prediction experiments). The dataset was divided into four parts, each part consisting of 624 segments. Each observer was assigned to two parts of the database, and thus annotated in total 2×624 segments, see Fig. 6.15. Of the 624 segments in each part, 24 segments occurred twice and were used to assess the rating consistency of the observer (intra-rater agreement) him/herself. For each observer, it took approximately 4 to 5 hours to complete the annotation of all 1248 segments, including breaks. That means that the annotation was carried out at a rate of approximately 6 times real-time.

The OTHER.3 annotations represent the annotations of 3 separate annotators (a 3×2400-matrix). In order to have a reference annotation of the OTHER.3-ratings that can be used for the prediction experiments, the ratings of the three annotators were averaged for each segment. We will refer to these ratings as OTHER.AVG: 'AVG'

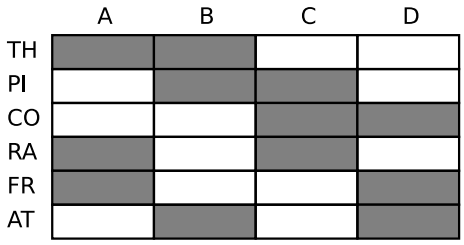


Figure 6.15: *Division of dataset into several overlapping parts, each observer annotated two cells (each cell contains 624 segments) such that each segment is annotated by 3 different observers.*

stands for 'averaged'. These OTHER.AVG-annotations (a 1×2400-matrix) were used as reference annotations in the prediction experiments. The histograms in Fig. 6.16 and Fig. 6.17 show the distributions of the averaged ratings OTHER.AVG: it seems that the majority of the segments were judged, on average, as Neutral (or in Neutral's vicinity) by the observers, more than the SELF-raters have done. The differences between the SELF-annotators and the OTHER.AVG-annotators become even more clear when we compare the 2-dimensional histograms based on SELF-annotations and OTHER.AVG-annotations, shown in Fig. 6.18 and Fig. 6.19. The SELF-annotators appear to have selected more extreme values for their own felt emotions than the observers who seemingly did not perceive these emotions as such and who mostly selected values in the vicinity of Neutrality. In addition, the pull towards Neutrality is also be caused by averaging the annotations.

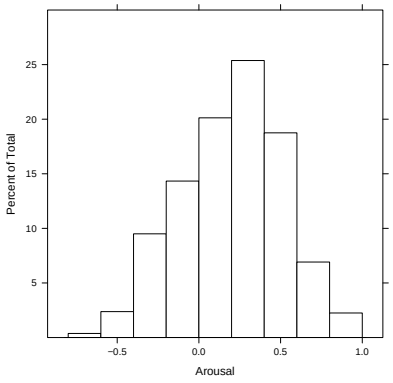


Figure 6.16: *Histogram of the Arousal* OTHER.AVG*-ratings for the 2400 selected speech segments.*

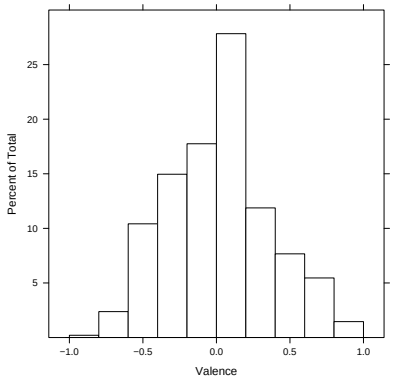


Figure 6.17: *Histogram of the Valence* OTHER.AVG*-ratings for the 2400 selected speech segments.*

6.4.4 *Reliability of* SELF*-annotations,* OTHER.3*-annotations and* OTHER.AVG*-annotations*

In this Section, we describe how the SELF-annotations, OTHER.3-annotations and OTHER.AVG-annotations were compared to find out how these types of annotations

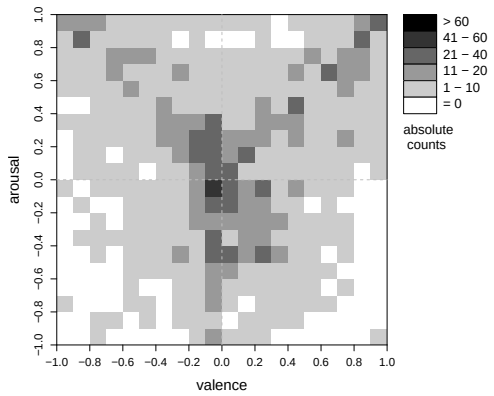


Figure 6.18: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the* SELF*-ratings.*

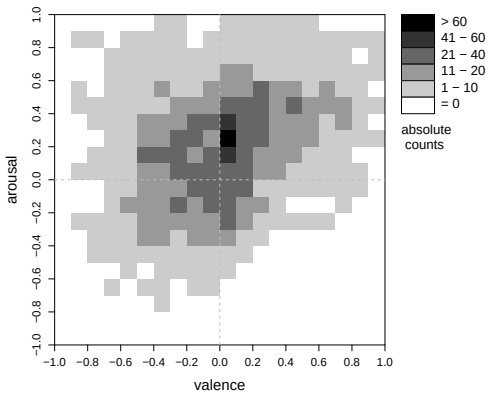


Figure 6.19: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the* OTHER.AVG*-ratings.*

relate to each other.

Similar to Experiment I, we used Krippendorff's α (Krippendorff [98]) (ordinal) to calculate intra-rater and inter-rater agreement figures. Surprisingly, scaling or normalizing (normalizing to $\mu = 0$ and $\sigma = 1$) all ratings increased agreement minimally (1%–4%) or even decreased agreement, so we decided to use the RAW-ratings. We also report the correlation coefficient Pearson's ρ to allow for comparison with other studies; ρ runs in the range of [-1,1]. In addition, if the number of observers is 2, Kappa κ (equal weights) is reported as well. For the computation of α and κ, the ratings were discretized into 5 classes with boundaries at -0.6, -0.2, 0.2, and 0.6.

	Krippendorff's $\alpha_{\mathrm{ord},5}$		Pearson's ρ		$\kappa_{\mathrm{equal},5}$	
Rater	Valence	Arousal	Valence	Arousal	Valence	Arousal
TH	0.90	0.44	0.91	0.55	0.79	0.35
PI	0.78	0.42	0.84	0.46	0.60	0.27
CO	0.81	0.52	0.87	0.64	0.71	0.42
RA	0.78	0.45	0.84	0.56	0.62	0.33
FR	0.66	0.55	0.73	0.54	0.53	0.36
AT	0.85	0.49	0.91	0.64	0.71	0.37
mean	0.80	0.48	0.85	0.57	0.66	0.35

Table 6.8: *Intra-rater agreement, based on 48 double-rated segments.*

In order to assess the rating consistency of individual annotators, we inserted 2×24 segments that were rated twice by the annotators. The intra-rater agreement figures of each annotator are presented in Table 6.8. Similar to the findings in Experiment I, we found that Valence is easier to rate than Arousal when audiovisual segments are provided. In general, the annotators were relatively consistent: α ranges from 0.66 to 0.90, and from 0.42 to 0.55 for Valence and Arousal respectively. Given

these relatively good intra-rater agreement figures, we considered the annotators reliable and hence, none of the annotators were replaced.

The results for the inter-rater agreement among the observers analyses are shown in Table 6.9, Table 6.10, and 6.11. Firstly, we focused on the OTHER.3-annotations: Table 6.9 shows how the inter-observer agreement is affected among the 3 or 4 annotators when the SELF-rater is added. Table 6.9 presents the alphas obtained with different discretization steps (3, 4, 5 or 10) on ordinal and nominal scales. Since we work with ordinal Arousal and Valence scales, we can adopt an ordinal α computation, and we continue to base α on a discretization of 5 classes (similar to Experiment I). Similar to the results in Experiment I, see Table 6.6, we observe in Table 6.9 that when the SELF-rater is added to the group of OTHER.3-raters, α_5 decreases slightly for Arousal (-0.03) and more substantially for Valence (-0.13). In order to have a feeling for the range of α when an annotator is added, we show in Table 6.10 the behavior of α when an annotator is added who perfectly agrees or disagrees with one of the 3 annotators. These figures suggest that the SELF-rater indeed slightly disagrees with the three annotators: in any case, the SELF-rater does not contribute to *more* agreement among the annotators.

		OTHER.3							
	Dim.	α_3		α_4		α_5		α_{10}	
			+SELF		+SELF		+SELF		+SELF
ordinal	Arousal	0.23	0.20	0.25	0.23	**0.28**	**0.25**	0.30	0.26
	Valence	0.45	0.36	0.50	0.39	**0.57**	**0.44**	0.58	0.45
nominal	Arousal	0.12	0.11	0.09	0.08	0.09	0.08	0.04	0.04
	Valence	0.31	0.25	0.23	0.18	0.27	0.19	0.11	0.09

Table 6.9: *Human inter-rater agreement among* OTHER.3*-raters, without and with the* SELF*-raters, for several discretization steps (3, 4, 5, and 10), and on an ordinal or nominal scale.*

	OTHER.3	OTHER.3 + 1 perfect agreement	OTHER.3 + 1 perfect *dis*agreement	OTHER.3 + SELF
Arousal	0.28	0.37/0.42/0.41	-0.11/-0.14/-0.18	0.25
Valence	0.57	0.64/0.63/0.65	-0.14/-0.15/-0.15	0.44

Table 6.10: *Behavior of $\alpha_{\mathrm{ord},5}$ (ordinal α with 5 discretization steps) when 1 annotator is added who perfectly agrees or disagrees with 1 of the 3 annotators.*

Secondly, we focused on the OTHER.AVG-annotations that were derived from the OTHER.3-annotations, by taking the average over the 3 observers' ratings (per segment). Consequently, the inter-rater agreement between the SELF-ratings and OTHER.AVG-ratings is relatively low as can be seen in Table 6.11. The plots in Fig. 6.20 indeed show that there is low correlation between the SELF-ratings and OTHER.AVG-ratings. These observations confirm our findings of Experiment I, namely that there is a discrepancy between self-judged and observed (perceived) emotions. In the following

Sections, we investigated what the effects are of using self-judged emotion annotations (=SELF) or observed emotion annotations (=OTHER.AVG) for the development of speech-based emotion recognizers.

	Krippendorff's $\alpha_{\mathrm{ord},5}$	Kappa $\kappa_{\mathrm{equal},5}$	Pearson's ρ
Arousal	0.27	0.16	0.33
Valence	0.36	0.25	0.41

Table 6.11: *Inter-rater agreement between the* SELF*-ratings and the* OTHER.AVG*-ratings.*

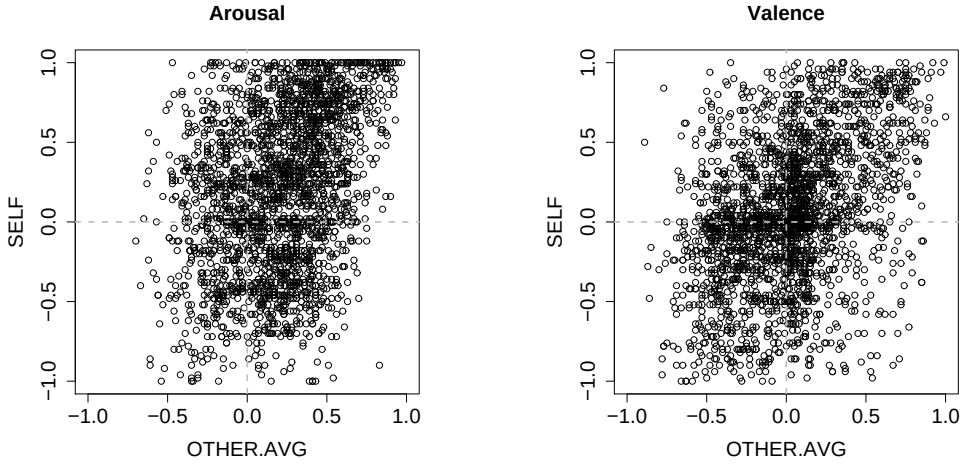


Figure 6.20: OTHER.AVG*-annotations plotted against* SELF*-annotations, left=Arousal, right=Valence.*

6.4.5 Features and Method

In this Section, we describe the speech features and methods used to model affect in terms of scalar Arousal and Valence values. Support Vector Regression was used as a learning method in combination with acoustic *and* lexical features.

Method: Support Vector Regression

Since our goal is to predict real-valued output rather than discrete classes, we used a learning algorithm based on regression. Support Vector Regression (SVR, see Smola and Schölkopf [178] and Section 2.4.3) was employed to train regression models that can predict Arousal and Valence scalar values on a continuous scale. Similar to SVMs, SVR is a kernel-based method and allows the use of the kernel trick to transform the original feature space to a higher-dimensional feature space through a (non-linear) kernel function. We used ϵ-SVR available in *libsvm* ([37]) to train our models. In SVR, a margin ϵ is introduced and SVR tries to construct a discriminative hyperplane that has at most ϵ deviation from the original training samples. In our emotion prediction experiments, the RBF kernel function was used, and the parameters c (cost), ϵ (the ϵ

of the loss function), and γ were tuned on a development set. The parameters were tuned via a simple grid search procedure that evaluates all possible combinations of c (with exponentially growing values between 2^{-4} and 2^{4}), ϵ (with exponentially growing values between 10^{-3} and 10^{0}), and γ (with exponentially growing values between 2^{-10} and 2^{2}).

Features: acoustic features

The acoustic feature extraction was performed with Praat (Boersma and Weenink [23]). Prior to feature extraction, a voiced-unvoiced detection algorithm (available in Praat) was applied to find the voiced units. To avoid the use of ASR, that can provide word alignments, the features were extracted over each *voiced unit* of a segment. We made a selection of features based on previous studies (e.g., Batliner et al. [18], Banse and Scherer [12]), and grouped these into features related to *pitch* information, *energy/intensity* information, and information about the *distribution of energy in the spectrum*. The spectral features MFCCs as commonly used in ASR were also included. And finally, global information calculated over the whole segment (instead of per voiced unit) about the speech rate and the intensity and pitch contour was included. An overview of the features used is given in Table 6.12.

Level	Features		N_{feat}
voiced unit	Pitch (PITCH)	mean, standard deviation, range (max-min), mean absolute pitch slope	4
voiced unit	Intensity (INTENS)	Root-Mean-Square (RMS), mean, range (max-min), standard deviation	4
voiced unit	Distribution energy in spectrum (ESPECTR)	slope Long-Term Averaged Spectrum (LTAS), Hammarberg index, standard deviation, center of gravity (cog), skewness	5
voiced unit	MFCC (MFCC)	12 MFCC coefficients, 12 deltas (first order derivatives)	24
whole segment	other	speech rate1, speech rate2, mean positive slope pitch, mean negative slope pitch, mean positive slope intensity, mean negative slope intensity	6

Table 6.12: *Acoustic features used for emotion prediction with SVR.*

Pitch and energy/intensity information are known to be useful in emotion recognition and are thus very commonly used. MFCCs are powerful speech features and are commonly used in ASR and speaker and language recognition technologies. The distribution of energy in the spectrum can give information about the vocal effort: in general, when speakers increase their vocal effort, the energy in the higher frequency regions of the long-term spectrum increase which results in a less steep spectral slope. The Hammarberg index is a measure that measures differences of the energy in different frequency regions of the long-term spectrum: it is defined here as the maximum energy measured in the frequency region 0–2000 Hz minus the maximum energy

measured between 2000 and 4000 Hz. The features 'speech rate1' and 'speech rate2' are calculated per segment and are defined as the number of voiced units divided by the segment duration without and with silences respectively. The mean positive and negative slopes of pitch and intensity are calculated by summing and averaging all the positive and negative changes in pitch and intensity measured framewisely over the voiced parts.

The majority of our acoustic features were measured per voiced unit. The features extracted on voiced-unit-level were aggregated to segment-level by taking the **mean**, **minimum**, and **maximum** of the features over the voiced units. Hence, we obtain per segment a feature vector with $(3 \times (4 + 4 + 5 + 24)) + 6 = 117$ dimensions. These features were normalized by transforming the features to z-scores: $z = (x - \mu)/\sigma$, with μ and σ calculated over a development set.

Features: lexical features

As SVMs (and SVRs) do not take raw text (words) as input, we used lexical features that are based on a continuous representation of the textual input. The textual input in our case is a manual word-level transcription made by the author herself. A fairly standard method to build features from textual input, and that has successfully been applied to text and document classification/retrieval (see e.g., Salton and Buckley [156], Joachims [87]) was employed, namely a *tf-idf* weighting scheme (**t**erm **f**requency-**i**nverse **d**ocument **f**requency). The *term frequency* $tf_{w,s}$ is defined as the number of times a given word w appears in a segment s (i.e., an utterance) and reflects its importance to that specific segment. The *document frequency* df_w is defined as the number of segments containing word w. The tf-idf weight for each word w is then computed by:

$$\text{tf-idf}_{w,s} = tf_{w,s} \times idf_w = tf_{w,s} \times log(\frac{N}{df_w}) \tag{6.6}$$

where N is the total number of segments in the training set. The weights tend to filter out common words. Words that appear frequently in one utterance ($= tf$), but rarely in the whole collection of utterances ($= idf$) are more likely to be relevant to that utterance and thus have a high tfidf weight. In addition, to adjust for differences in utterance length, the feature vectors were normalized to unit length by L2-normalization.

$$x_n = \frac{x_n}{\sqrt{\sum_{i=0}^{N} x_i^2}} \tag{6.7}$$

where x_n is a value in a vector with N dimensions. To give an idea of the size of N, the number of unique words in the whole corpus is 1963.

