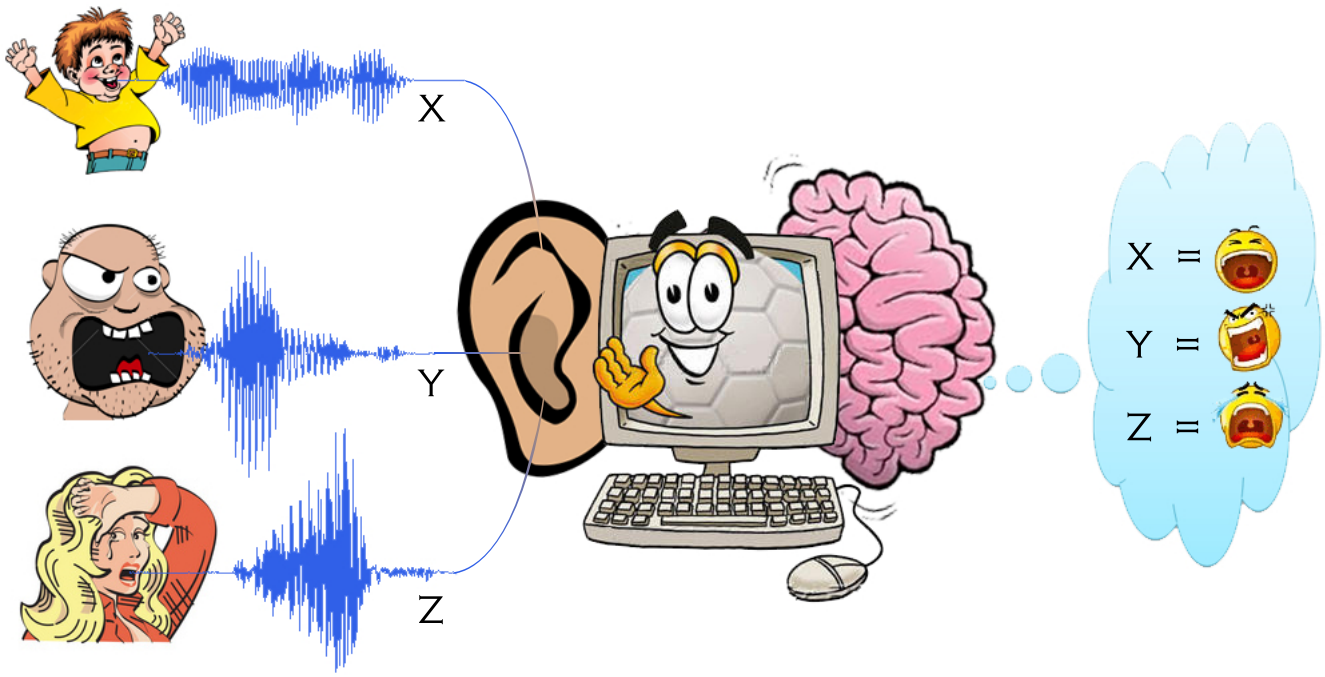
Delft University of Technology

MSc THESIS

Automatic Emotion Analysis Based on Speech

Iulia Chiriacescu



Faculty of Electrical Engineering, Mathematics and Computer Science
July 2009

Automatic Emotion Analysis Based on Speech

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

MEDIA AND KNOWLEDGE ENGINEERING

by

Iulia Chiriacescu
born in Brasov, Romania

Media and Knowledge Engineering
Department of Electrical Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Automatic Emotion Analysis Based on Speech

by Iulia Chiriacescu

Abstract

The focus of this thesis is on emotion recognition based on the speech signal. The state of the art is being reviewed. Based on models in psychology and the requirements of automatic systems, models for emotion recognition from speech are proposed and a most appropriate one for automatic detection is chosen.

The aim is to search for methods suitable for enhancing the generality, portability and robustness of emotion recognition systems. For this purpose we introduce a minimal set of features that provides recognition capabilities. The set has been tested on multiple databases and the results show that it provides capabilities for emotion discrimination.

An experiment including more databases was designed, aiming to get insight into the generalization capabilities of systems trained on extended corpora, as well as their portability. The results did however not give a clear indication in that sense.

Several classification techniques along with different feature types were fused together and better performing systems were generated. The best results were determined by the use of logistic regression fusion on t-normalized data.

Besides a list of experiments on acted corpora, two databases containing real speech have been used. The framework generated by real data was different than for acted data. The problem of working with emotions in a continuous space was also touched.

The knowledge gathered for all the experiments was used for building EmoReSp, a real-time system for emotion recognition based on speech.