6.4.6 Experiments and Results

We first describe the experimental setup and performance metrics used. Subsequently, the results obtained with the emotion predictors developed are presented. In addition, we present results of a comparison made between acoustic feature sets, results of a

Gender	N_{segments}	N_{fold}	Splits (approximately) in training-development-testing sets
Female	1048	11	80%-10%-10%
Male	1352	17	87%-8%-5%

Table 6.13: *Experimental setup of the material for N-fold cross-validation experiments.*

comparison between human and machine performance, and results of a comparison between acted and spontaneous emotional speech.

Experimental setup

The automatic emotion prediction experiments were carried out speaker-independently, but separately for female and male speakers. We performed N-fold cross-validation, where in each fold, we leave out one specific speaker for testing. In each fold, the data set was divided into three sets: a training, development and test set (see Table 6.13), where the training and test sets are disjoint. The test set consists of speech segments from a specific speaker that is excluded from the training and development set. The development set is comprised of randomly picked segments, drawn from the remaining segments after the test speaker has been filtered out.

The development set is used for parameter tuning and z-scoring. The features were normalized by z-scoring ($z = (x - \mu)/\sigma$) where the μ and σ were calculated on the development set. In parameter tuning, the parameter set that achieved the lowest error rate, averaged over N folds, was selected to use in the final testing. The computation of the error rate is explained in Chapter 6.4.6, and is shown in Eq. (6.9).

We performed two types of prediction experiments. One is based on the SELF-annotations, and the other one is based on the OTHER.AVG-annotations. The SELF-annotations refer to the annotations that were made by the gamers themselves which are most likely to reflect 'felt' emotions. The OTHER.AVG-annotations refer to the averaged annotations of 3 different observers who annotated the 'observed' emotions of the gamers. With these two experiments, we compared the added value of annotation of felt emotion versus annotation of perceived emotion.

Evaluation metric

Because there are various evaluation metrics applicable to this emotion prediction task, we report several evaluation metrics. Firstly, we used a relatively simple evaluation metric (similar to Grimm et al. [69]) that measures the absolute difference between the predicted output and the reference input:

$$e_i = |x_i^{\text{pred}} - x_i^{\text{ref}}| \tag{6.8}$$

$$e_{\text{avg}} = \frac{1}{N}\sum_{i}^{N} e_i \tag{6.9}$$

and we report the e that is averaged over a total of N segments. The lower e_{avg}, the better the performance. Secondly, we calculated Krippendorff's α between the human

reference and the machine predictions in order to evaluate machine performance in terms of human-machine agreement. Thirdly, Pearson's ρ correlation coefficient was also reported. Finally, if possible, equally weighted Kappa κ was calculated.

Results acoustic and lexical emotion predictors

Here, we present the results of our emotion prediction experiments which were performed separately for female and male data, and separately for Arousal and Valence dimensions. One experiment employs SELF-annotations as reference, and the other one employs OTHER.AVG-annotations as reference. The emotion predictors were developed with either acoustic information or lexical information. The main evaluation metrics are e_{avg} and $\alpha_{ord,5}$. The results for the acoustic and lexical Arousal and Valence predictors are presented in Table 6.14 and 6.15.

	Gender	Reference	Test_{SVR}				Baseline	
			e_{avg}	$\alpha_{\text{ord},5}$	$\kappa_{\text{equal},5}$	ρ	e_{avg}	$\alpha_{\text{ord},5}$
ARO	F	SELF	0.46	0.09	0.05	0.11	0.46	-0.03
	M	SELF	0.36	0.32	0.19	0.37	0.44	-0.10
	F	OTHER.AVG	0.20	0.37	0.27	0.48	0.32	-0.28
	M	OTHER.AVG	0.22	0.44	0.31	0.57	0.31	-0.12
	ALL	**SELF**	**0.41**	**0.22**	**0.13**	**0.25**	**0.45**	-0.07
	ALL	**OTHER.AVG**	**0.21**	**0.42**	**0.30**	**0.55**	**0.31**	-0.18
VAL	F	SELF	0.37	0.07	0.04	0.12	0.37	0.00
	M	SELF	0.34	0.12	0.08	0.20	0.35	-0.02
	F	OTHER.AVG	0.25	0.23	0.17	0.35	0.25	0.00
	M	OTHER.AVG	0.27	0.31	0.22	0.45	0.29	0.00
	ALL	**SELF**	**0.36**	**0.10**	**0.06**	**0.18**	**0.36**	-0.01
	ALL	**OTHER.AVG**	**0.26**	**0.28**	**0.20**	**0.41**	**0.28**	0.00

Table 6.14: *Results of the acoustic Arousal and Valence predictors: the last two columns under 'Baseline' represent results from a baseline predictor that always predicts Neutrality.*

According to the results shown in Table 6.14, the Arousal scale is better modeled than the Valence scale with acoustic information. Furthermore, performance is consistently higher when OTHER.AVG-annotations were used rather than SELF-annotations. There seems to be a small advantage for the male data which is most of the time slightly better modeled than the female data. Finally, although the performances are relatively low, the predictors developed at least perform better than a baseline predictor that always chooses Neutrality, see the last two columns in Table 6.14.

In contrast with the acoustically based predictors, the lexically based predictors appear to be able to model Valence better than Arousal, see Table 6.15. Similar to the acoustically based predictors, performance is higher with OTHER.AVG-annotations than with SELF-annotations, and performance is above baseline.

In Fig. 6.21, the machine predicted output is plotted against the human reference input. These plots reflect the human-machine agreement and correlation figures as

	Gender	Reference	Test_{SVR}				Baseline	
			e_{avg}	$\alpha_{\text{ord},5}$	$\kappa_{\text{equal},5}$	ρ	e_{avg}	$\alpha_{\text{ord},5}$
ARO	F	SELF	0.48	-0.07	-0.05	-0.06	0.46	-0.03
	M	SELF	0.40	0.03	0.01	0.08	0.44	-0.10
	F	OTHER.AVG	0.23	0.18	0.13	0.27	0.32	-0.28
	M	OTHER.AVG	0.26	0.16	0.09	0.27	0.31	-0.12
	ALL	**SELF**	**0.44**	**-0.01**	**0.01**	**0.01**	**0.45**	-0.07
	ALL	**OTHER.AVG**	**0.24**	**0.19**	**0.12**	**0.29**	**0.31**	-0.18
VAL	F	SELF	0.37	0.07	0.04	0.16	0.37	0.00
	M	SELF	0.35	0.08	0.05	0.11	0.35	-0.02
	F	OTHER.AVG	0.20	0.47	0.38	0.61	0.25	0.00
	M	OTHER.AVG	0.23	0.48	0.36	0.62	0.29	0.00
	ALL	**SELF**	**0.36**	**0.07**	**0.04**	**0.14**	**0.36**	-0.01
	ALL	**OTHER.AVG**	**0.21**	**0.48**	**0.37**	**0.62**	**0.28**	0.00

Table 6.15: *Results of the lexical Arousal and Valence predictors:the last two columns under 'Baseline' represent results from a baseline predictor that always predicts Neutrality.*

can be observed in Table 6.14 and Table 6.15: the figures and plots show that there is relatively low agreement and correlation between human and machine.

The main reason that we have reported several different performance metrics, is that we wanted to be able to make comparisons with other studies. Grimm et al. [69] and Grimm et al. [68] report mean absolute errors between 0.13 and 0.28, and Pearson's ρ correlation coefficients of 0.75–0.82 and 0.28–0.46 for Arousal and Valence prediction respectively. Laskowski and Schultz [104] report human inter-rater agreement κ between 0.15 and 0.67 for 3 human annotators, and Douglas-Cowie et al. [54] report human inter-rater agreement κ between 0.37 and 0.54, based on category emotion annotation. Although real valid comparisons are not possible, e.g., due to different databases used, different annotation procedures used etc., a comparison with our results can give us a rough idea of how our results relate to other studies. With the emotion predictors developed, we obtained mean absolute errors e_{avg} ranging from 0.21 to 0.26 (see Table 6.14 and 6.15). The correlation coefficients ρ between man and machine obtained lie in the range 0.29–0.55 and 0.41–0.62 for Arousal and Valence respectively (with $kappa$s between 0.12 and 0.37). These results suggest that the machine performances achieved in our experiments lag behind that of the machine and human performance reported in other studies.

The main results are summarized in Fig. 6.22 where the female and male results are combined together, and where the performances based on the OTHER.AVG-annotations are shown. To summarize, performance was higher with the OTHER.AVG-annotations than with the SELF-annotations, which suggest that 'felt' emotions are more difficult to recognize than 'perceived' emotions. We further discuss this observation and its implications in Section 6.5. Secondly, in spontaneous emotional speech data, Valence information can be best captured and modeled by lexical features (rather than acoustic features). Arousal can be better modeled by acoustic fea-

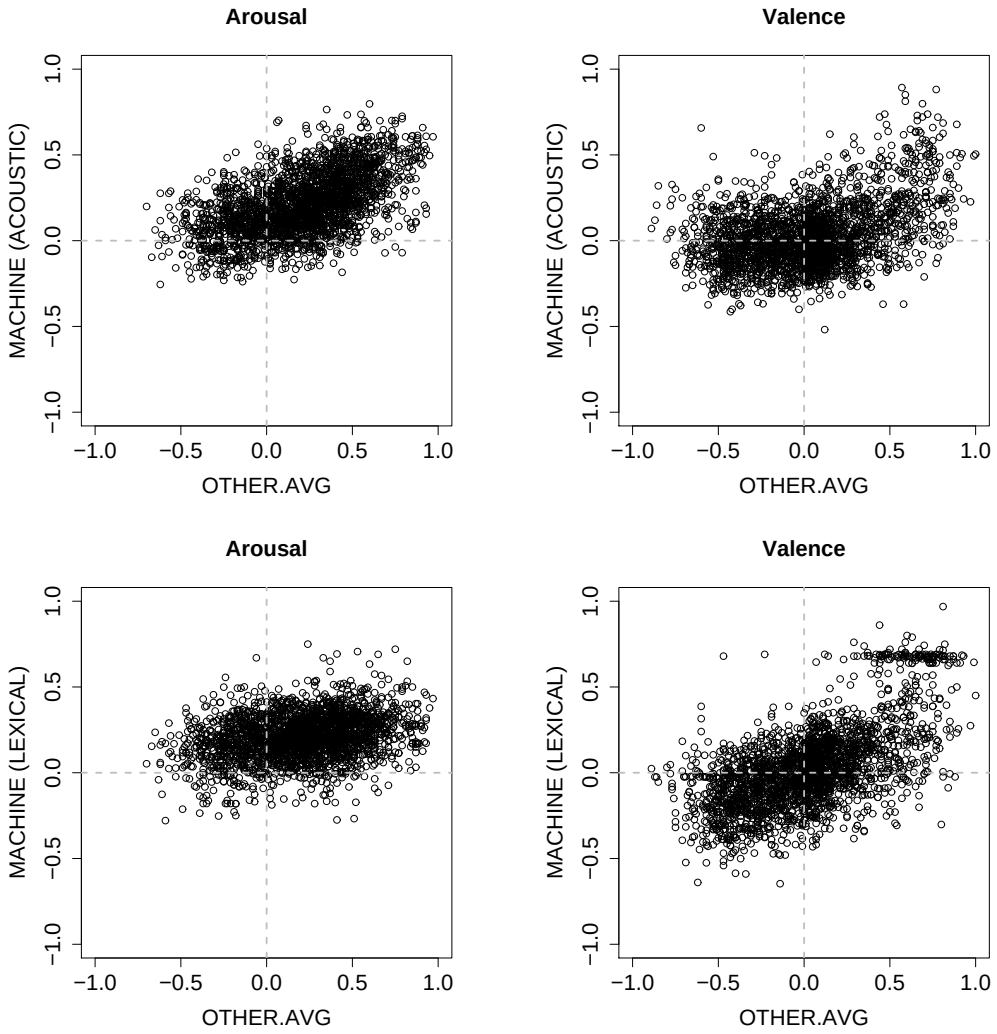


Figure 6.21: *Human reference plotted against* MACHINE*-predictions for acoustic (top) and lexical (bottom) predictors on Arousal (left) and Valence (right) scales.*

tures. Thirdly, the machine performance and human-machine agreement is relatively low. However, note that the prediction experiments were carried out on relatively short segments of speech (averaged duration of 1.50 s). In the following experiments, we compared several types of feature sets for Arousal and Valence prediction, and we tried to improve machine performance with a reduced set of acoustic features. In addition, we compare machine performance to human performance, and interpret machine performance in terms of human-machine agreement.

Comparison types of acoustic feature sets

In order to investigate what types of acoustic features contribute the most to the modeling of Arousal and Valence, we carried out prediction experiments with the separate feature sets, described in Table 6.12: PITCH, INTENS, EPSPECTR, and MFCC

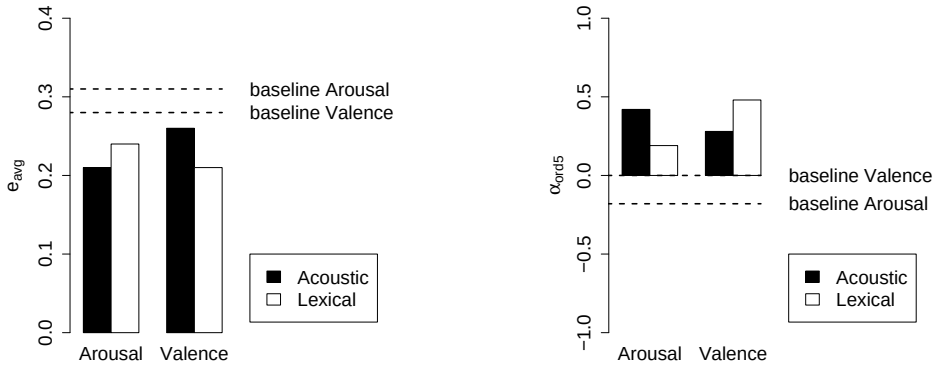


Figure 6.22: e_{avg} *(left) and* $\alpha_{\text{ord},5}$ *(right) performances of female and male results together, based on* OTHER.AVG*-annotations. Baseline is based on a predictor that always predicts Neutrality.*

('other' was not included because it consists of several different types of acoustic features). Rather than performing an extensive search on individual acoustic features, we preferred to bundle closely related features and perform prediction experiments with these groups of features. We are more interested in the importance of types of features rather than the selection of important individual features that, to a certain extent, is data-dependent. However, an extensive search on individual acoustic features may improve prediction performance, hence, we also applied a stepwise method in order to obtain a reduced set of acoustic features (referred to as REDUCED) with the aim to improve performance.

Traditionally, in acted emotional speech, the literature reports that intensity and pitch features are most important for modeling emotions, more specifically Arousal. For Valence, quality-related features like MFCCs usually perform better. Since there appear to be differences in the production of acted and spontaneous emotional speech (Wilting et al. [210]), it is likely that the importance of feature types also differs between acted and spontaneous emotional speech (Vogt and André [203]).

Gender	Dim.	Features						
ALL	ARO	ESPECTR	$>$	INTENS	$>$	MFCC	$>$	PITCH
		0.40	$>$	0.40	$>$	0.33	$>$	0.25
ALL	VAL	MFCC	$>$	ESPECTR	$>$	PITCH	$=$	INTENS
		0.25	$>$	0.19	$>$	0.16	$=$	0.16

Table 6.16: *Ranking of feature sets by* $\alpha_{\text{ord},5}$.

Table 6.16 and Fig. 6.23 show the ranking of the acoustic features for Arousal and Valence. It appears that Arousal can be best modeled by energy-related features like ESPECTR and INTENS, while Valence can be best modeled by MFCCs and ESPECTR. Surprisingly, PITCH features are less important than expected. This is in line with

the results as reported in Vogt and André [203]: they found that for spontaneous emotional speech, MFCCs rather than pitch-related features, are more important for automatic emotion classification.

To see whether the performance of the emotion predictors could be improved by feature selection, we applied a stepwise regression method (Venables and Ripley [198]) to select the best features. Based on a regression model, stepwise regression at each stage adds or removes a variable according to some criterion, in this case, the Akaike Information Criterion (AIC, see Venables and Ripley [198]). During our K-fold cross-validation prediction experiments with the development sets, we collect K number of different feature sets, selected by the stepwise regression method. Since we aim to have 1 fixed feature set that is applied to all K folds, we derive from the K different feature sets 1 fixed feature set. The size N of this fixed feature set is determined by the averaged size of the K feature sets that were obtained via stepwise regression. Subsequently, the frequency of each individual acoustic feature that is selected by the stepwise procedure is counted over all K folds, and the top N of most frequently selected features are included in the fixed feature set which will be referred to as the REDUCED feature set. In Fig. 6.23 and Table 6.17, we can observe that the REDUCED set of acoustic features obtained with stepwise regression outperforms all other feature sets, including the full set of features.