Department : Man Machine Interaction
Student Number : 1300695

Committee Members :

Member:	Prof. Drs. Dr. L. J. M. Rothkrantz, TUDelft
Member:	Prof. Dr. Ir. D. A. van Leeuwen, TNO
Member:	Dr. Ir. P. Wiggers, TUDelft
Member:	Ir. H. Geers, TUDelft
Member:	MSc. M. C. Popa, TUDelft

Contents

List of Figures

List of Tables

Acknowledgements

The thesis your are reading is describing my graduation project as a Master's student at Delft University of Technology. I expressively thank TNO for giving me the opportunity to follow the Master's studies at TUDelft by granting me a TNO Excellence Launch Scholarship. Part of the research presented in this thesis was done at TUDelft and the other part in the framework of an internship at TNO Human Factors, Soesterberg. I consider myself lucky to have had this experience, since I could benefit from a lot of guidance and I learnt to look at the problems from different angles.

I would like to take this opportunity and thank those who helped me and influenced me these past two years. I am grateful to Leon Rothkrantz for all the support he offered me, for the time and energy he invested in this thesis and not only. He has a subtle way of helping me clear my thoughts every time I get lost in details, and also of encouraging me and giving me more confidence. Furthermore, I would like to thank Pascal Wiggers for all his support, for giving me valuable advice and for being able to find error in this thesis even after is was reviewed many times. I am very happy that David van Leeuven was my supervisor from behalf of TNO. I would like to thank him for showing me a different perspective of things, for showing a lot of patience and for introducing me to Linux.

Besides my supervisors, I would like to thank Dragos, Mirela, Bogdan, Alin, Khiet and Javier for always finding the time to help me and giving me lots of hints. Also, special thanks go to Ruud and Bart who were always there for me and provided professional technical support.

Of course, non of this would have been possible without "a little help from my friends". I thank them all for making the weight on my shoulders lighter and the times spent here unforgettable. I was happy to integrate in the student group from the MMI and CG group. Thanks for all the good times and of course for the Dutch lessons.

I thank my parents for always being there for me despite the distance and for always helping me to keep an optimistic mind. Finally, I thank Mihai for believing in me and for making my time brighter.

Iulia Chiriacescu
Delft, The Netherlands
July 23, 2009

1 Introduction

Whatever people do, whatever they try to achieve, in one way or another they communicate. As Paul Watzlawick said, "one cannot not communicate" [Watzlawick *et al.*, 1967]. More and more often, people communicate with machines. Whether the communication partner is another human or is a machine, the problem of transmitting the correct message and having the right interpretation from the conversation partner is of main importance.

Humans, by nature, use all their available senses for a maximum perception of the received message. They hear the sound, they read lips, they interpret gestures and facial expression and of course they discover the semantics of the utterance. Using all these senses, people can perceive a phrase as being amusing, they can understand irony and sarcasm, and they can show an appropriate reaction. Through all the mentioned senses, people actually sense the emotional state of the conversation partner and therefore are able to adapt to it. As human computer interaction (HCI) is evolving, many efforts concentrate on enabling computers to recognize human emotions and to react accordingly. Even though emotion detection comes as something natural for humans (at least to some extent), this is a very challenging task for computers.

The final purpose of emotion recognition systems is the application of emotion-related knowledge in such a way that human computer communication will be enhanced and furthermore the user's experience will become more satisfying. Surely, most people have experienced many times frustration caused by faulty computer programs or bad communication with the system. By enabling computers to sense the emotional state of the user and react accordingly, this communication can be transformed to a pleasant and productive one. However, improving the interaction with computers is not the only application of emotion recognition. Specialized systems can be used for even more serious problems like aggression detection, stress detection, or frustration detection.

A very common application of emotion recognition from speech is related to call-centers. Often call-center operators need to deal with unsatisfied clients, or the answering machine should transfer the caller to an operator whenever the stress level is perceived to be high. It is the same case with emergency call centers like 112 or 911, where calls could be sorted by priority and answered accordingly.

Keyboard and mouse are still dominating the human-machine communication. Enhancing this communication in such a way that it will become closer and closer to human-human communication can also have an important impact on the life of people with disabilities. Just imagine an environment designed to adapt to the specific need of people just by sensing their emotional state!

Many steps towards emotion recognition have already been taken by researchers. The approaches vary a lot when it comes to the source modality. Emotion recognition has been carried out from audio, video, linguistic information, electroencephalogram (EEG)

signal, etc. Later on, the trend was to fuse information of more of these channels of communication, and just like humans do, gather most information from what is being perceived given all input modalities.

1.1 Problem Definition

The work of this thesis focuses on recognizing emotions from one of the previously mentioned modalities: speech. Speech is probably the most common means of communication, especially when it comes to conveying a semantic message. Furthermore, one's voice can easily betray emotions. A stronger vibration, laughter, a sigh, a raise in voice, can give insights in the emotional state of the communication partner. However, the paralinguistic information is not the only kind of emotional clue that can be extracted from speech; the choice of words and especially the semantic content are also strong indications. Therefore, it is only natural to consider speech as a rich modality for emotion recognition that can lead to two different paths for analysis: acoustic analysis and natural language processing (NLP).