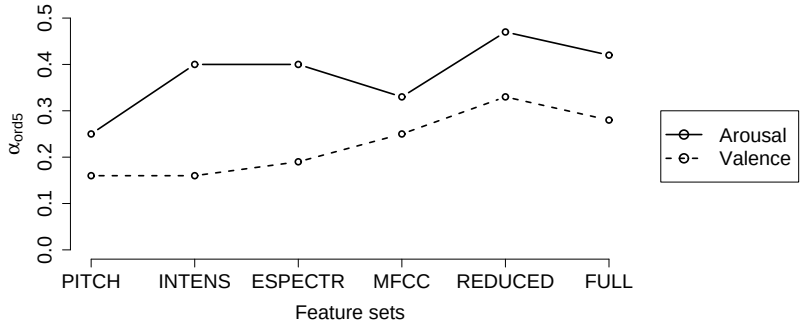


Figure 6.23: $\alpha_{\mathrm{ord},5}$ *obtained with several types of acoustic feature sets, based on* OTHER.AVG*-annotations as reference.*

Comparison between human-machine agreement and human-human agreement

So far, we have mainly considered the emotion predictors' performances from a machine learning point of view: given a set of labeled data, what is the lowest error rate we can achieve and is it better than another predictor's performance? We have concluded that machine performance was relatively low from a classifier's perspective. Another way to assess performance is to compare machine performance to human performance. In Table 6.18 and Fig. 6.24, we show the human-machine agreement obtained with the automatic emotion predictors, next to the human-human agreement, obtained with the human annotation process.

Although the comparison is not completely fair in Table 6.18 (the human-human

Feature set	N_{feat}	Gender	Dim.	e_{avg}	Pearson's ρ	$\alpha_{\text{ord},5}$	$\kappa_{\text{equal},5}$
REDUCED	37	F	ARO	0.19	0.52	0.38	0.27
	34	M	ARO	0.20	0.61	0.51	0.36
	48	F	VAL	0.23	0.43	0.32	0.24
	54	M	VAL	0.25	0.46	0.33	0.26
	37, 34	**ALL**	**ARO**	**0.20**	**0.58**	**0.47**	**0.33**
	48, 54	**ALL**	**VAL**	**0.25**	**0.44**	**0.33**	**0.25**
FULL	117, 117	ALL	ARO	0.21	0.55	0.42	0.30
	117, 117	ALL	VAL	0.26	0.41	0.28	0.20

Table 6.17: *Results obtained with a reduced acoustic feature set, using* OTHER.AVG*-annotations as reference.*

	Krippendorff's α			
	Human-machine agreement between OTHER.AVG-ratings and MACHINE-predictions		**Human-human** agreement among OTHER.3-ratings	**Human-human** agreement between SELF-ratings and OTHER.AVG-ratings
	Acoustic (REDUCED)	Lexical		
Arousal	0.47	0.19	0.28	0.27
Valence	0.33	0.48	0.57	0.36

Table 6.18: *Human-machine agreement and human-human agreement, female and male results combined.*

agreement shown in the second column of Table 6.18 is based on OTHER.3-ratings rather than OTHER.AVG-ratings), it gives an idea of how machines perform in comparison with humans. For example, in Table 6.18, if we compare the human-machine agreement shown in the first column, to the human-human agreement shown in the last column, we can observe that an acoustic Arousal predictor shows more agreement with OTHER.AVG-annotations than a human SELF-rater does. Similarly, a lexical Valence predictor outperforms a human SELF-rater. When we compare the human-machine agreement figures to the human-human agreement figures shown in the middle column, we find that the machine predictors can produce emotion predictions that approach or even surpass the levels of agreement among humans. In the case of acoustic Arousal prediction, our system can give predictions that surpass the level of human-human agreement. In the case of lexical Valence prediction, the system's predictions approach human-level performance. These observations have also been made visible in Fig. 6.24. In Fig. 6.24, the inter-rater agreement among OTHER.3-raters is plotted, including the agreement when the SELF-ratings, or the MACHINE-predictions are added. In addition, 3 upper and 3 lower inter-rater agreement boundaries are plotted which represent the highest agreements possible (when 1 annotator is added who perfectly agrees with 1 of the OTHER.3-raters, 'best-case' scenario) and the low-

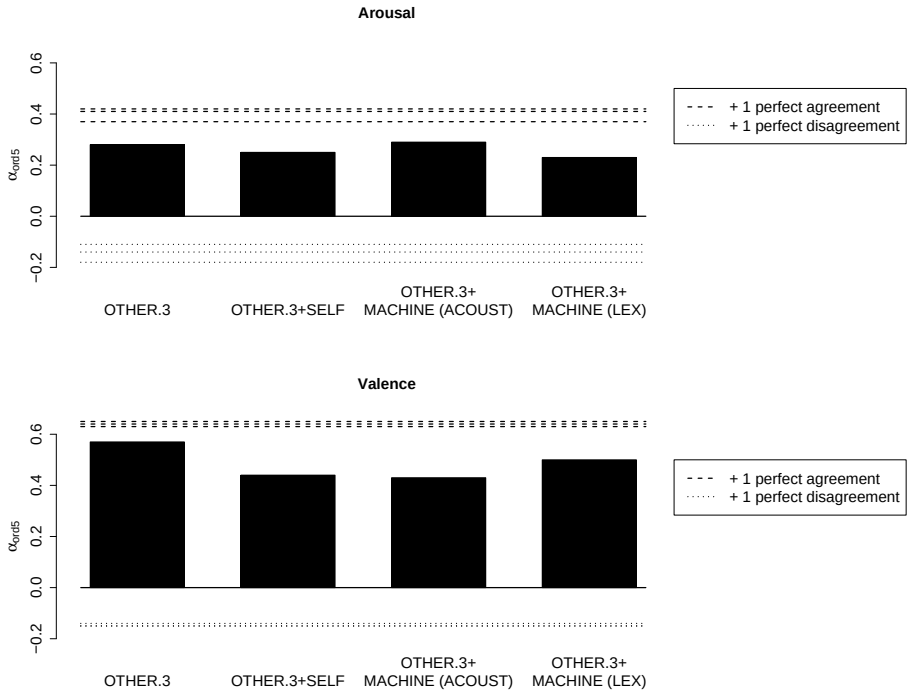


Figure 6.24: *Inter-rater agreement among* OTHER.3*-ratings plus* SELF*-ratings, or plus* MACHINE*-predictions (top: Arousal, bottom: Valence).*

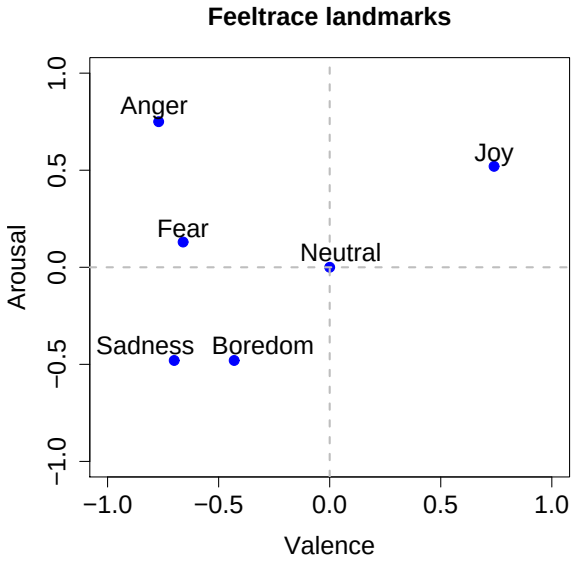
est agreements possible (when 1 annotator is added who 'perfectly' disagrees with 1 of the OTHER.3-raters, 'worst-case' scenario) respectively. In the case of Arousal, we can observe that when the acoustic machine's predictions are added to the OTHER.3-ratings, the inter-rater agreement does not decrease while in the case of Valence, when the lexical machine's predictions are added to the OTHER.3-ratings, the inter-rater agreement does decrease. These observations support our previous ones, namely a) that our acoustic Arousal predictor can perform at human-level performance in terms of agreement, and b) that our lexical Valence predictor can approach human performance in terms of agreement, but still needs improvement.

In summary, from a machine learning perspective, the emotion predictors developed have a relatively low performance. However, from a human's perspective, it seems that machines predict emotions (almost) just as badly (or as well) as humans do.

6.4.7 Comparison with acted emotional speech

In order to show that this SVR method combined with the features selected *can* work under other conditions, the method was also applied to an acted emotional speech database, the BERLIN Emotional Speech database (Burkhardt et al. [25]). However, the BERLIN database contains emotional speech that is organised in discrete emotion categories, and hence, lacks Arousal and Valence ratings. Therefore, in order to obtain these ratings, each discrete emotion category was replaced by an Arousal and

Valence landmark rating as given by the Feeltrace tool (the squared version), see Fig. 6.25 (note that this Section shows some resemblance with Section 4.7, however, the purpose of this experiment is different from the one described in Section 4.7). The emotion Disgust was discarded because there are no Feeltrace landmark ratings available for this emotion.



Emotion	Arousal	Valence
Anger	0.75	-0.77
Joy	0.52	0.74
Fear	0.13	-0.66
Neutral	0	0
Sadness	-0.48	-0.70
Boredom	-0.48	-0.43

Figure 6.25: *Discrete emotion categories from the* BERLIN *database with their corresponding Feeltrace ratings.*

The exact same SVR method, acoustic features and development procedures were applied to the BERLIN database. Female and male models were trained and tested separately, and speaker-independently. The distribution of the samples is as follows:

	Anger	Joy	Fear	Neutral	Sadness	Boredom	Total
Female	67	40	29	40	36	45	257
Male	60	24	26	38	17	34	199

Table 6.19: *Number of samples from the* BERLIN *database used in SVR experiment.*

The results of the SVR experiments carried out on the BERLIN database are shown in Table 6.20 and in Fig. 6.26. According to the figures in Table 6.20, Arousal is much better modeled in the Berlin database than in the TNO-GAMING database. This is not the case for Valence. But this can be explained by the fact that the spread of positive and negative emotions is poor in the BERLIN database (as can be seen in Fig. 6.26): there is only one positive emotion class in the BERLIN dataset, namely Joy. These results imply that when the emotional speech data is neatly arranged, i.e., acted full-blown emotions, good spread of various emotions, clean speech signal etc., the SVR method in combination with the features selected can be used to predict Arousal with a relatively good performance, in comparison with a spontaneous database like TNO-GAMING.

		Test_{SVR}		
		e_{avg}	$\alpha_{\text{ord},5}$	Pearson's ρ
Berlin (acoustic FULL)	Arousal	**0.19**	**0.81**	**0.87**
	Valence	0.38	0.17	0.35
TNO-GAMING (acoustic FULL)	Arousal	0.21	0.43	0.56
	Valence	0.26	0.28	0.40

Table 6.20: *Comparing results from prediction experiments with acted emotional speech (*BERLIN *and spontaneous emotional speech (*TNO-GAMING*, based on* OTHER.AVG*-annotations).*

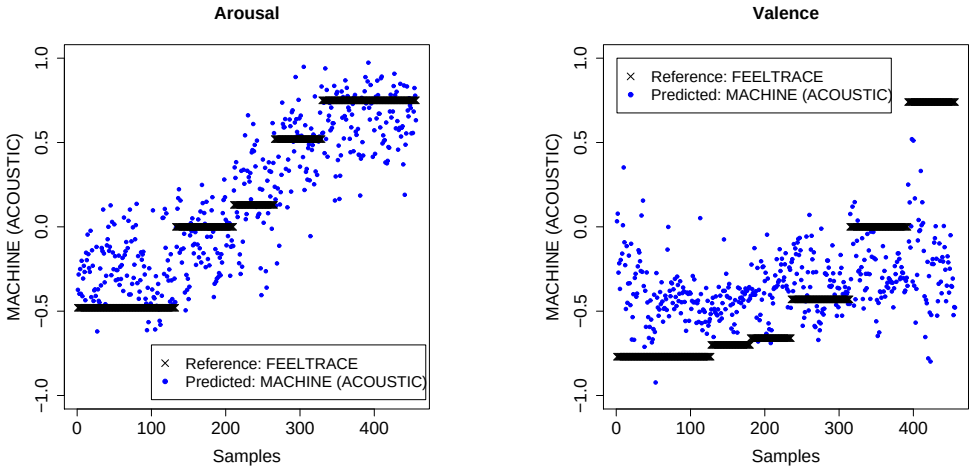


Figure 6.26: *Predictions of the SVR method applied to* BERLIN *database for Arousal (left) and Valence (right) prediction.*

6.4.8 Conclusions

In Experiment II, we developed and tested emotion recognizers that can predict Arousal and Valence scalar values using acoustic and lexical speech features. A subset of the dataset (= 2400 segments) was re-annotated by 6 observers such that each segment of the subset was annotated by 3 observers and the SELF-rater. As reference, we used the SELF-annotations of the gamers themselves or the annotations of the 6 observers (OTHER.AVG-annotations). In a reliability analysis, we found, similar to the findings in Experiment I, that there are differences between SELF-raters' emotion judgments and emotion judgments from several observers. Consequently, these differences affected the performances of the automatic emotion predictors. It appeared that the emotion recognizers achieve a better performance with the OTHER.AVG-annotations than with the SELF-annotations: this can be seen as an indication that emotion recognizers can model *observed* emotions better than *felt* emotions. Furthermore, we confirmed that in spontaneous emotional speech, Arousal can be better modeled by acoustics and Valence can be better modeled by lexical features. Moreover, Arousal was best modeled by the acoustic feature sets ESPECTR and INTENS, while for Valence, the MFCCs outperformed the other acoustic feature sets.

Unexpectedly, PITCH was the worst performing feature set. A reduced set of acoustic features comprised of features selected by stepwise regression yielded slightly better results than the FULL set of features. Finally, although from a machine learning point of view, the emotion recognizers seem to perform moderately, these performances are not so poor from a human's perspective: the human-machine agreement for Arousal prediction based on acoustic features is on par with the human-human agreement, while the human-machine agreement for Valence prediction based on lexical features lags slightly behind that of human-human agreement.

6.5 Discussion and conclusions

The human assessment of spontaneous emotion is a complex process, and to a certain extent, subjective given the relatively low agreement scores reported in previous studies and the current study. It is expected that agreement among humans increases when there is more 'information' available, i.e., when more multimodal sources are available to base the emotion assessment on. In Experiment I, we found that this is indeed the case: the agreement figures were higher in the audiovisual condition than in the audio only or visual only conditions. However, when context information (i.e., the video stream of the game itself) was added, the agreement did not always increase. One of the reasons why the addition of context information did not consistently increase agreement, might be that we provided context information in a way that did not help the observers but rather made the task slightly more difficult. The game content was shown next to the visual channel with the consequence that the observer had to divide his/her attention between two screens. In addition, the movie clips provided to the observers were cut from their 'contextual flow', so that it may have been difficult to place the video clips in their exact context. The value of an additional source for emotion assessment seems to depend much on *what* type of information is added and *how* it is added. However, it is not clear yet how these multimodal sources interact with each other and how emotion is processed in a multimodal way, especially in cases where multimodal emotional expressions are incongruent.

We also found in Experiment I that there are differences between SELF-annotations and OTHER-annotations. A minor drawback of this analysis is that we do not have information about the annotation skills of the SELF-raters and the OTHER-raters in Experiment I (due to practical limitations). In other words, we do not know their intra-rater reliabilities. It would have provided more insight into whether humans are capable of judging their own emotions.

In Experiment II, automatic speech-based emotion recognizers were developed with SELF-annotations and OTHER.AVG-annotations. One of the main conclusions of Experiment II is that the MACHINE performance obtained with the predictors were worse when these were trained with SELF-annotations as reference, than when trained with OTHER.AVG-annotations as reference. There are several explanations possible for this result. Firstly, it is possible that 'felt' emotions cannot be captured through acoustic and lexical features, while 'observed' emotions can. This suggests that the recognition technology is not mature and advanced enough to recognize 'felt' emotions, and that we should aim at the recognition of perceived emotions first. Secondly,

we have to note that in the SELF-condition, the experiments were performed speaker-independently *and* 'annotator-independently' since in this condition, the speaker is the same person as the annotator. Whereas in the OTHER.AVG-condition, the annotators were drawn from the same pool in training and testing. Thirdly, we have to keep in mind that the SELF-annotations and OTHER.AVG-annotations were carried out in slightly different ways which may have resulted in 'noisier' SELF-annotations. The SELF-annotators were provided with all possible information available, audiovisual recordings of the webcam and contextual information (the video stream of the game), but they could not pause or re-wind the video. The annotation took place right after the game was finished because we wanted the gamers to be able to remember how they felt during the game. Therefore, there was no time to segment the video into smaller segments. The observers on the other hand could pause or re-view the pre-segmented video segments and were provided audiovisual information only. It may appear as if the observers were advantaged in their annotation procedure but is it not more advantageous when one can judge his/her own emotion that she/he has just experienced? It remains a debate how to annotate emotion, but from our experiments, we have learned that the SELF-annotations have lead to worse machine performance than the observers' annotations.

Another conclusion of Experiment II is that in terms of agreement, the automatic emotion recognizer performs on human-level. For Arousal, α for human-human agreement among OTHER.3-raters is 0.28 whereas the emotion predictor based on acoustic features achieves a human-machine agreement of 0.47. So the machine agrees better with a human than other humans among each other do. From a machine learning point of view, this performance is rather moderate, because on an α-scale from [-1,1], the score is 0.43. For Valence, α for human-human agreement among OTHER.3-raters is 0.57 whereas the predictor based on lexical features achieves a human-machine agreement of 0.48: the predictor seems to lag behind human performance. A possible way to improve the human-machine agreement for Valence is to include a predictor based on facial expressions. Recall that the OTHER.AVG-annotations were based on audiovisual information; the visual part, which we have not dealt with at all, could supplement and improve the current human-machine agreement. From a machine learning point of view, the Valence predictor based on lexical features performs moderately. But how badly is it really for the automatic emotion recognizers to perform moderately? Because compared to human performance, which is rather low, the machine performance does not seem too bad. In addition, the machine performance seems to depend on the type of data used. In acted emotional speech data, where the annotation is given, machine performances are much better. But in spontaneous emotional speech, where emotions are more 'shadier' than in acted emotional speech, humans do not easily agree about the observed emotions. If humans are bad in judging real affect, then how can we expect from a machine to do it better? Yet, in human-human communication we do not seem to have any problems interpreting social or emotional cues. Perhaps an automatic emotion analyzer should aim at recognizing those social and affective cues in human-human conversation, taking into account the unwritten conversational rules that serve as a 'context'.

Chapter 7

Conclusions

From Terminator 2 (1991):
———
The Terminator: "It has to end here."
John Connor: "I order you not to go. I order you not to go, I ORDER YOU NOT TO GO!"
[John starts to cry]
The Terminator: "I know now why you cry," [Terminator wipes Johns tears]
The Terminator: "but it is something I can never do."
———

At the end of the movie Terminator 2, the Terminator appears to have acquired some emotional intelligence: it now understands why John, the human, cries. Somehow, the affective system built within the cyborg must have learned why humans cry.

Our work described in this thesis was focused on a specific aspect of building affective systems: we investigated the effects of using real affective speech data on affect recognition in speech. By performing affect recognition experiments on several different types of emotional speech data, we acquired knowledge about how the use of natural affective data affects the way affect recognizers can be developed. In Section 7.1, we summarize our findings, acquired by experimenting, and recapitulate the research questions stated in Section 1.4.1, and we discuss the conclusions drawn. Finally, in Section 7.2, we give recommendations for future research.

7.1 Research questions

The main aim of the work described in this thesis was to develop speech-based affect recognizers for real affective speech. Using spontaneous speech material, we have developed affect recognizers for the detection of laughter, and sentiments and opinions in the context of meetings (see Chapter 5). In the context of gaming, speech-based affect recognizers were developed that can predict Arousal and Valence scalar values (see Chapter 6). Since the use of acted emotional speech for the development of affect recognizers is still an attractive option that has been frequently chosen by researchers, we employed acted emotional speech to illustrate alternative evaluation

methodologies, borrowed from other similar recognition technologies, that are also very well applicable to the emotion recognition task and that have the possibility to emulate 'real-life' situations (see Chapter 4). Finally, since the way in which the spontaneous emotional speech data is acquired and annotated plays an important role in the development of affect recognizers, efforts were taken to acquire such data in the field (see Chapter 3), and a new corpus containing spontaneous emotional speech was recorded to explore how differences in annotation may affect the recognizer's performance (see Chapter 6). Along with the development of these speech-based recognizers of *real* affect, several interesting research questions that were stated in Chapter 1 could be addressed.

More than for other similar recognition technologies, such as speaker or language recognition, the development of affect recognition techniques is strongly dependent on the nature of the available speech material. Since most of the affect recognition systems are most likely to be exposed to natural affect, the use of acted emotional speech data in emotion classification experiments reduces the relevance of the outcomes for real-life data: natural affect contains much more subtle emotion expressions that often cannot be classified in one of the Big Six universal emotion categories. One of the major difficulties in emotional speech research is the scarcity of labeled real affective speech material. It is a very time and labor consuming process to acquire a substantial amount of natural affective speech data that is labeled, and that can be used to develop automatic affect recognition systems. This is in contrast with speaker or language recognition where there are hundreds hours of speech available for the development of speaker and language recognition systems. An additional difficulty is that there is no consensus on how to describe and annotate real affect, and that there is no 'ground truth'. We hypothesized that the type and strength of emotion expression is strongly dependent on the naturalness of the context in which these emotions are expressed, and that the description and annotation of affect is heavily dependent on the naturalness of emotions expressed and the context in which these were expressed.

Hence, the first two research questions, that were stated in Chapter 1, are related to the naturalness and the description of **emotional speech data**:

Research question 1: How does the speech data's level of naturalness used in speech-based affect recognition affect the task and performance of the recognizer?

Research question 2: How does the description and annotation of emotional speech data that is used in speech-based affect recognition, affect the task and performance of the recognizer?

In this research, several emotional speech databases were used in our emotion recognition experiments. In terms of naturalness of emotions occurring in these databases, the data used ranged from acted emotions (see Chapter 4), to emotions (see Chapter 6), to natural emotions (see Chapter 5). For acted emotional speech, the emotion labels given to the emotional speech signals are usually straightforward: it is the emotion category that the actor was asked to perform. Usually, full-blown and basic universal emotions (e.g., Anger, Joy, Fear, Disgust, Sadness, Boredom) are involved. Human and machine recognition performances on these type of datasets are usually

relatively good. In the context of basic emotions, our experimental results obtained with the BERLIN database, see Chapter 4, indicate that Sadness appears to be well detectable by machines. Anger is also relatively easy to detect, however, it is often confused with Joy. Similar to humans, machines have difficulty detecting Fear and Disgust. Although the recognition of full-blown emotions is still not perfect, the procedure we proposed for the development of such an emotion recognizer can run relatively smoothly, and is relatively straightforward. This is in contrast with the effort that is needed to build an affect recognizer that can recognize natural, spontaneous emotions.

Subsequently, in Chapter 5, we performed emotion recognition experiments with natural emotional speech data which, in this case, was extracted from natural meetings. As is known, expressions of full-blown emotions are rare in natural, daily-life situations. Researchers who have put an effort in annotating and describing natural (meeting) data, have found very few occurrences of full-blown or basic emotions. Hence, the description of natural (meeting) data requires a different approach. Researchers have annotated meeting data in terms of 'emotionally colored behavior', focusing more on human behavior and interaction that is most likely triggered by some affective event during conversation. Assigning a good descriptive label to a speech signal in these natural contexts is complex: who or what decides what a good descriptive label is for a specific speech signal? These observations have inspired us to employ a more 'indirect' description and annotation of emotion. We decided to focus on emotionally colored phenomena that are somehow related to the expression of emotion. Firstly, we developed detectors for the recognition of laughter. The task of the laughter detector was simply to detect laughter, without *interpreting* the laughter, or without recovering the meaning and function behind the laughter (for example, laughter can be expressed out of politeness or as a reaction to a joke). According to our detection experiments, laughter can be relatively easily discriminated from speech: EERs and C_{det} ranged between 3% and 10%. Secondly, we developed detectors for the recognition of subjective content. The task of the subjectivity detector was to detect subjective clauses, and to detect whether the subjective clause was positively or negatively charged. The underlying assumption was that if somebody has an opinion, he/she will utter it with more affect than when it is a factual statement. Subjectivity and the polarity of subjectiveness was much more difficult to detect: C_{det} was around 26%. One of the reasons why laughter is easier to detect than subjectiveness, could be that laughter is more directly linked to affective behavior than subjectiveness. Subjectiveness comprises a relatively broad concept, that is more expressed through linguistic than acoustic features, and that is only indirectly linked to affective behavior: our assumption that subjectiveness is expressed with more affect has appeared to be a very weak one.

As an intermediate between the use of very artificial, shorter and distinct emotions and very natural, but mostly less distinct and more subtle emotions, we decided to use elicited emotional speech data with gamers who annotated their own felt emotions on emotion dimensions of Arousal and Valence. In Chapter 6, we presented our results of emotion detection experiments carried out with emotional speech data from our own collected TNO-GAMING corpus. Since the data was annotated on con-

tinuous Arousal and Valence dimensions, the task of the recognizer was to estimate real-valued Arousal and Valence scalars. The disadvantage is that the performance of this regression-based method is difficult to compare with our discrete emotion recognition methods. According to the results of our emotion recognition experiments, the person who performs the annotation also plays an important role in the task description and performance of the affect recognizer. We found that the recognition performance is much higher when the data is annotated by observers rather than the gamers themselves. This suggests that it is more difficult to detect *felt* emotions than *expressed* emotions. However, one can imagine that the ultimate goal is not only to detect what is expressed but also to detect what is felt. In addition, our results showed that Arousal can be much better predicted than Valence when only acoustic features are used. Valence was much better modeled than Arousal using lexical features. Further, our emotion prediction method performed better in acted emotional speech than in spontaneous emotional speech, which supports the notion that acted emotional speech is easier to model.

In short, with respect to RQ1 and RQ2, our emotion recognition experiments performed on several sets of emotional speech data that differ in naturalness and emotion description have given us insight in how much the task and performance of an affect recognizer is dependent on the data that it is trained with. In general, acted emotional speech implies expressions of emotions that are full-blown and perhaps even exaggerated, which appear to be easier to recognize than spontaneous emotional expressions. The use of natural emotional speech data usually means that a more subtle description of emotion is required which, most often, leads to emotion descriptions that are less direct and more attuned to human behavior. In real-life, full-blown emotions are replaced with subtle emotionally colored behavior, regulated by social rules, that may or may not be suppressed. It appears that with the current technology, these types of subtle emotions are hard to detect automatically with the exception of laughter, which could be detected with an acceptable accuracy. Furthermore, the task of an affect recognizer is also determined by the person who annotated the data. If the annotator is the same person who has undergone the emotion, the annotations will very likely reflect felt emotions. Our experiments have shown that our current technology is not ready yet to detect felt emotions. If the annotator is not the same person as the one who has undergone the emotional experience, the annotations will reflect the emotions as perceived by the annotator. According to our experiments, perceived emotions are much better to predict than felt emotions.

Our third research question is related to **method and features** used for the development of speech-based affect recognition systems:

Research question 3: What features and modeling techniques can best be used to automatically extract information from the speech signal about the speaker's emotional state?

In Chapters 4, 5, and 6 we have performed emotion recognition experiments in which we systematically compared different sets of features and recognition techniques to each other. In Chapter 4, we worked with acted emotional speech and basic emotions to illustrate the workings of several recognition and fusion techniques for basic

emotion recognition in a detection framework. We made several combinations of features and modeling techniques, and combined the two best performing systems with a linear weighted sum rule or an LDA. The three systems tested were: Standard GMMs using RPLP ('Standard-GMM-rplp'), GMM supervector based SVM using RPLP ('GMM-SV-SVM-rlp'), and an SVM using a set of prosodic features ('SVM-Praat'). The best performing system was GMM-SV-SVM-rplp, followed by SVM-Praat. Hence, these two systems were combined with each other on decision-level which yielded the best performance. Our results obtained were in line with previously reported results obtained with the same data set.

For laughter detection (see Chapter 5), the lowest EERs were achieved with a combination of spectral and prosodic features, and a combination of GMMs and SVMs. For subjectivity and polarity recognition (see Chapter 5), we also employed textual features: we tested word n-grams, character n-grams, phone n-grams and prosodic features. The prosodic features appeared to be less powerful than the textual features. According to our recognition experiments, subjectivity seems to be more apparent in the words used, than in the prosodics. A carefully linear weighted decision-level fusion (rather than an unweighted feature-level fusion) between the separate detectors yielded significantly lower error rates.

Lexical and acoustic features were both also used in the prediction of Arousal and Valence in the speech of gamers, see Chapter 6. Lexical features were shown to be better predictors of Valence information than acoustic features, while acoustic features were better predictors of Arousal information than lexical features. In addition, in a comparison between several types of acoustic features, energy-related features and MFCCs proved to be the best predictors of Arousal and Valence respectively. Although the literature cites pitch as one of the main features that carries information about the speaker's emotional state, our models trained achieve higher performances with the other types of features, i.e. MFCC, INTENS, and ESPECTR.

In short, with respect to RQ3, it is clear that (decision-level) fusion between different types of systems and different types of features can improve recognition performance significantly. Combining generative (GMM) and discriminative (SVM) learning systems has shown to yield significantly better results. Short-term spectral features (typically used in ASR) should be combined with long-term prosodic features that can capture the typical slow varying emotional characteristics in prosody. These type of fusions help, most of the time, because each separate system can provide uncorrelated, complementary information.

Our fourth research question was related to **performance evaluation** of speech-based affect recognition systems:

Research question 4: How can the current evaluation methodology for affect recognition in the lab be improved to match more closely the real-life, field situation in which affect occurs?

In Chapter 4, we proposed to develop and evaluate emotion recognizers in a detection framework, similar to speaker and language recognition, that provides standardized evaluation tools and performance metrics. Although we need to assume discrete emotion categories in order for this evaluation framework to work, we believe that by

adopting this evaluation, emotion recognition technology can profit from the many advantages this type of framework offers. First of all, the detection framework fits the typical emotion recognition problem better (conceptually) than the traditional multi-class classification paradigm: most of the time, it is more realistic (with respect to the application one has in mind) and advantageous to develop binary-class detectors that detect if a person is for instance angry or not, than to force a multiclass classifier to select one of the pre-fixed classes that may be chosen rather arbitrarily. For instance, a multiclass classification experiment based on a database that contains Anger, Sadness, Joy and Neutral, informs us about the *discriminability between* these emotions rather than the *recognizability in general* of these emotions. Results of such multiclass classification experiments are therefore highly dependent on the type and number of emotion classes available in the database. Therefore, we introduced an evaluation scheme that simulates a so-called 'open-set' situation. In this 'open-set' detection evaluation scheme, we tested how good a specific target emotion can be detected without having prior information about the potential non-target emotions available. In other words, for each target emotion, we tested on samples with 'unseen' emotions uttered by 'unseen' speakers. These experiments have shown that Sadness, in the context of the other basic emotions, is a very distinct emotion that is easy to detect, even when there is no prior information available about the potential non-target emotions.

So, how **does** real affect affect affect recognition in speech? The use of real natural affective speech data instead of artificial and acted emotional speech data has a substantial effect on every aspect in the development of a speech-based affect recognizer. Real affect initiates the use of more complex annotation procedures and emotion descriptions. The emotion annotation requires much human labor to reach consensus on emotion labels. Simple category labels do not suffice and are not able to capture the subtlety with which affect is expressed in real-life situations. In natural settings, it is more appropriate to describe affect in terms of Arousal and Valence, or to describe affect in terms of conversational behavior. Real affect triggers the use of different acoustic features: feature selection methods in real affective speech have shown that the type of features selected are different from the ones selected in acted emotional speech. In addition, in real-life, the expression of affect is a multimodal process that involves not only vocal expressions, but also involves, e.g., the choice of words or facial or bodily expressions. For speech- based affect recognition, a combination of lexical and acoustic information will improve recognition performance. Finally, one needs to be aware that shared databases, tasks, and proper evaluation methodologies can help affect recognition technology to advance to a higher level of innovation and performance. Perhaps this can be achieved by adopting existing evaluation methodologies from similar recognition technologies, such as language recognition, like the detection framework.