It is interesting to note how many times humans get most of their understanding of emotions from the spoken content and of course they combine it with the paralinguistic cues (see Figure 1.1). Many times, people experience difficulties when they are asked to name emotions from utterances in languages that they cannot understand, just because the semantic content is not available and suddenly the problem is not that easy, since sometimes the sound can be misleading. Whether emotions are dependent on culture is a topic for which a high amount of research has been dedicated from people interested in facial expressions [Ekman, 1994]. Such kind of discussions can also be also very interesting, but we will leave that for psychologists. We will just try to see the impact in getting some information from the spoken words can have on the overall recognition.

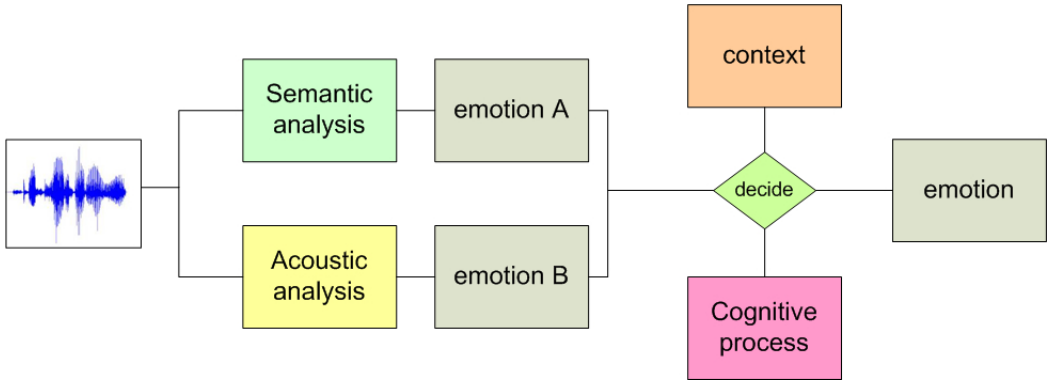


Figure 1.1: Model for emotion recognition from speech for humans

Previous research in speech emotion recognition proves that there is still not an available "recipe" that is expected to work in most problems. Research in natural language processing is based on spotting emotional keywords, parsing the text and doing semantic analysis. This is however beyond the scope of this thesis. For emotion recognition based on the acoustics signal, researchers follow in broad lines the following approach:

- Consider an emotional model (e.g., discrete or continuous),
- Start with analysing one or more of the available databases (however the efforts concerning more databases are still small in amount),
- Extract a set of features (which can be acoustic, linguistic or both),
- Train a classifier or use some reasoning techniques or probabilistic approaches in order to be able to make statements on test data.

Each of these steps is actually a point where a decision needs to be made. The first two problems can be regarded as a whole, since there is no much availability in databases, so the emotional model in most cases will be the one used for the recording of the used database.

There have been a lot of debates regarding databases. The main problem is that most available databases contain recordings of actors that have to portray a given emotion. As the final target is of course to detect emotions from the real world, it is often the case that these databases will be insufficient. A step further was the construction of elicited databases, which have the advantage that the subjects will be induced in a specific state before they will start acting and in this manner the data will be closer to reality. However, there exist a number of databases of real speech, but privacy issues keep most of them from being available and the fact that they contain real data also makes them more difficult to be used.

When it comes to feature extraction, the methods differ with regard to feature type (e.g., prosodic, spectral, and linguistic) and to the unit of analysis. Some researchers use features extracted frame wise, some use entire utterances and some take some in-between, for instance segments between pauses or words. It was proven that many times fusion of features from different levels can lead to improved recognition.

In early work, researchers had the tendency to use either frame level spectral features, or prosodic features from which some statistics are extracted over a longer period of time, like for instance mean fundamental frequency. However, the number of different features was not too high. In more recent works, it seems that the trend is to extract a large number of features, and especially a large number of statistics, and use an algorithm that selects the most relevant ones, in such a way that the remaining set will be very effective and with no redundancy. The problem with this approach, despite the improved performance on the initial database, is the lack of generalization capabilities. It was already shown that the selected best features are not the same for individual databases, so in the end we can expect the performance on real data to be worse.

Choosing a classification technique based on literature is also not trivial. So far researchers have used different classification methods, ranging from linear discriminant analysis (LDA) or k-nearest neighbours (kNN) to support vector machines (SVM) or probabilistic models like hidden Markov models (HMM) or Bayesian networks (BN). Studies have also approached combinations of multiple classification techniques, and many times the results were better or close to the ones of the best individual classifier.

Since everybody is doing emotion recognition in their own way, it should be possible to compare all approaches and to find advantages and disadvantages of each, and later

on come with a version that takes the best decisions at each point. However, coming to a conclusion from the previous work is not straightforward. This is mainly because researchers use different databases, many of them unavailable for all researchers, and after using different features and classification techniques they report the results in different manners, so it becomes impossible to know for sure which method was better.