7.2 Future research

Based on our experiments, we can make some recommendations for future research. It is clear that *real* affect recognition is still a research area under development. First of all, one needs to be more aware of the fact that spontaneous affect involves multi-

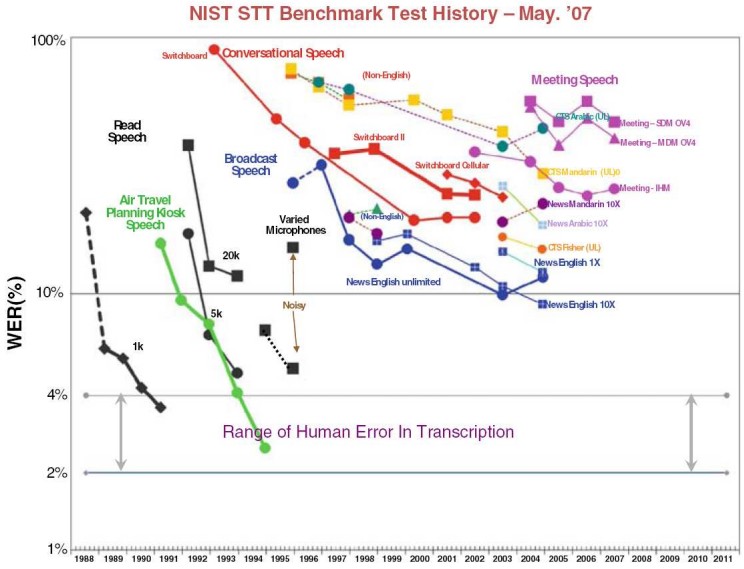


Figure 7.1: *History of NIST Benchmark test, showing general decrease in Word Error Rate (WER) on a logarithmic scale, as a function of date (figure adopted from Fiscus et al. [63]).*

ple modalities and that we do not know yet, how these modalities interact with each other and what the relations between these modalities are. This is especially important when people use multiple modalities to express incongruent emotions, which can occur in real-life. If more is known about these interrelations, we can develop multimodal affect recognizers that can cope with this and detect real multimodal affect. Secondly, in comparison with automatic speech recognition, speaker and language recognition, affect recognition is seemingly much more person dependent. Nowadays, personalization plays an increasingly important role, so it makes sense to investigate how speaker specific methods (e.g., speaker adaptation) can work for affect recognition. Rather than trying to detect a 'universal' concept of affect that is the same for each person, one can also try to fit personalized models to each person who inherently expresses affect differently from another person. Thirdly, one of the main application areas of affect recognition is that of intelligent interactive interfaces: affect recognition can be employed to make man-machine interaction more intelligent and effective. For that purpose, affect should be more investigated in its context in its broadest sense, i.e., affect-in-interaction. For example, dialog acts like agreement or disagreement are also related to affect. Laughter is a beautiful example of affect-in-interaction: it (usually) occurs as a reaction to the person who you are socially interacting with. From a methodology, and technology perspective, affective events (or affect bursts Schröder [167]) like laughter are still interesting events to detect. These events can be relatively distinctively defined, can be relatively good detected, and are important bearers of affect information. Finally, we believe that having shared databases, common tasks and common evaluation protocols will help to advance affect recognition technology. Developing spontaneous affective speech databases is a very time and human labor consuming process. But it would help the affect recogni-

tion community substantially when these databases are also made publicly available. Sharing research tools allows for much easier comparisons and consequently, will lead to increased competition and motivation to develop improved recognition technology.

Affect recognition is a relatively young research area that is gradually advancing towards maturity. Drawing the parallel with the history of automatic speech recognition, we may conclude that affect recognition technology is in the stage where automatic speech recognition technology was about 20 years ago, see Fig. 7.1. About 20 years ago, ASR technology started with the recognition of read digits, and read speech which had a relatively good performance. 20 years later, researchers have moved towards the use of spontaneous speech, recognizing broadcast speech and natural meeting speech, achieving decreased word error rates. Hopefully, in the near future, a benchmark test chart can be created for speech-based affect recognition as well, that shows positive developments in technology and performance.

Bibliography

[1] Kerstens, J. and Ruys, J. E. and Zwarts, J. Ed. URL `http://www2.let.uu.nl/Uil-OTS/Lexicon/`.

[2] NIST. URL `http://www.nist.gov`.

[3] SPRACHcore. URL `http://www.icsi.berkeley.edu/ dpwe/projects/sprach/`.

[4] VicarVision. URL `http://www.vicarvision.nl`.

[5] E. Alpaydin. *Introduction to machine learning*. The MIT Press, Cambridge, Massachusetts, 2004.

[6] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 2037–2040, 2002.

[7] Y. Ang, J. Liu and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1061–1064, 2005.

[8] V. Aubergé, N. Audibert, and A. Rilliard. Auto-annotation: an alternative method to label expressive corpora. In *Proceedings of International Conference on Lanugage Resources and Evaluation (LREC)*, 2006.

[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verifcation systems. *Digital Signal Processing*, 10:42–54, 2000.

[10] J.-A. Bachorowski and M. J. Owren. Not all laughs are alike: voiced but not unvoiced laughter readily elicits positive affect. *Psychologocial Science*, 12:252–257, 2001.

[11] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110:1581–1597, 2001.

[12] R. Banse and K. R. Scherer. Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70:614–636, 1996.

[13] D. Baron, E. Shriberg, and A. Stolcke. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 949–925, 2002.

[14] R. Barra, J. M. Montero, J. Macías-Guarasa, L. F. D'Haro, R. San-Segundo, and R. Córdoba. Prosodic and segmental rubrics in emotion identification. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1085–1088, 2006.

[15] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication*, 40:117–143, 2003.

[16] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. We are not Amused - But how do you know? User states in a multi-modal dialogue system. In *Proceedings of Eurospeech*, pages 733–736, 2003.

[17] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of the International Conference of Language Resources and Evaluation (LREC)*, pages 171–174, 2004.

[18] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. Combining efforts for improving automatic classification of emotional user states. In *Language Technologies (IS-LTC)*, pages 240–245, 2006.

[19] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The impact of F0 extraction errors on the classification of prominence and emotion. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS)*, pages 2201–2204, 2007.

[20] A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl. Mothers, adults, children, pets - towards the acoustics of intimacy. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4497–4500, 2008.

[21] C. Bickley and S. Hunnicutt. Acoustic analysis of laughter. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 927–930, 1992.

[22] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the Institute of Phonetic Sciences*, volume 17, pages 97–110, 1993.

[23] P. Boersma and D. Weenink. Praat: doing phonetics by computer. URL `http://www.praat.org`.

[24] S. Burger, V. MacLaren, and H. Yu. The ISL meeting corpus: the impact of meeting type on speech style. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 301–304, 2002.

[25] F. Burkhardt, A. Paeschke, M. Rolfes, and M. Sendlmeier. A database of German emotional speech. In *Proceedings of Interspeech*, pages 1517–1520, 2005.

[26] C. Busso and S. S. Narayanan. The expression and perception of emotions: Comparing assessments of Self versus Others. In *Proceedings of Interspeech*, pages 257–260, 2008.

[27] R. Cai, L. Lie, H.-J. Zhang, and L.-H. Cai. Highlight sound effects detection in audio stream. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 37–40, 2003.

[28] N. Campbell. The recording of emotional speech: JST/CREST database research. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2002.

[29] N. Campbell. On the use of nonverbal speech sounds in human communication. In *Proceedings of International Workshop on Paralinguistic Speech ParaLing*, 2007.

[30] N. Campbell, H. Kashioka, and R. Ohara. No laughing matter. In *Proceedings of Interspeech.*, pages 465–468, 2005.

[31] W. M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 161–164, 2002.

[32] W. M. Campbell, D. A. Reynolds, and J. P. Campbell. Fusing discriminative and generative methods for speaker recognition: experiments on Switchboard and NFI/TNO field data. In *Proceedings of Odyssey 2004: The Speaker and Language Recognition Workshop*, pages 41–44, 2004.

[33] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311, 2006.

[34] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 97–100, 2006.

[35] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41:181–190, 2007.

[36] O. Cetin and E. Shriberg. Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition. In *Proceedings of Interspeech*, pages 293–296, 2006.

[37] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001. URL `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

[38] F. Chen, A. Li, H. Wang, T. Wang, and Q. Fang. Acoustic analysis of friendly speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 559–572, 2004.

[39] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette. The SAFE corpus: illustrating extreme emotions in dynamic situations. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2006.

[40] C. Clavel, L. Devillers, G. Richard, I. Vasilescu, and T. Ehrette. Detection and analysis of abnormal situations through fear-type acoustic manifestations. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 21–24, 2007.

[41] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

[42] R. Collobert and S. Bengio. SVMTorch: support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[43] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–498, 1998.

[44] R. Cowie and M. Schröder. Piecing together the emotion jigsaw. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 3361/2005 of *Lecture Notes in Computer Science*, pages 305–317. Springer, Berlin/Heidelberg, 2005.

[45] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. FEELTRACE: an instrument for recording perceived emotion in real time. In *Proceedings of ISCA ITRW on Speech and Emotion*, pages 19–24, 2000. URL `http://www.dfki.de/ schroed/feeltrace/`.

[46] C. Cox. Sensitive artificial listener induction techniques. In *HUMAINE Network of Excellence Summer School.*, 2004. URL `http://emotion-research.net/ws/summerschool1/SALAS.ppt259`.

[47] C. Darwin. *The expression of the emotions in man and animals*. John Murray, London,

1872.

[48] D. Datcu and L. J. M. Rothkrantz. The recognition of emotions using Gentleboost Classifier. In *Proceedings of International Conference on Computer Systems and Technologies (CompSysTech)*, 2006.

[49] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28)(4):357–366, 1980.

[50] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[51] M. den Uyl and H. van Kuilenberg. The FaceReader: Online facial expression recognition. In *Proceedings of Measuring Behavior*, pages 598–590, 2005.

[52] L. Devillers and L. Vidrascu. Real-life emotions detections with lexical and paralinguistic cues on human-human call center dialogs. In *Proceedings of Interspeech*, pages 801–804, 2006.

[53] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1–2):33–60, 2003.

[54] E. Douglas-Cowie, L. Devillers, J.-C. Martin, R. Cowie, S. Davvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: facing up to complexity. In *Proceedings of Interspeech*, pages 813–816, 2005.

[55] R. Drullman, J. M. Festen, and R. Plomp. Effect of termporal envelope smearing on speech reception. *Journal of the Acoustical Society of America*, 95:1053–1064, 1994.

[56] P. Ekman. Universals and cultural differences in facial expressions of emotion. In J. Cole, editor, *Nebraska Symposium on Motivation*, pages 207–283. University of Nebraska Press, Lincoln, Nebraska, 1972.

[57] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[58] P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[59] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: categories, origins, usage, and encoding. *Semiotica*, 1:49–98, 1969.

[60] A. El Hannani and D. Petrovska-Delcretaz. Exploiting high-level information provided by ALISP in speaker recognition. In *Proceedings of the Non-Linear Speech Processing Workshop (NOLISP)*, pages 19–24, 2005.

[61] Noldus FaceReader. URL `http://www.noldus.com/site/doc200705001`.

[62] R. Fernandez and R. W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40(1–2):145–159, 2003.

[63] J. G. Fiscus, J. Ajot, and J. S. Garofolo. The Rich Transcription 2007 meeting recognition evaluation. In *The Joint Proceedings of the CLEAR 2007 and RT 2007 Evaluation Workshops*, volume 4625 of *Lecture Notes in Computer Science*, pages 373–389. Springer, 2007.

[64] Y. Freund and R. E. Shapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1990.

[65] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic

dependencies. In *Proceedings of the Meeting of the Association for Computational Linguistics (ACL)*, pages 669–676, 2004.

[66] L. Gillick and S. Cox. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 532–535, 1989.

[67] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar. Combining prosodic, lexical and cepstral systems for deceptive speech detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1033–1036, 2006.

[68] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, pages 787–800, 2007.

[69] M. Grimm, K. Kroschel, and S. Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1085–1088, 2007.

[70] M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German audio-visual emotional speech database. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, 2008.

[71] M. Grootjen, M. A. Neerincx, J. C. M. van Weert, and K. P. Truong. Measuring cognitive task load on a naval ship: Implications of a real world environment. In *Foundations of Augmented Cognition*, volume 4565/2007 of *Lecture Notes in Computer Science*, pages 147–156. Springer, Berlin/Heidelberg, 2007.

[72] S. Hahn, R. Ladner, and M. Ostendorf. Agreement/disagreement classification: exploiting unlabeled data using contrast classifiers. In *Proceedings of the Human Language Technology Conference of the NAACL*, 2006.

[73] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin. Perceptual and acoustic correlates of abnormal voice qualities. *Acta Otolaryngologica*, 90:441–451, 1980.

[74] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7:143–154, 2005.

[75] J. H. L. Hansen and S. E. Bou-Ghazale. Getting started with SUSAS: a speech under simulated and actual stress database. In *Proceedings of Eurospeech*, pages 1743–1746, 1997.

[76] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[77] H. Hermansky. Perceptual Linear Predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87:1738–1752, 1990.

[78] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2:578–589, 1994.

[79] D. Heylen, D. Reidsma, and R. J. F. Ordelman. Annotating state of mind in meeting data. In *Proceedings of the Language Resources and Evaluation Conference (LREC): workshop on corpora for research on emotion and affect*, pages 84–87, 2006.

[80] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies conference*

(HTL-NAACL), 2003.

[81] H. Hu, M.-X. Xu, and W. Wu. GMM supervector based SVM with spectral features for speech emotion recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416, 2007.

[82] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177, 2004.

[83] A. Hua, D. J. Litman, K. Forbes-Riley, M. Rotaru, J. Tetreault, and A. Purandare. Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of Interspeech*, pages 1682–1684, 2006.

[84] A. Ito, W. Xinyue, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Proceedings of International Conference on Cyberworlds*, pages 437–444, 2005.

[85] A. Janin, J. Ang, R. Bhagat, S. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede. The ICSI Meeting project: resources and research. In *NIST ICASSP 2004 Meeting Recognition Workshop*, 2004.

[86] F. Jelinek and R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of Workshop on Pattern Recognition in Practice*, pages 381–397, 1980.

[87] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, 1998.

[88] T. Johnstone. Emotional speech elicited using computer games. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 1985–1988, 1996.

[89] S. Kajarekar, N. Scheffer, M. Gracianera, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet. The SRI NIST 2008 speaker recognition evaluation system. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[90] L. S. Kennedy and D. P. W. Ellis. Laughter detection in meetings. In *Proceedings of NIST ICASSP 2004 Meeting Recognition Workshop*, 2004.

[91] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner. Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Proceedings of Interspeech*, pages 809–812, 2005.

[92] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the International conference on Computational Linguistics (COLING)*, page 1367, 2004.

[93] D. Klein, J. Smarr, H. Nguyen, and C. D. Manning. Named entity recognition with character-level models. In *Proceedings of the conference on Natural language learning at HLT-NAACL*, pages 180–183, 2003.

[94] M. T. Knox and N. Mirghafori. Automatic laughter detection using neural networks. In *Proceedings of Interspeech*, pages 2973–2976, 2007.

[95] M. T. Knox, N. Morgan, and N. Mirghafori. Getting the last laugh: automatic laughter segmentation in meetings. In *Proceedings of Interspeech*, pages 797–800, 2008.

[96] J. Krajewski and B. Kröger. Using prosodic and spectral characteristics for sleepiness detection. In *Proceedings of Interspeech*, pages 1841–1844, 2007.

[97] K. Krippendorff. Reliability in content analysis. *Human Communication Research*, 30(3)):411–433, 2004.

[98] K. Krippendorff. Computing krippendorff's alpha-reliability. URL `http://www.asc.upenn.edu/usr/krippendorff/webreliability.doc`.

[99] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee. Emotion recognition by speech signals. In *Proceedings of Eurospeech*, pages 125–128, 2003.

[100] W. Labov. *Language in the Inner City*. University of Pennsylvania Press, Philadelphia, 1972.

[101] P. J. Lang. *Technology in mental health care delivery systems*, chapter Behavioral treatment and bio-behavioral assessment: computer applications, pages 119–137. Ablex, Norwood, NJ, 1980.

[102] P. J. Lang. The emotion probe - studies of motivation and attention. *American Psychologist*, 50:371–385, 1995.

[103] K. Laskowski and S. Burger. Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2006.

[104] K. Laskowski and T. Schultz. Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 5237 of *Lecture Notes in Computer Science*, pages 149–160. Springer, Berlin/Heidelberg, 2008.

[105] N. Lazarro. Why we play games: 4 keys to more emotion without story. In *Game Developers Conference*, 2004.

[106] S. C. Levinson. *Pragmatics*. Cambridge University Press, United Kingdom, 1983.

[107] R. P. Lippmann, L. Kukolich, and E. Singer. LNKnet: Neural Network, Machine-Learning, and Statistical Software for Pattern Classification. *Lincoln Laboratory Journal*, 6:249–268, 1993.

[108] J. Liscombe, J. Hirschberg, and J. Venditti. Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech*, pages 1837–1840, 2005.

[109] D. J. Litman and K. Forbes-Riley. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590, 2006.

[110] A. Lockerd and F. Mueller. LAFCam - leveraging affective feedback camcorder. In *Proceedings of CHI Conference on Human Factors in Computing Systems*, pages 574–575, 2002.

[111] M. Lugger and B. Yang. The relevance of voice quality features in speaker independent emotion recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 17–20, 2007.

[112] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech*, pages 1895–1898, 1997.

[113] R. Matejka, L. Burget, P. Schwarz, O. Glembek, M. Karafiat, F. Grezl, J. Cernocky, D. A. van Leeuwen, N. Brümmer, and A. Strasheim. STBU System for the NIST 2006 Speaker Recognition Evaluation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 221–224, 2007.

[114] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The AMI Meeting Corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005.

[115] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve. Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of ISCA ITRW on Speech and Emotion*, pages 207–212, 2000.

[116] A. Mehrabian. *Silent Messages*. Wadsworth, Belmont, California, 1971.

[117] W. A. Melder, K. P. Truong, M. den Uyl, D. A. van Leeuwen, M. A. Neerincx, L. R. Loos, and B. S. Plum. Affective multimodal mirror: sensing and eliciting laughter. In *Proceedings of the International workshop on Human-Centered Multimedia (HCM)*, pages 31–40, 2007.

[118] P. P. A. B. Merkx, K. P. Truong, and M. A. Neerincx. Inducing and measuring emotion through a multiplayer first-person shooter computer game. In *Proceedings of the Computer Games Workshop*, 2007.

[119] N. Morgan and H. Bourlard. Automatic speech recognition: An auditory perspective. Springer, 2004.

[120] D. E. Mowrer, L. L. LaPointe, and J. Case. Analysis of five acoustic correlates of laughter. *Journal of Nonverbal Behavior*, 11(3):191–199, 1987.