The metrics to rate an emotion recognition system are usually dominated by the accuracy of detecting the right emotion. However, these metrics can vary with the application type. Some applications try to detect six basic emotions, some aim at a broader ranges, reaching to even 21 emotion classes in the Man-Machine-Interaction group from Delft University of Technology, and some focus on distinguishing only one emotion which is of main importance for their application, like for example call-centers or stress detectors.

Figure 1.2: General model for speech emotion recognition

To summarize, our purpose is to create a system that is able to perform emotion recognition from speech, as shown in Figure 1.2, and to tackle at least part of the problems that arise at each of the previously described steps. There are many methods, many approaches, many experimental settings, so taking the right decision becomes a very challenging task. We are in search for a good model for emotion recognition, that can benefit from the information from previous research and therefore is robust and general.

1.2 Research Goals

The purpose of this thesis is to experiment with both modalities enabled by speech: the way of speaking and what is spoken. The main focus however is on acoustic analysis while the combination with linguistic content can be regarded as a final touch. As we take each step needed for the development of the system, there are more questions that need to be answered in order to make a decision. In this section a number of goals that we plan to research on will be formulated.

1.2.1 Designing a Model for Emotion Recognition Based on Speech

In order to find a model for emotion recognition from speech, it is important to ask ourselves a few questions. First of all, it is important to know which features are relevant and give a good indication of the emotional state of a speaker. For example, when we are angry we speak louder, so the amplitude of the speech signal will be higher. When we are frightened and we scream, the pitch is very high. Also, when we are under a lot of stress, our voice is trembling because there is a tremor is our muscles which influence the vibrations in the vocal folds, and this result is called jitter.

If we want to have a good model for building an emotion recognizer, it is important to choose a representative set of features. The next step is finding a way to come up with a decision for an emotional output based on the behavior of the chosen features. In other words, we need to find a relation between emotions and features. For the purpose of this thesis we will use classifiers in order to model the relation between emotions and features.

1.2.2 Combining Individual Databases into a Corpus with Generalization Capabilities

Choosing a database to work with is not a fairly difficult decision, since there is no large offer of speech emotional databases. This of course results in a limitation of the system, because it will be tailored for that database. For the purpose of this work, the choice was to gather all available databases, and experiment with all of them. We aim at finding a way for combining the databases in such a way that the emotion recognition system becomes more robust and shows improved accuracies. The ability of emotion recognition systems to generalize has been only little investigated, and it can be a very accessible source for improvement.

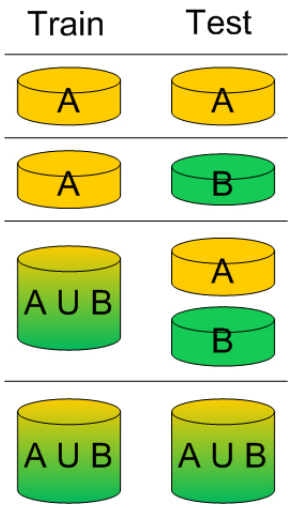


Figure 1.3: Different methods for combining databases

Figure 1.3 gives an insight into different possibilities for training and testing involving more databases. The first method is the one currently used in emotion recognition research. The second method can be regarded as the main challenge in real applications, since the training data will not be the same as the test data, and this usually leads to problems. The difference between the last two approaches is in the way of looking at testing results.

Our aim is to find a good way of combining more databases, that is the most beneficial for the emotion recognition system, from the accuracy, robustness and portability points of view.

1.2.3 Increasing the Recognition Capabilities by Fusion of More Classifiers

There are more contributions in this field that address the problems of classification and it appears to be the case that a fusion of more classifiers can lead to better results. The idea behind using more classifiers is that they are expected to make different mistakes and have different internal models for the data. That is why, a combination of more classifiers is expected to be more reliable. The choice of classifiers in the fusion should be made in such a way that they complement each other. We plan to experiment with using more classifiers and to see which type of fusion will give better results.

Fields of study similar to emotion recognition are speaker recognition and language recognition. The main part of these research areas is also related to feature extraction form speech and classification. Some recent efforts concentrate on different modalities of fusing classifiers, some of them with very promising results. We are also interested is using these approaches for improving the system.

1.2.4 Key-Parameters for Research

There are a number of details which can have a great influence on the performance of an emotion recognizer. One of them is the unit of analysis. By unit of analysis we refer to the length of the speech segment that is used for making a decision on the emotional state of the speaker. This could be phoneme, frame, word, sentence or utterance.

Choosing the unit of analysis is closely related to the feature types used, because some features are extracted for instance frame-wise, and some are computer over longer time intervals. Another interesting topic is seeing how the combination of different granularities (frame level and utterance level) of features influence the accuracy.