[121] G. Murray, S. Renals, and J. Carletta. Extractive summarization of meeting recordings. In *Proceedings of Interspeech*, pages 593–596, 2005.

[122] G. Murray, P.-Y. Hsueh, S. Tucker, J. Kilgour, J. Carletta, J. D. Moore, and S. Renals. Automatic segmentation and summarization of meeting speech. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies conference (NAACL-HLT)*, 2007.

[123] I. R. Murray and J. L. Arnott. Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, 93:1097–1108, 1993.

[124] D. Neiberg, K. Elenius, and K. Laskowski. Emotion recognition in spontaneous speech using GMMs. In *Proceedings of Interspeech*, pages 1581–1584, 2006.

[125] T. L. Nwe, S. W. Foo, and L. C. De Silva. Speech emotion recognition using Hidden Markov Models. *Speech Communication*, 41(4):603–623, 2003.

[126] E. E. Nwokah, P. Davies, A. Islam, H.-C. Hsu, and Fogel. A. Vocal affect in three-year-olds: a quantitative acoustic analysis of child laughter. *Journal of the Acoustical Society of America*, 94:3076–3090, 1993.

[127] N. Oostdijk. The Spoken Dutch corpus: overview and first evaluation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 887–894, 2000.

[128] C. E. Osgood. The nature and measurement of meaning. *Psychological Bulletin*, 49(3): 197–237, 1952.

[129] C. E. Osgood, W. H. May, and M. S. Miron. *Cross-Cultural Universals of Affective Meaning*. University of Illinois Press, Urbana, 1975.

[130] S. Petridis and M. Pantic. Audiovisual discrimination between laughter and speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5117–5120, 2008.

[131] S. Petridis and M. Pantic. Fusion of audio and visual cues for laughter detection. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR)*, pages 329–337, 2008.

[132] S. Petridis and M. Pantic. Audiovisual laughter detection based on temporal features. In *Proceedings of the International Conference on Multimodal Interfaces (ICMI)*, pages 37–44, 2008.

[133] V. Petrushin. Emotion in speech: Recognition and application to call centers. In *Proceedings of Artificial Neural Networks in Engineering (ANNIE)*, 1999.

[134] R. W. Picard. *Affective computing*. The MIT Press, Cambridge, Massachusetts, 1997.

[135] C. G. Pittam, C. Gallois, and V. Callan. The long-term spectrum and perceived emotion. *Speech Communication*, 9(3):177–187, 1990.

[136] R. Plutchik. *The Emotions: Facts, Theories, and a New Model*. Random House, New York, 1962.

[137] R. Plutchik. *The Psychology and Biology of Emotion*. Harper Collins, New York, 1994.

[138] R. Plutchik. The nature of emotions. *American Scientist*, 89:344, 2001.

[139] A.-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT)*, pages 339–346, 2005.

[140] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. *A comprehensive grammar of the English Language*. Longman, New York, 1985.

[141] S. Raaijmakers. Sentiment classification with interpolated information diffusion kernels. In *Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising (AD-KDD)*, 2007.

[142] S. Raaijmakers and W. Kraaij. A shallow approach to subjectivity classification. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2008.

[143] S. Raaijmakers, K. P. Truong, and T. Wilson. Multimodal subjectivity analysis of multiparty conversation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 466–474. Association for Computational Linguistics, 2008. URL `http://www.aclweb.org/anthology/D08-1049`.

[144] L. Rabiner and B. Juang. An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3:4–16, 1986.

[145] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, pages 257–286, 1989.

[146] D. Reidsma, D. Heylen, and R. J. F. Ordelman. Annotating emotions in meetings. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, pages 1117–1122, 2006.

[147] B. Reuderink, M. Poel, K. P. Truong, R. W. Poppe, and M. Pantic. Decision-level fusion for audio-visual laughter detection. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 5237 of *Lecture Notes in Computer Science*, pages 137–148. Springer, Berlin/Heidelberg, 2008.

[148] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using Gaussian Mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3 (1):72–83, 1995.

[149] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.

[150] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 105–112, 2003.

[151] A. Rosenberg and J. Hirschberg. Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of Interspeech*, pages 513–516, 2005.

[152] H. Rothganger, G. Hauser, A. C. Cappellini, and A. Guidotti. Analysis of laughter and speech sounds in italian and german students. *Naturwissenschaften*, 85:394–402, 1998.

[153] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 6:1161–1178, 1980.

[154] J. A. Russell, A. Weiss, and G. A. Mendelsohn. Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57(3):493–502, 1989.

[155] J. A. Russell, J.-A. Bachorowski, and J.-M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Pscyhology*, 54:329–349, 2003.

[156] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[157] A. Salway and M. Graham. Extracting information about emotions in films. In *Proceedings of the ACM Conference on Multimedia*, pages 299–302, 2003.

[158] F. Schaeffler, V. Kempe, and S. Biersack. Comparing vocal parameters in spontaneous and posed child-directed speech. In *Proceedings of the 3rd Speech Prosody Congress*, 2006.

[159] K. Scherer. Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99:143–165, 1986.

[160] K. R. Scherer. On the Nature and Function of Emotion: A Component Process Approach. In K. R. Scherer and P. Ekman, editors, *Approaches to Emotion*, pages 293–317. Erlbaum, Hillsdale, NJ, 1984.

[161] K. R. Scherer. Affect bursts. In S. H. M. van Goozen, N. E. van de Poll, and J. A. Sergeant, editors, *Emotions*, pages 161–193. Lawrence Erlbaum, Hillsdale, NJ, 1994.

[162] K. R. Scherer. Psychological models of emotion. In J. C. Borod, editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press, New York, 2000.

[163] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256, 2003.

[164] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[165] H. A. Schlosberg. Three dimensions of emotion. *Psychological Review*, 44:229–237, 1954.

[166] M. Schröder. Emotional speech synthesis - a review. In *Proceedings of Eurospeech*, pages 561–564, 2001.

[167] M. Schröder. Experimental study of affect bursts. *Speech Communication*, 1–2:99–116, 2003.

[168] M. Schröder. Expressing degree of activation in synthetic speech. *IEEE Transactions on*

Audio, Speech, and Language Processing, 14(4):1128–1136, 2006.

[169] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *Proceedings of Eurospeech*, pages 87–90, 2001.

[170] B. Schuller and G. Rigoll. Timing levels in segment-based speech emotion recognition. In *Proceedings of Interspeech*, pages 1695–1698, 2006.

[171] B. Schuller, S. Reiter, R. Müller, M. Al-Hames, M. Lang, and G. Rigoll. Speaker independent speech emotion recognition by ensemble classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 864–867, 2005.

[172] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Proceedings of Interspeech*, pages 2253–2256, 2007.

[173] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl. Towards more reality in the recognition of emotional speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 941–944, 2007.

[174] M. Shami and W. Verhelst. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication*, 49:201–212, 2007.

[175] R. E. Shapire and Y. Singer. BoostTexter: a boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.

[176] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the 5th Workshop on Discourse and Dialogue (SIGdial)*, pages 97–100, 2004.

[177] A. M. C. Sluijter and V. J. van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *Psychological Bulletin*, 100(4):2471–2485, 1996.

[178] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, GMD FIRST, Berlin, Germany, 1998. URL `http://www.svms.org/regression/SmSc98.pdf`. Produced as part of the ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2 27150.

[179] M. J. Smoski and J.-A. Bachorowski. Antiphonal laughter between friends and strangers. *Cognition and Emotion*, 17(2):327–340, 2003.

[180] S. Somasundaran, J. Ruppenhofer, and J. Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the Workshop on Discourse and Dialogue (SIGdial)*, 2007.

[181] S. Steininger, F. Schiel, and A. Glesner. User-state labeling procedures for the multimodal data collection of SmartKom. In *Proceedings of the International Conference of Language Resources and Evaluation (LREC)*, 2002.

[182] R. Tato, R. Santos, R. Kompe, and J. M. Pardo. Emotional space improves emotion recognition. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pages 2029–2032, 2002.

[183] S. S. Tomkins. *Affect, Imagery, Consciousness. The Positive Affects*, volume 1. Springer, New York, 1962.

[184] J. Trouvain. Segmenting phonetic units in laughter. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, pages 2793–2796, 2003.

[185] K. P. Truong and S. Raaijmakers. Automatic recognition of spontaneous emotions in speech using acoustic and lexical features. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 5237/2008 of *Lecture Notes in Computer Science*, pages 161–172. Springer, Berlin/Heidelberg, 2008.

[186] K. P. Truong and D. A. van Leeuwen. Automatic laughter detection. In *Proceedings of Interspeech*, pages 485–488, 2005.

[187] K. P. Truong and D. A. van Leeuwen. Automatic discrimination between laughter and speech. *Speech Communication*, 49:144–158, 2007.

[188] K. P. Truong and D. A. van Leeuwen. An open-set detection evaluation methodology for automatic emotion recognition in speech. In *Proceedings of International Workshop on Paralinguistic Speech (ParaLing)*, pages 5–10, 2007.

[189] K. P. Truong and D. A. van Leeuwen. Visualizing acoustic similarities between emotions in speech: an acoustic map of emotions. In *Proceedings of Interspeech*, pages 2265–2668, 2007.

[190] K. P. Truong and D. A. van Leeuwen. Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features. In *Proceedings of the Workshop on the Phonetics of Laughter*, 2007.

[191] K. P. Truong, M. A. Neerincx, and D. A. van Leeuwen. Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. In *Proceedings of Interspeech*, pages 318–322, 2008.

[192] K. P. Truong, M. A. Neerincx, and D. A. Leeuwen van. Measuring spontaneous vocal and facial emotion expressions in real world environments. In *Proceedings of Measuring Behavior*, 2008.

[193] R. Van Bezooijen. *Characteristics and recognizability of vocal expressions of emotions*. Floris, Dordrecht, 1984.

[194] D. van Leeuwen and M. Huijbregts. The AMI speaker diarization system for NIST RT06s meeting data. In *NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation (RT06s)*, volume 4299 of *Lecture Notes in Computer Science*, pages 371–385. Springer Verlag, Berlin, 2007.

[195] D. A. van Leeuwen and K. P. Truong. An open-set detection evaluation methodology applied to language and emotion recognition. In *Proceedings of Interspeech*, pages 338–341, 2007.

[196] D.A. van Leeuwen and N. Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker Classification I*, volume 4343/2007 of *Lecture Notes in Computer Science*, pages 330–353. Springer, Berlin/Heidelberg, 2007.

[197] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, NY, USA, 1995.

[198] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S, fourth edition*. Springer, 2002.

[199] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, pages 341–344, 2004.

[200] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, 2006.

[201] L. Vidrascu and L. Devillers. Detection of real-life emotions in call centers. In *Proceedings of Interspeech*, pages 1841–1844, 2005.

[202] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Combining frame and turn-level information for robust recognition of emotions within speech. In *Proceedings of Interspeech*, pages 2249–2252, 2007.

[203] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 474–477, 2005.

[204] J. Wagner, J. Kim, and E. André. From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 940–943, 2005.

[205] D. Watson, L. A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[206] C. E. Williams and K. N. Stevens. On determining the emotional state of pilots during flight: an exploratory study. *Aerospace Medicine*, 40:1369–1372, 1969.

[207] C. E. Williams and K. N. Stevens. Emotions and speech: some acoustical correlates. *Journal of the Acoustical Society of America*, 52:1238–1250, 1972.

[208] T. Wilson. Annotating subjective content in meetings. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2008.

[209] T. Wilson and S. Raaijmakers. Comparing word, character, and phoneme *n*-grams for subjective utterance recognition. In *Proceedings of Interspeech*, pages 1614–1617, 2008.

[210] J. Wilting, E. Krahmer, and M. Swerts. Real vs. acted emotional speech. In *Proceedings of Interspeech*, pages 1093–1096, 2006.

[211] B. Wrede and E. Shriberg. Spotting "hot spots" in meetings: Human judgements and prosodic cues. In *Proceedings of Eurospeech*, pages 2805–2808, 2003.

[212] W. Wundt. *Grundzüge der physiologischen Psychologie [Fundamentals of Physiological Psychology]*. Engelmann, Leipzig, fifth edition, 1905.

[213] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen. Two-stage classification of emotional speech. In *Proceedings of International Conference on Digital Telecommunications*, pages 32–37, 2006.

[214] S. Yacoub, X. Simske, S. Lin, and J. Burns. Recognition of emotions in interactive voice response systems. In *Proceedings of Eurospeech*, pages 729–732, 2003.

[215] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, and C. Busso. An acoustic study of emotions expressed in speech. In *Proceedings of Interspeech*, pages 2193–2196, 2004.

[216] S. Yildirim, C. M. Lee, and S. Lee. Detecting politeness and frustration state of a child in a conversational computer game. In *Proceedings of Interspeech*, pages 2209–2212, 2005.

[217] T. Yingthawornsuk, H. Kaymaz Keskinpala, D. M. Wilkes, R. G. Shiavi, and M. Salomon. Direct acoustic feature using iterative EM algorithm and spectral energy for classifying suicidal speech. In *Proceedings of Interspeech*, pages 766–769, 2007.

[218] H. Yu and V. Hatzivassiloglou. Towards answering opinion questions: separating facts

from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 129–136, 2003.

[219] G. Zhou, J. H. L. Hansen, and J. F. Kaiser. Nonlinear features based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3):201–216, 2001.

[220] M. Zimmermann, A. Stolcke, and E. Shriberg. Toward joint segmentation and classification of dialog acts in multiparty meetings. In *Machine Learning for Multimodal Interaction (MLMI)*, volume 3869/2006 of *Lecture Notes in Computer Science*, pages 187–193. Springer, Berlin/Heidelberg, 2006.

Summary

The automatic analysis of affect is a relatively new and challenging multidisciplinary research area that has gained a lot of interest over the past few years. The research and development of affect recognition systems has opened many opportunities for improving the interaction between man and machine. Although affect can be expressed through multimodal means like hand gestures, facial expressions, and body postures, this dissertation has focused on speech (i.e., vocal expressions) as the main carrier of affect. Speech carries a lot of 'hidden' information. By hearing a voice only, humans can guess *who* is speaking, what *language* he/she is speaking (or accent or dialect), what *age* he/she is etc. The goal of automatic speech recognition (ASR) is to recognize *what* is said. In automatic speech-based emotion recognition, the goal is to recognize *how* something is said. In this work, several experiments are described which were carried out to investigate how affect can be automatically recognized in speech.

One of the first steps in developing speech-based affect recognizers involves finding a spontaneous speech corpus that is labeled with emotions. Machine learning techniques, that are often used to build these recognizers, require these data to learn how to associate specific speech features (e.g., pitch, energy) with certain emotions. However, collecting and labeling real affective speech data has appeared to be difficult. Efforts to collecting affective speech data in the field have been described in this work.

As an alternative, speech corpora that contain acted emotional speech (actors are asked to portray certain emotions) have often been used. Advantages of these corpora are that the recording conditions can be controlled, the emotions portrayed can be clearly associated with an emotion label, the costs and effort required to collect such corpora are relatively low, and the recordings are usually made available to the research community. In this work, an acted emotional speech corpus (containing basic, universal emotions like Anger, Boredom, Disgust, Fear, Happiness, Neutral, and Sadness) was used to explore and apply recognition techniques and evaluation frameworks, adopted from similar research areas like automatic speaker and language recognition, to automatic emotion recognition. Recognizers were evaluated in a detection framework, and an evaluation for handling so-called 'out-of-set' emotions (unknown emotions that were not present in the training data, but which can occur in real-life situations) was presented. Partly due to lack of standardization and shared databases, the evaluation of affect recognizers remains somewhat problematic. While evaluation is an important aspect in development, it has been a relatively underexposed topic of investigation in the emotion research community.

The main objections against the use of acted emotional speech corpora are that the expressions are not 'real' but rather portrayals of prototype emotions (and hence, expressed rather exaggeratedly), and the emotions portrayed do not often occur in real life situations. Therefore, in this work, spontaneous data has also been used and methods were developed to recognize spontaneous, vocal expressions of affect, like laughter. The task of the laughter detector was to recognize audible laughter in meeting speech data. Using a combination of Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs), and a combination of prosodic and spectral speech features, relatively low error rates between 3%–12% were achieved. Although the detector did not interpret the affective meaning of the laughter, the detection of laughter alone was informative enough. Part of these findings were used to build a so-called 'Affective Mirror' that successfully elicited and recognized laughter with different user groups.

Other speech phenomena related to vocal expressions of affect, also in the context of meeting speech data, are the expressions of opinions and sentiments. In this work, it was assumed that opinions are expressed differently from factual statements in terms of tone of voice, and the words used. Classification experiments were carried out to find the best combination of lexical and prosodic features for the discrimination between subjective and non-subjective clauses. As lexical features, word-level, phone-level, and character-level n-grams were used. The experiments showed that a combination of all features yields the best performances, and that the prosodic features were the weakest of all features investigated. In addition, a second task was formulated, namely the discrimination between positive subjective clauses and negative subjective clauses. Similar results for this task were found. The relatively high error rates for both tasks, $C_{\text{det}} = 26\%$–30%, indicate that these are more difficult recognition problems than that of laughter: the relation between prosodic and lexical features, and subjectivity and polarity (i.e., positive vs. negative), is not as clear as is in the case of laughter.