1.2.5 Designing a Real-Time Automated Emotion Recognizer Based on Speech

Most previous work focused on offline classification of databases. As the final purpose of the emotion related research from the human computer interaction community is aiming in applying emotion recognition to enhance existing systems and improve communication, it is important to be able to apply all the modalities in real-time. As a final step in our research we plan to design and implement a real-time emotion recognizer.

Building a real-time system is not straight-forward. There are many problems that arise when an online analysis is intended. We plan to explore these problems and come up with suitable solutions at least to some extent.

The approach followed for preparing this thesis and the experiments on which it is based are summarized in Table 1.1.

1.3 Thesis Outline

This thesis is organized as follows. The research subject and the main problems from the field together with the main research goals as well as the societal relevance of the subject were presented in the introduction.

Table 1.1: Action and goals of our research

Actions	**Goals**
literature research	learn about the state of the art
use more databases	increase portability
use different feature granularities	increase robustness
use different feature types	increase robustness
fuse more classifiers	increase recognition capabilities
emotional keywords spotting	add another channel of information
build a real-time emotion recognizer	experiment the challenges of a real-time framework and put in practice the experience accumulated so far

The second chapter gives an overview of the related work. In the first part the most important emotion theories are presented, as well as model of emotion which are of importance for the emotion-related research. As in practice building an emotion recognition system that has speech as an input is a problem that very much depends on learning from examples, databases of emotional speech and the recent controversies from that domain are introduced. Capturing information from the speech signal is of course one of the major steps in the emotion recognition process. In the third section of the second chapter, several ways of describing the speech signal are presented. This is done in terms of speech features: prosodic, spectral, voice quality and linguistic. A section describing some of the frequently used machine learning methods used so far in emotion recognition from speech follows. We would like to stress that this chapter only presents the machine learning techniques as a link to the emotion recognition field and does not provide full explanation of the concepts. For the machine learning techniques that we have used, please see chapter three. The second chapter ends with a conclusion on the main problems present in the emotion recognition area.

After the main topics in the emotion recognition research based on speech are presented, it is time to give more insight into what our research focuses on. The third chapter presents the resources we have used. This includes a description of the databases, features, feature extraction tools, classification methods and classification tools which we have used.

In the fourth chapter we introduce a couple of models from psychology of how humans perceive emotion. Based on them, we propose other three models that can be used in an automated emotion recognizer. Each of the model is analysed according to its advantages and disadvantages, and finally a decision is made for one model that is used within the experiments of this thesis.

Moving on, the fifth chapter is the one in which the experiments are described. A

first section gives an short overview over the entire flow of the experiments, highlighting the main directions the research focuses on. Later on, the methodology used for experimenting is explained, including the evaluation measures. Then, the setup of each experiment is described and the results are presented and interpreted. The experiments are of high importance to our research because they are meant to clarify the questions we ask ourselves in the research goals section. They span several directions, but the final goal of all of them is to acquire more insight into what methods can lead to a robust and general emotion recognizer based on speech.

The experiments lead to more knowledge about the expected behaviour of an emotion recognition system and more insight into the problems that one might expect. As a final challenge, a real-time emotion recognizer based on speech signal was developed and implemented. In the sixth chapter the main challenges of building such a real-time system are explained, along we the decisions we made. Design and implementation details, as well as conclusions are discussed.

The thesis ends with our conclusions. It is a chapter about lessons learnt and confronting problems as well as unanswered questions and solution still to be found. Direction for further research are given within this last chapter.

2 Related Work

This chapter can be regarded as a presentation of the state of the art in emotion recognition from speech. However, the notion of state of the art is perhaps not totally suitable, since a comparison of the different approaches is difficult and there is not one best recipe. We can consider this chapter as a journey across the main ideas behind the process of recognizing emotion. At each stop the concepts are presented, along with the relevant contributions researchers made so far.

The first section discusses the main emotion theories and emotion models. The second section presents the work developed with regards to speech emotional databases, as well as the main problems in the field. The overview continues with presenting features used in classification and presenting the variation in these features when emotions appear. Classification approaches are also described and the results of their application are discussed. The chapter ends with a short conclusion regarding the problems from the studied literature.

2.1 Emotions

2.1.1 Emotion Theories

Theories about emotions date back to the Hellenistic philosophers and continue up to present, gaining complexity and broadness. Notions about the history of emotions appear in a myriad of articles. Here we follow mostly the description in [Sousa, 2008].

Theories about emotion can have the tendency to cluster into several directions. One approach is to treat emotions as one of the faculties of mind. For example Plato suggests that the soul has a tripartite structure consisting of cognition, emotion and motivation, and that these components are clearly separated. Later on, Aristotle disagreed with this separation, considering that there is actually interaction between the three components, and regarding emotion as a capacity to learn to experience the right feelings given certain stimuli. Later on, Hume emphasised the strong influence of passion upon reasoning. Other theories consider three or two faculties of mind while the other one is contained in one of these two. The Stoics thought of emotions as judgements about the values of things, and that we should remain indifferent to them.

In the seventeenth century, Decartes was the author of a revolution in emotion theory. He proposed to study both mental and physiological phenomena at the same time. Decartes believed that emotions had an impact on rationality, that they were affecting thoughts and decisions. Also, he considered that there exists a set of basic emotions from which the others are compounded. Later on, Charles Darwin classified several bodily traits characteristic for several emotions.

The evolutionary approach considers that emotions are caused by adaptation efforts

for solving ecological problems that appeared in the evolution of species. Charles Darwin emphasizes the survival relatedness of emotions, and therefore expects that emotions are more or less the same for all humans, independent of their culture. Also Darwin was the first to describe facial expressions and bodily movements that correspond to certain emotions. His theories have been further developed by researchers like Ekman, Izard, Tomkins. Also the studies of Ekman and Friesen show that there is a set of basic emotions which have universally recognizable facial expressions. These emotions are: happiness, sadness, fear, anger, surprise, and disgust. However, surprise and disgust are considered too simple to be called emotions by [Panksepp, 1998]. In particular the work of Ekman who postulated the universality of the six basic emotions has been extensively used in research concerning emotions. However, this approach neglects emotions that involve higher cognitive processes, like boredom, jealousy or envy.

Other theories postulate that emotions are a specific kind of feeling. The most famous of these theories was formulated by William James and Carl Lange, and regards emotions as feelings formed as a reaction to certain physiological stimuli. In the opinion of James, we cry and then we feel sad because we cry, we scream and then we feel angry, we laugh or smile and then we fell joy [James, 1884]. Problems with this concept were first signalized by Walter Cannon who objected that it is not possible to differentiate emotions on these bases, for example having the same instinctive reaction for fear and anger. The theory of Cannon resembles the results of an experiment by Stanley Schachter and Jerome Singer [Schacter & Singer, 1962], that proved that the differences between specific emotions are not physiological, but cognitive or something else.

The effects of emotion on behaviour and on decision making are also highlighted by Bechara and Damasio in the somatic-marker hypothesis:"An emotional mechanism that rapidly signals the prospective consequences of an action, and accordingly assists in the selection of an advantageous response option" [Bechara & Damasio, 2005]. Frijda also regarded emotions as tendencies to adopt certain decisions and behaviors according to one's needs [Frijda, 1986].

The psychological approach highlights the fact that emotions appear as an effect of an evaluation of certain stimuli in accordance to personal interest. This process of determining the personal meaning of a situation to an individual was first introduced in psychology by Magna Arnold under the name of appraisal. Another sustainer of this theory is Richard Lazarus who believes that appraisals are the necessary and sufficient condition for emotion, and mainly classifies appraisals into attraction and aversion [Lazarus, 1999]. A very complex model was elaborated based on appraisal theory by Scherer et al. 2001, which locates emotions on 18 or more dimensions of appraisal. This multidimensional space was later on simplified to two or three dimensions, mainly focusing on valence and arousal.

Separately from the theories mentioned above, there is another view on emotion, which relates to social and cultural aspects. This constructivist approach highlights the role of the cortical processes that arise during social behavior and mainly advocates that emotions are strongly related to social and cultural life of the individual, and that emotions can be best understood and recognized between members of the same culture.

2.1.2 Models of Emotion

The previous section gave an insight into philosophies of emotion. This section will continue with a more practical view: the description of several emotional models that provide a basis for human computer interaction related research. The most used models of emotion are the discrete model and the dimensional model. We will describe them in turn, and then we will also mention some other models.

2.1.2.1 The Discrete Model

Considering the existence of only a finite set of emotions is one of the most popular concepts used in emotion recognition. The number of these emotions varies between six and twenty-one, and the emotions are easily conceptualized because of their labels representing commonly occurring emotions. This approach, also known as the basic emotions approach, is based on the evolutionary theory of Darwin, where he claims that due to species' adaptation for survival, a finite number of fundamental emotions were developed, and that facial expression generated by experiencing these emotions are universal.

Even though Darwin is generally considered the first to claim the universality of several facial expressions corresponding to fundamental emotions, this idea has been postulated before by Bain, Bell, Duchenne de Boulogne, Piderit and Spencer [Russell, 1980]. The work of Darwin was further continued by many researchers who searched for empirical evidence of the basic emotions, their relation to facial expressions and the universality of these emotions. Tomkins argued for the existence of nine basic emotions that are recognizable by specific facial expression [Tomkins, 1984]. Ekman and Friessen recognized only six basic emotions, namely anger, fear, disgust, happiness, sadness and surprise. They conducted extensive studies on proving the universality of these emotions and the corresponding facial expressions. Izard focused on the discrete emotions theory, which states that emotions are innate and distinct from one another from a very early age, and each emotion is described by specific facial and bodily expressions [Izard, 1977].