As an intermediate between real affective expressions and acted expressions, elicited affective expressions were employed in this dissertation in several human perception and classification experiments. To this end, a multimodal corpus with elicited affect was recorded. Affective vocal and facial expressions were elicited via a multiplayer first-person shooter video game (Unreal Tournament) that was manipulated by the experimenter. These expressions were captured by close-talk microphones and high-quality webcams, and were afterwards rated by the players themselves on Arousal (active-passive) and Valence (positive-negative) scales. After post-processing the data, perception and classification experiments were carried out on this data. The first experiment carried out with this unique kind of data tried to answer the question how the level of agreement between observers on the perceived emotion is affected when audio-only, video-only, audiovisual, or audiovisual + context information clips containing affective expressions are shown. The observers were asked to rate each clip on Arousal and Valence scales. The results showed that the agreement among human observers was highest when audiovisual clips were shown. Furthermore, the observers reached higher agreement on Valence judgments than Arousal judgments. Additionally, the results indicated that the 'self'-ratings of the gamers

themselves differed somewhat from the 'observed'-ratings of the human observers. This finding was further investigated in a second experiment. Six raters re-annotated a substantial part of the corpus. The results confirmed that there is a discrepancy between what the 'self'-raters (i.e., the gamers themselves) experienced/felt and what observers perceive based on the gamers' vocal and facial expressions. This finding has consequences for the development of automatic affect analyzers that use these ratings: the goal of affect analyzers can be to recognize 'felt' affect, or to recognize 'observed/perceived' affect. Two different types of speech-based affect recognizers were developed in parallel to recognize either 'felt' or 'perceived' affect on continuous Arousal and Valence scales. The results showed that 'felt' emotions are much harder to predict than 'perceived' emotions. Although these recognizers performed moderately from a classification perspective, the recognizers did not perform too bad in comparison to human performance. The recognizers developed depend much on how the affect data is rated by humans; if this data reflects moderate human judgments of affect, then it can be difficult for the machine to perform well (in an absolute sense).

The work presented in this dissertation shows that the automatic recognition of affect in speech is complicated by the fact that real affect, as encountered in real-life situations, is a very complex phenomenon that sometimes cannot be described straightforwardly in ways that can be useful for computer scientists (who would like to build affect recognizers). The use of real affect data has led to the development of recognizers that are more targeted toward affect-related expressions. Laughter and subjectivity are examples of such affect-related expressions. The Arousal and Valence descriptors offer a nice way to describe the meaning of these affective expressions. The relatively high error rates obtained for Arousal and Valence prediction, suggest that the acoustic correlates used in this research only partly capture the characteristics of real affective speech. The search for stronger acoustic correlates or vocal profiles for specific emotions continues. This search is partly complicated by the 'noise' that comes with real affect which remains a challenge for the research community working toward automatic affect analyzers.

Samenvatting

The automatische analyse van emotie herkenning is een relatief jong en uitdagend multidisciplinair onderzoeksgebied waar de laatste jaren veel interesse voor is. Het onderzoek de ontwikkeling van systemen die emoties kunnen herkennen maken innovatieve applicaties mogelijk die als doel hebben de interactie tussen mens en machine te verbeteren. Hoewel emoties via verschillende modaliteiten getoond kunnen worden, bijvoorbeeld via handgebaren, gezichtsexpressies, en lichaamshoudingen, wordt er in deze dissertatie gefocused op spraak (de stem). In spraak zit veel 'verborgen' informatie. Aan iemands spraak kunnen mensen vaak een schatting maken van *wie* er aan het woord is, in welke *taal* of *dialect* er gesproken wordt, wat de *leeftijd* is van de spreker etc. Het doel van automatische spraakherkenning (ASH) is te herkennen *wat* er gezegd wordt. Het doel van spraak-gebaseerde emotieherkenning is te herkennen *hoe* iets gezegd wordt. In deze dissertatie worden enkele experimenten beschreven die uitgevoerd zijn om te onderzoeken hoe emotie in spraak automatisch herkend kan worden.

Een van de eerste stappen in de ontwikkeling van een spraak-gebaseerde emotieherkenningssysteem is het verkrijgen van een spontane spraakdatabase die gelabeled is op emotie. De algoritmes die vaak gebruikt worden om de herkenners te ontwikkelen hebben deze gelabelde data nodig om te leren hoe bepaalde spraak (bijv. toonhoogte, amplitude) elementen geassocieerd kunnen worden met specifieke emoties. Helaas is gebleken dat het verzamelen en het labelen van spontane emotionele spraak een moeizaam en complex proces is. In deze dissertatie zijn enkele inspanningen om spontane emotionele spraak op te nemen en te labelen beschreven.

Als een alternatief worden er ook vaak spraakdatabases gebruikt die geacteerde emotionele spraak bevatten (acteurs worden gevraagd om bepaalde emoties uit te beelden). Aan het gebruik van dit soort databases zitten duidelijke voordelen: de opnames vinden plaats in een gecontroleerde omgeving, de uitgebeelde emoties zijn makkelijk te labelen, de inspanningen om een dergelijke database op te zetten zijn relatief laag, en de opnames kunnen meestal beschikbaar gemaakt worden voor de onderzoeksgemeenschap. In dit werk is ook gebruik gemaakt van een geacteerde emotionele spraakdatabase (deze bevatte spraak uitgesproken in de basis en universele emoties Boosheid, Verveling, Walging, Blijheid, Neutraal, en Droevigheid) om herkenningstechnieken en evaluatie schema's uit soortgelijke onderzoeksgebieden, zoals automatische spreker- en taalherkenning, toe te passen op emotieherkenning. Herkenners werden geevalueerd in een detectie schema, en een evaluatie schema die rekening houdt met 'verrassings emoties' (onbekende emoties die niet aanwezig waren in de trainingsdatabase, en dus niet gemodelleerd zijn, maar die wel kunnen voorkomen

in de werkelijkheid) werd in dit werk gepresenteerd. Deels door het ontbreken aan standaardisering en gedeelde datasets is de evaluatie van emotieherkenningssystemen enigszins problematisch. Terwijl de evaluatie van emotieherkenningssystemen een belangrijk onderdeel is van de ontwikkeling van deze systemen, is dit een relatief onderbelicht onderwerp gebleven in de onderzoeksgemeenschap.

De belangrijkste bezwaren tegen het gebruik van geacteerde emotionele spraak zijn dat de expressies niet 'echt' zijn maar stereotype uitingen van prototype emoties (en mogelijk dus overdreven uitgebeeld), en dat deze uitgebeelde emoties niet vaak voorkomen in de werkelijkheid. Daarom is er in deze dissertatie ook gebruik gemaakt van spontane data, en zijn er methoden ontwikkeld voor de herkenning van spontane, nonverbale expressies van emotie, zoals gelach. De taak van de lach detector was het herkennen van gelach in vergaderingen. Een combinatie van Gaussian Mixture Models (GMMs) en Support Vector Machines (SVMs), en een combinatie van prosodische en spectrale spraakfeatures resulteerden in relatief lage fouten percentages van tussen 3%–12%. Hoewel de lach detector de emotionele betekenis van het gelach niet interpreteerde, was het detecteren van het gelach alleen al informatief genoeg. Een gedeelte van deze bevindingen werd gebruikt voor het bouwen van een 'Affective Mirror' die gelach herkent en uitlokt bij gebruikers.

Andere spraakuitingen die emotioneel geladen kunnen zijn, zijn expressies van opinies en sentiment in de context van vergaderingen. In dit werk is de aanname gedaan dat opinies (subjectiviteit) anders worden 'gebracht' dan feiten in termen van de wijze waarop een uiting wordt uitgesproken en de woorden die gebruikt worden. Classificatie experimenten werden uitgevoerd om te onderzoeken welke combinatie van lexicale en prosodische features de beste prestatie leverde in het onderscheiden van subjectieve en niet-subjectieve uitingen. Als lexicale features werden woord, foneem, en letter n-grammen gebruikt. De experimenten lieten zien dat een combinatie van alle features leidde tot de beste prestatie, en dat de prosodische features de zwakste groep was die weinig onderscheidend vermogen toonde. De tweede taak bestond uit de discriminatie tussen positieve subjectieve en negatieve subjectieve uitingen. Soortgelijke resultaten werden verkregen. De relatief hoge fouten percentages voor beide taken, $C_{\mathrm{det}} = 26\%$–30%, geven aan dat dit complexere herkenningstaken zijn dan die van gelach: de relatie tussen prosodische en lexicale features, en subjectiviteit en polariteit (positief vs. negatief) is minder sterk dan in het geval van gelach.

Als een tussenliggend alternatief tussen spontane en geacteerde emotie uitingen werd in deze dissertatie uitgelokte emotionele uitingen gebruikt in verschillende perceptie en classificatie experimenten. Hiervoor werd een multimodale database met uitgelokte emotie uitingen opgenomen. Affectieve spraakuitingen en gezichtsexpressies werden uitgelokt door mensen een 'multiplayer first-person shooter' video game (Unreal Tournament) te laten spelen die gemanipuleerd was door de onderzoeker. Alles werd opgenomen met behulp van microfoontjes en webcams, en achteraf werden alle opgenomen spraak en gezichtsexpressies beoordeeld op Arousal (actief-passief) en Valence (positief-negatief) schalen door de spelers zelf. Het eerste experiment uitgevoerd met deze unieke dataset richtte zich op het beantwoorden van de vraag hoe mensen affectieve uitingen beoordelen, en in hoeverre mensen het met

elkaar eens zijn over de waargenomen affectieve uitingen wanneer deze getoond worden in verschillende condities: alleen audio, alleen video, audiovisueel of audiovisueel+context informatie. De mensen (=de observeerders) moesten elke clip beoordelen op een Arousal en Valence schaal. De resultaten lieten zien dat mensen het meer met elkaar eens zijn in de audiovisuele conditie, en mensen zijn het meer met elkaar eens over de Valence beoordelingen dan de Arousal beoordelingen. Verder lieten de resultaten zien dat de 'zelf'-beoordelingen (van de spelers zelf) wat verschilden van de 'waargenomen' beoordelingen van de observeerders. Dit resultaat werd verder onderzocht in een tweede experiment. Zes andere observeerders hebben een deel van de database opnieuw beoordeeld op Arousal en Valence schalen. De resultaten bevestigden dat de 'zelf'-oordelen van de spelers die aangaven wat ze voelden, verschillen van de waargenomen oordelen van de observeerders, gebaseerd op de spraak en de gezichtsexpressies van de spelers. Dit gegeven heeft gevolgen voor de ontwikkeling van automatische emotieherkenners die deze Arousal en Valence oordelen gebruiken: het doel van een automatische emotieherkenner kan zijn het herkennen van 'gevoelde' emotie, of het herkennen van 'waargenomen' emotie. Twee spraak-gebaseerde emotieherkenners werden in parallel ontwikkeld die als doel hadden of 'gevoelde' emotie of 'waargenomen' emotie te herkennen op continue Arousal en Valence schalen. The resultaten lieten zien dat 'gevoelde' emoties veel moeilijker te herkennen zijn dan 'waargenomen' emoties. Hoewel de herkenners niet erg goed presteerden vanuit een classificatie oogpunt, presteerden ze in vergelijking met menselijke prestaties redelijk. De prestatie van de ontwikkelde herkenners hangt af van hoe de affectieve data beoordeeld is door mensen; als mensen weinig overeenstemming hebben bereikt over de affectieve data, dan kan het moeilijk voor de herkenner zijn om goed te presteren (absoluut gezien).

Het werk dat gepresenteerd is in deze dissertatie laat zien dat de automatische herkenning van emotie in spraak bemoeilijkt wordt door het feit dat 'echte' emotie, zoals het voorkomt in de werkelijkheid, een erg complex fenomeen is dat soms niet op zo'n eenduidige manier beschreven kan worden dat het ook bruikbaar is voor informatici (die automatische emotieherkenningssystemen willen bouwen). Het gebruik van spontane affectieve data heeft geleid tot de ontwikkeling van herkenners die meer gefocused zijn op emotie-gerelateerde expressies. Gelach en subjectiviteit zijn voorbeelden van zulke emotie-gerelateerde uitingen. De Arousal en Valence schalen bieden een flexibele manier aan om de betekenis van zulke emotieuitingen te beschrijven. De relatief hoge fouten percentages suggereren dat de akoestische correlaten die gebruikt zijn in dit onderzoek niet sterk en volledig genoeg waren om alle karakteristieken van emotionele spraak te beschrijven. De zoektocht naar sterker akoestische correlaten of akoestische profielen voor bepaalde emoties gaat door. Deze zoektocht wordt enigszins bemoeilijkt door de 'ruis' die erbij komt kijken wanneer echte, spontane affectieve data wordt gebruikt wat een een uitdaging blijft voor onderzoekers die hun onderzoek wijden aan het ontwikkelen van automatische emotieherkenners.

SIKS dissertation series

Since 1998, all dissertations written by Ph.D. students who have conducted their research under auspices of a senior research fellow of the SIKS research school are published in the SIKS Dissertation Series. This thesis is the 231st in the series.

2009-33 Khiet Truong (UT), *How Does Real Affect Affect Affect Recognition In Speech?*

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU), *Architectural Knowledge Management: Supporting Architects and Auditors*

2009-31 Sofiya Katrenko (UVA), *A Closer Look at Learning Relations from Text*

2009-30 Marcin Zukowski (CWI), *Balancing vectorized query execution with bandwidth-optimized storage*

2009-29 Stanislav Pokraev (UT), *Model-Driven Semantic Integration of Service-Oriented Applications*

2009-27 Christian Glahn (OU), *Contextual Support of social Engagement and Reflection on the Web*

2009-26 Fernando Koch (UU), *An Agent-Based Model for the Development of Intelligent Mobile Services*

2009-25 Alex van Ballegooij (CWI), *RAM: Array Database Management through Relational Mapping*

2009-24 Annerieke Heuvelink (VUA), *Cognitive Models for Training Simulations*

2009-23 Peter Hofgesang (VU), *Modelling Web Usage in a Changing Environment*

2009-22 Pavel Serdyukov (UT), *Search For Expertise: Going beyond direct evidence*

2009-21 Stijn Vanderlooy (UM), *Ranking and Reliable Classification*

2009-20 Bob van der Vecht (UU), *Adjustable Autonomy: Controling Influences on Decision Making*

2009-19 Valentin Robu (CWI), *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*

2009-18 Fabian Groffen (CWI), *Armada, An Evolving Database System*

2009-17 Laurens van der Maaten (UvT), *Feature Extraction from Visual Data*

2009-16 Fritz Reul (UvT), *New Architectures in Computer Chess*

2009-15 Rinke Hoekstra (UVA), *Ontology Representation - Design Patterns and Ontologies that Make Sense*

2009-14 Maksym Korotkiy (VU), *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*

2009-13 Steven de Jong (UM), *Fairness in Multi-Agent Systems*

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin), *Operating Guidelines for Services*

2009-11 Alexander Boer (UVA), *Legal Theory, Sources of Law & the Semantic Web*

2009-10 Jan Wielemaker (UVA), *Logic programming for knowledge-intensive interactive applications*

2009-09 Benjamin Kanagwa (RUN), *Design, Discovery and Construction of Service-oriented Systems*

2009-08 Volker Nannen (VU), *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*

2009-07 Ronald Poppe (UT), *Discriminative Vision-Based Recovery and Recognition of Human Motion*

2009-06 Muhammad Subianto (UU), *Understanding Classification*

2009-05 Sietse Overbeek (RUN), *Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality*

2009-04 Josephine Nabukenya (RUN), *Improving the Quality of Organisational Policy Making using Collaboration Engineering*

2009-03 Hans Stol (UvT), *A Framework for Evidence-based Policy Making Using IT*

2009-02 Willem Robert van Hage (VU), *Evaluating Ontology-Alignment Techniques*

2009-01 Rasa Jurgelenaite (RUN), *Symmetric Causal Independence Models*

2008-35 Ben Torben Nielsen (UvT), *Dendritic morphologies: function shapes structure*

2008-34 Jeroen de Knijf (UU), *Studies in Frequent Tree Mining*

2008-33 Frank Terpstra (UVA), *Scientific Workflow Design; theoretical and practical issues*

2008-32 Trung H. Bui (UT), *Toward Affective Dialogue Management using Partially Observable Markov Decision Processes*

2008-31 Loes Braun (UM), *Pro-Active Medical Information Retrieval*

2008-30 Wouter van Atteveldt (VU), *Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content*

2008-29 Dennis Reidsma (UT), *Annotations and Subjective Machines – Of Annotators, Embodied Agents, Users, and Other Humans*

2008-29 Dennis Reidsma (UT), *Annotations and Subjective Machines – Of Annotators, Embodied Agents, Users, and Other Humans*

2008-28 Ildiko Flesch (RUN), *On the Use of Independence Relations in Bayesian Networks*

2008-27 Hubert Vogten (OU), *Design and Implementation Strategies for IMS Learning Design*

2008-26 Marijn Huijbregts (UT), *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*