Another discrete model, the circuit model, was proposed by Panksepp. The main idea is also evolutionary-related, and mentions the existence of four circuits that enable mammals and primates to cope with different threats for survival. These circuits correspond to rage, fear, expectancy and panic, but Panksepp also mentions the existence of secondary emotional states that are generated by mixing several primary states [Panksepp, 1982].

As the impressive number of human emotions cannot be differentiated and expressed with only a discrete set of basic emotions, research focused on blending the basic emotions and obtaining other secondary emotions. Plutchik considered eight basic emotions. Opposite emotions in the graphical representation called the solid of emotions (see Figure 2.1), neutralize each other. Blending the adjacent basic emotions results in primary dyads, like for example mixing surprise and sadness leads to disappointment. Secondary and tertiary dyads are also used; however emotion names fitting the blended emotions were not always found [Plutchik, 1980].

The vertical axis from Plutchik's solid represents intensity, for example Rage, Anger and Annoyance are three similar emotions but with different intensities. Besides mixtures

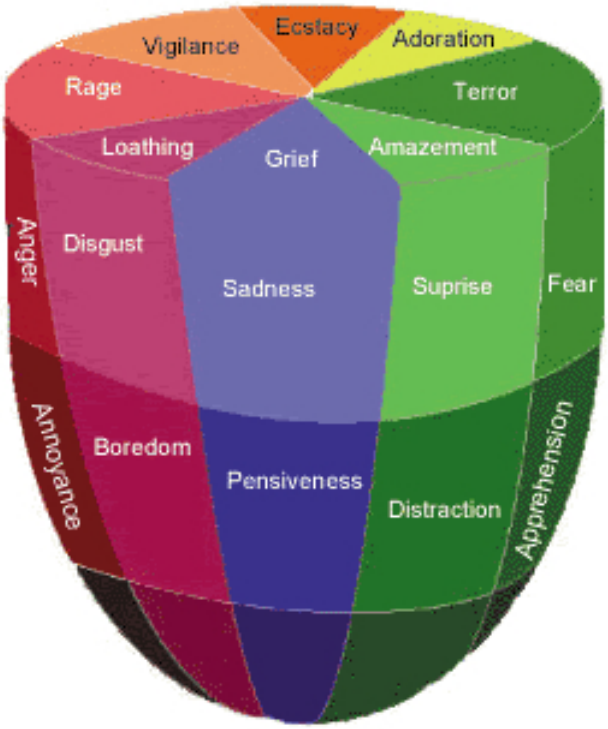


Figure 2.1: Plutchik's solid of emotion

of two emotions, blending three emotions is also possible, resulting in emotional triads.

The basic emotions were extensively criticized by Russell, who denied their universality and the grounds of the studies of Ekman and others. Russell argued for the use of the valence-arousal model which will be, among other dimensional models, described in the next section.

2.1.2.2 The Dimensional Model

This approach is based on the principle that emotions can be regarded as points in an n-dimensional space. There were several attempts to classify emotions on one or on a small number of dimensions. Emotions were even classified on one dimension, pleasantness or arousal, for example Duffy considered strictly excitation. Schlosberg proposed three such dimensions: pleasantness, activation, and attention-rejection [Posner *et al.*, 2005].

However, it seems that the most popular model is the two dimensional one, with the dimensions pleasantness (the degree of perceived pleasure) and activation (the degree of excitation), or valence and arousal. Through time, the effectiveness of these two dimensions was pointed out by Schlosberg, Russell, Plutchick and Cowie et al. [Steidl *et al.*, 2008]. Even though the names of the dimensions vary slightly, it can be noticed that in general one dimension describes whether the perceived emotion is positive or negative, which can also be regarded as appraisal, and the other dimension describes the intensity of emotion, which can also be regarded as the strength of the

action tendency [Russell & Barrett, 1999].