2008-25 Geert Jonker (UU), *Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency*

2008-24 Zharko Aleksovski (VU), *Using background knowledge in ontology matching*

2008-23 Stefan Visscher (UU), *Bayesian network models for the management of ventilator-associated pneumonia*

2008-22 Henk Koning (UU), *Communication of IT-Architecture*

2008-21 Krisztian Balog (UVA), *People Search in the Enterprise*

2008-20 Rex Arendsen (UVA), *Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven.*

2008-19 Henning Rode (UT), *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*

2008-18 Guido de Croon (UM), *Adaptive Active Vision*

2008-17 Martin Op 't Land (TUD), *Applying Architecture and Ontology to the Splitting and Allying of Enterprises*

2008-16 Henriëtte van Vugt (VU), *Embodied agents from a user's perspective*

2008-15 Martijn van Otterlo (UT), *The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains.*

2008-14 Arthur van Bunningen (UT), *Context-Aware Querying; Better Answers with Less Effort*

2008-13 Caterina Carraciolo (UVA), *Topic Driven Access to Scientific Handbooks*

2008-12 József Farkas (RUN), *A Semiotically Oriented Cognitive Model of Knowledge Representation*

2008-11 Vera Kartseva (VU), *Designing Controls for Network Organizations: A Value-Based Approach*

2008-10 Wauter Bosma (UT), *Discourse oriented summarization*

2008-09 Christof van Nimwegen (UU), *The paradox of the guided user: assistance can be counter-effective*

2008-08 Janneke Bolt (UU), *Bayesian Networks: Aspects of Approximate Inference*

2008-07 Peter van Rosmalen (OU), *Supporting the tutor in the design and support of adaptive e-learning*

2008-06 Arjen Hommersom (RUN), *On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective*

2008-05 Bela Mutschler (UT), *Modeling and simulating causal dependencies on process-aware information systems from a cost perspective*

2008-04 Ander de Keijzer (UT), *Management of Uncertain Data – towards unattended integration*

2008-03 Vera Hollink (UVA), *Optimizing hierarchical menus: a usage-based approach*

2008-02 Alexei Sharpanskykh (VU), *On Computer-Aided Methods for Modeling and Analysis of Organizations*

2008-01 Katalin Boer-Sorbán (EUR), *Agent-Based Simulation of Financial Markets: A modular, continuous-time approach*

2007-25 Joost Schalken (VU), *Empirical Investigations in Software Process Improvement*

2007-24 Georgina Ramírez Camps (CWI), *Structural Features in XML Retrieval*

2007-23 Peter Barna (TUE), *Specification of Application Logic in Web Information Systems*

2007-22 Zlatko Zlatev (UT), *Goal-oriented design of value and process models from patterns*

2007-21 Karianne Vermaas (UU), *Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005*

2007-20 Slinger Jansen (UU), *Customer Configuration Updating in a Software Supply Network*

2007-19 David Levy (UM), *Intimate relationships with artificial partners*

2007-18 Bart Orriëns (UvT), *On the development an management of adaptive business collaborations*

2007-17 Theodore Charitos (UU), *Reasoning with Dynamic Networks in Practice*

2007-16 Davide Grossi (UU), *Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems*

2007-15 Joyca Lacroix (UM), *NIM: a Situated Computational Memory Model*

2007-14 Niek Bergboer (UM), *Context-Based Image Analysis*

2007-13 Rutger Rienks (UT), *Meetings in Smart Environments; Implications of Progressing Technology*

2007-12 Marcel van Gerven (RUN), *Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty*

2007-11 Natalia Stash (TUE), *Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System*

2007-10 Huib Aldewereld (UU), *Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols*

2007-09 David Mobach (VU), *Agent-Based Mediated Service Negotiation*

2007-08 Mark Hoogendoorn (VU), *Modeling of Change in Multi-Agent Organizations*

2007-07 Nataša Jovanović (UT), *To Whom It May Concern – Addressee Identification in Face-to-Face Meetings*

2007-06 Gilad Mishne (UVA), *Applied Text Analytics for Blogs*

2007-05 Bart Schermer (UL), *Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance*

2007-04 Jurriaan van Diggelen (UU), *Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach*

2007-03 Peter Mika (VU), *Social Networks and the Semantic Web*

2007-02 Wouter Teepe (RUG), *Reconciling Information Exchange and Confidentiality: A Formal Approach*

2007-01 Kees Leune (UvT), *Access Control and Service-Oriented Architectures*

2006-28 Börkur Sigurbjörnsson (UVA), *Focused Information Access using XML Element Retrieval*

2006-27 Stefano Bocconi (CWI), *Vox Populi: generating video documentaries from semantically annotated media repositories*

2006-26 Vojkan Mihajlović (UT), *Score Region Algebra: A Flexible Framework for Structured Information Retrieval*
2006-25 Madalina Drugan (UU), *Conditional log-likelihood MDL and Evolutionary MCMC*
2006-24 Laura Hollink (VU), *Semantic Annotation for Retrieval of Visual Resources*
2006-23 Ion Juvina (UU), *Development of Cognitive Model for Navigating on the Web*
2006-22 Paul de Vrieze (RUN), *Fundaments of Adaptive Personalisation*
2006-21 Bas van Gils (RUN), *Aptness on the Web*
2006-20 Marina Velikova (UvT), *Monotone models for prediction in data mining*
2006-19 Birna van Riemsdijk (UU), *Cognitive Agent Programming: A Semantic Approach*
2006-18 Valentin Zhizhkun (UVA), *Graph transformation for Natural Language Processing*
2006-17 Stacey Nagata (UU), *User Assistance for Multitasking with Interruptions on a Mobile Device*
2006-16 Carsten Riggelsen (UU), *Approximation Methods for Efficient Learning of Bayesian Networks*
2006-15 Rainer Malik (UU), *CONAN: Text Mining in the Biomedical Domain*
2006-14 Johan Hoorn (VU), *Software Requirements: Update, Upgrade, Redesign – towards a Theory of Requirements Change*
2006-13 Henk-Jan Lebbink (UU), *Dialogue and Decision Games for Information Exchanging Agents*
2006-12 Bert Bongers (VU), *Interactivation – Towards an e-cology of people, our technological environment, and the arts*
2006-11 Joeri van Ruth (UT), *Flattening Queries over Nested Data Types*
2006-10 Ronny Siebes (VU), *Semantic Routing in Peer-to-Peer Systems*
2006-09 Mohamed Wahdan (UM), *Automatic Formulation of the Auditor's Opinion*
2006-08 Eelco Herder (UT), *Forward, Back and Home Again – Analyzing User Behavior on the Web*
2006-07 Marko Smiljanic (UT), *XML schema matching – balancing efficiency and effectiveness by means of clustering*
2006-06 Ziv Baida (VU), *Software-aided Service Bundling – Intelligent Methods & Tools for Graphical Service Modeling*
2006-05 Cees Pierik (UU), *Validation Techniques for Object-Oriented Proof Outlines*
2006-04 Marta Sabou (VU), *Building Web Service Ontologies*
2006-03 Noor Christoph (UVA), *The role of metacognitive skills in learning to solve problems*
2006-02 Cristina Chisalita (VU), *Contextual issues in the design and use of information technology in organizations*
2006-01 Samuil Angelov (TUE), *Foundations of B2B Electronic Contracting*
2005-21 Wijnand Derks (UT), *Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics*
2005-20 Cristina Coteanu (UL), *Cyber Consumer Law, State of the Art and Perspectives*
2005-19 Michel van Dartel (UM), *Situated Representation*
2005-18 Danielle Sent (UU), *Test-selection strategies for probabilistic networks*
2005-17 Boris Shishkov (TUD), *Software Specification Based on Re-usable Business Components*
2005-16 Joris Graaumans (UU), *Usability of XML Query Languages*
2005-15 Tibor Bosse (VU), *Analysis of the Dynamics of Cognitive Processes*
2005-14 Borys Omelayenko (VU), *Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics*
2005-13 Fred Hamburg (UL), *Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen*
2005-12 Csaba Boer (EUR), *Distributed Simulation in Industry*
2005-11 Elth Ogston (VU), *Agent Based Matchmaking and Clustering – A Decentralized Approach to Search*
2005-10 Anders Bouwer (UVA), *Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments*
2005-09 Jeen Broekstra (VU), *Storage, Querying and Inferencing for Semantic Web Languages*
2005-08 Richard Vdovjak (TUE), *A Model-driven Approach for Building Distributed Ontology-based Web Applications*
2005-07 Flavius Frasincar (TUE), *Hypermedia Presentation Generation for Semantic Web Information Systems*
2005-06 Pieter Spronck (UM), *Adaptive Game AI*
2005-05 Gabriel Infante-Lopez (UVA), *Two-Level Probabilistic Grammars for Natural Language Parsing*
2005-04 Nirvana Meratnia (UT), *Towards Database Support for Moving Object data*
2005-03 Franc Grootjen (RUN), *A Pragmatic Approach to the Conceptualisation of Language*
2005-02 Erik van der Werf (UM)), *AI techniques for the game of Go*
2005-01 Floor Verdenius (UVA), *Methodological Aspects of Designing Induction-Based Applications*
2004-20 Madelon Evers (Nyenrode), *Learning from Design: facilitating multidisciplinary design teams*
2004-19 Thijs Westerveld (UT), *Using generative probabilistic models for multimedia retrieval*
2004-18 Vania Bessa Machado (UvA), *Supporting the Construction of Qualitative Knowledge Models*
2004-17 Mark Winands (UM), *Informed Search in Complex Games*
2004-16 Federico Divina (VU), *Hybrid Genetic Relational Search for Inductive Learning*
2004-15 Arno Knobbe (UU), *Multi-Relational Data Mining*
2004-14 Paul Harrenstein (UU), *Logic in Conflict. Logical Explorations in Strategic Equilibrium*
2004-13 Wojciech Jamroga (UT), *Using Multiple Models of Reality: On Agents who Know how to Play*
2004-12 The Duy Bui (UT), *Creating emotions and facial expressions for embodied agents*
2004-11 Michel Klein (VU), *Change Management for Distributed Ontologies*
2004-10 Suzanne Kabel (UVA), *Knowledge-rich indexing of learning-objects*
2004-09 Martin Caminada (VU), *For the Sake of the Argument; explorations into argument-based reasoning*
2004-08 Joop Verbeek (UM), *Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politile gegevensuitwisseling en digitale expertise*

2004-07 Elise Boltjes (UM), *Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes*
2004-06 Bart-Jan Hommes (TUD), *The Evaluation of Business Process Modeling Techniques*
2004-05 Viara Popova (EUR), *Knowledge discovery and monotonicity*
2004-04 Chris van Aart (UVA), *Organizational Principles for Multi-Agent Architectures*
2004-03 Perry Groot (VU), *A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving*
2004-02 Lai Xu (UvT), *Monitoring Multi-party Contracts for E-business*
2004-01 Virginia Dignum (UU), *A Model for Organizational Interaction: Based on Agents, Founded in Logic*
2003-18 Levente Kocsis (UM), *Learning Search Decisions*
2003-17 David Jansen (UT), *Extensions of Statecharts with Probability, Time, and Stochastic Timing*
2003-16 Menzo Windhouwer (CWI), *Feature Grammar Systems – Incremental Maintenance of Indexes to Digital Media Warehouses*
2003-15 Mathijs de Weerdt (TUD), *Plan Merging in Multi-Agent Systems*
2003-14 Stijn Hoppenbrouwers (KUN), *Freezing Language: Conceptualisation Processes across ICT-Supported Organisations*
2003-13 Jeroen Donkers (UM), *Nosce Hostem – Searching with Opponent Models*
2003-12 Roeland Ordelman (UT), *Dutch speech recognition in multimedia information retrieval*
2003-11 Simon Keizer (UT), *Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks*
2003-10 Andreas Lincke (UvT), *Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture*
2003-09 Rens Kortmann (UM), *The resolution of visually guided behaviour*
2003-08 Yongping Ran (UM), *Repair Based Scheduling*
2003-07 Machiel Jansen (UvA), *Formal Explorations of Knowledge Intensive Tasks*
2003-06 Boris van Schooten (UT), *Development and specification of virtual environments*
2003-05 Jos Lehmann (UVA), *Causation in Artificial Intelligence and Law – A modelling approach*
2003-04 Milan Petković (UT), *Content-Based Video Retrieval Supported by Database Technology*
2003-03 Martijn Schuemie (TUD), *Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy*
2003-02 Jan Broersen (VU), *Modal Action Logics for Reasoning About Reactive Systems*
2003-01 Heiner Stuckenschmidt (VU), *Ontology-Based Information Sharing in Weakly Structured Environments*
2002-17 Stefan Manegold (UVA), *Understanding, Modeling, and Improving Main-Memory Database Performance*
2002-16 Pieter van Langen (VU), *The Anatomy of Design: Foundations, Models and Applications*
2002-15 Rik Eshuis (UT), *Semantics and Verification of UML Activity Diagrams for Workflow Modelling*
2002-14 Wieke de Vries (UU), *Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems*
2002-13 Hongjing Wu (TUE), *A Reference Architecture for Adaptive Hypermedia Applications*
2002-12 Albrecht Schmidt (Uva), *Processing XML in Database Systems*
2002-11 Wouter C.A. Wijngaards (VU), *Agent Based Modelling of Dynamics: Biological and Organisational Applications*
2002-10 Brian Sheppard (UM), *Towards Perfect Play of Scrabble*
2002-09 Willem-Jan van den Heuvel (KUB), *Integrating Modern Business Applications with Objectified Legacy Systems*
2002-08 Jaap Gordijn (VU), *Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas*
2002-07 Peter Boncz (CWI), *Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications*
2002-06 Laurens Mommers (UL), *Applied legal epistemology; Building a knowledge-based ontology of the legal domain*
2002-05 Radu Serban (VU), *The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents*
2002-04 Juan Roberto Castelo Valdueza (UU), *The Discrete Acyclic Digraph Markov Model in Data Mining*
2002-03 Henk Ernst Blok (UT), *Database Optimization Aspects for Information Retrieval*
2002-02 Roelof van Zwol (UT), *Modelling and searching web-based document collections*
2002-01 Nico Lassing (VU), *Architecture-Level Modifiability Analysis*
2001-11 Tom M. van Engers (VUA), *Knowledge Management: The Role of Mental Models in Business Systems Design*
2001-10 Maarten Sierhuis (UvA), *Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design*
2001-09 Pieter Jan 't Hoen (RUL), *Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes*
2001-08 Pascal van Eck (VU), *A Compositional Semantic Structure for Multi-Agent Systems Dynamics.*
2001-07 Bastiaan Schonhage (VU), *Diva: Architectural Perspectives on Information Visualization*
2001-06 Martijn van Welie (VU), *Task-based User Interface Design*
2001-05 Jacco van Ossenbruggen (VU), *Processing Structured Hypermedia: A Matter of Style*
2001-04 Evgueni Smirnov (UM), *Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets*
2001-03 Maarten van Someren (UvA), *Learning as problem solving*
2001-02 Koen Hindriks (UU), *Agent Programming Languages: Programming with Mental Models*
2001-01 Silja Renooij (UU), *Qualitative Approaches to Quantifying Probabilistic Networks*
2000-11 Jonas Karlsson (CWI), *Scalable Distributed Data Structures for Database Management*
2000-10 Niels Nes (CWI), *Image Database Management System Design Considerations, Algorithms and Architecture*
2000-09 Florian Waas (CWI), *Principles of Probabilistic Query Optimization*
2000-08 Veerle Coupé (EUR), *Sensitivity Analyis of Decision-Theoretic Networks*
2000-07 Niels Peek (UU), *Decision-theoretic Planning of Clinical Patient Management*

2000-06 Rogier van Eijk (UU), *Programming Languages for Agent Communication*

2000-05 Ruud van der Pol (UM), *Knowledge-based Query Formulation in Information Retrieval.*

2000-04 Geert de Haan (VU), *ETAG, A Formal Model of Competence Knowledge for User Interface Design*

2000-03 Carolien M.T. Metselaar (UVA), *Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief.*

2000-02 Koen Holtman (TUE), *Prototyping of CMS Storage Management*

2000-01 Frank Niessink (VU), *Perspectives on Improving Software Maintenance*

1999-08 Jacques H.J. Lenting (UM), *Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.*

1999-07 David Spelt (UT), *Verification support for object database design*

1999-06 Niek J.E. Wijngaards (VU), *Re-design of compositional systems*

1999-05 Aldo de Moor (KUB), *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*

1999-04 Jacques Penders (UM), *The practical Art of Moving Physical Objects*

1999-03 Don Beal (UM), *The Nature of Minimax Search*

1999-02 Rob Potharst (EUR), *Classification using decision trees and neural nets*

1999-01 Mark Sloof (VU), *Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products*

1998-05 E.W. Oskamp (RUL), *Computerondersteuning bij Straftoemeting*

1998-04 Dennis Breuker (UM), *Memory versus Search in Games*

1998-03 Ans Steuten (TUD), *A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective*

1998-02 Floris Wiesman (UM), *Information Retrieval by Graphically Browsing Meta-Information*

1998-01 Johan van den Akker (CWI), *DEGAS – An Active, Temporal Database of Autonomous Objects*