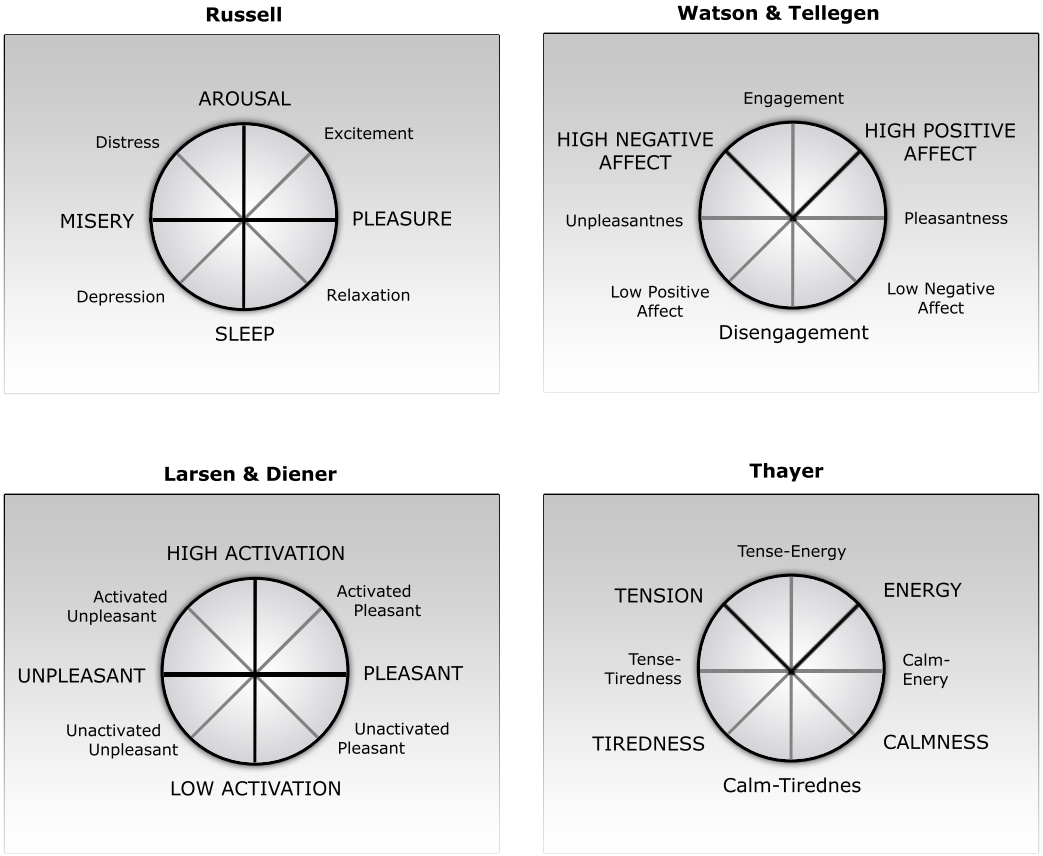


Figure 2.2: Descriptive models of affect [Russell & Barrett, 1999])

In Figure 2.2 four models of affect are presented, rotated and reoriented in such a way that the resemblance with the model in Figure 2.3 is emphasized. In Figure 2.3, the inner circle shows a schematic map of core affect, while on the outer circle, some basic emotions are mapped [Russell & Barrett, 1999].

The valence and arousal model is a very efficient way of describing emotions and therefore represents a very popular choice for the implementation of emotional related systems. However, it has a drawback due to the loss of information caused by reduction of the n-dimensional space to a two dimensional space. This drawback can be noticed when mapping emotion related words onto the two dimensional space and using Euclidian distance for distinguishing the similarities and differences between them. For example, anger and fear are mapped on two points that are close together and impossible to separate. A solution can be the addition of a new dimensions, like perceived control or inclination to engage, that results in a positive evaluation for anger and a negative evaluation for fear. However, this new dimension only solves a limited number of these confusions, and other dimensions might be necessary [Cowie & Cornelius, 2003].

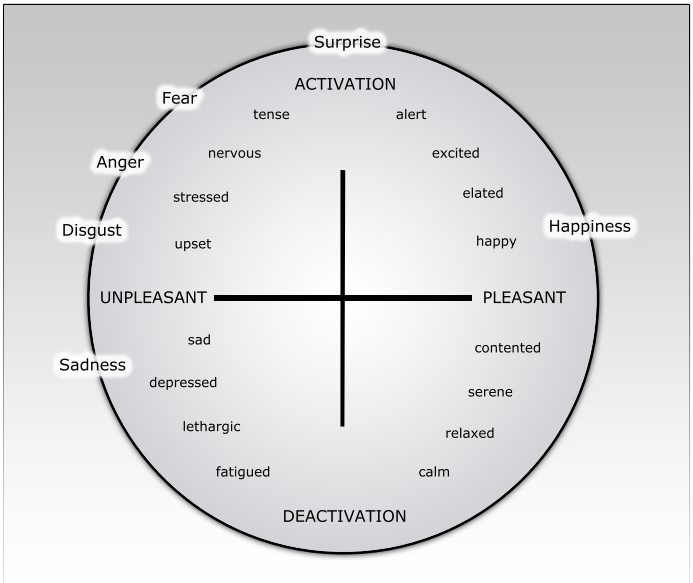


Figure 2.3: Schematic map of affect [Russell & Barrett, 1999]

Whissel and Plutchik also contributed to the popularity of the activation-evaluation approach. Whissel described an extensive set of emotional words in terms of their activation and evaluation. Both dimensions are ranging from 1 until 6, with [3,3] correspond to neutral. Plutchik argued that full blown emotions are not evenly distributed in the activation - evaluation space, and that they form a circular pattern. He provided angular measures for emotional words, depending on their activation and evaluation. These two solutions for measuring emotion are very practical and useful in emotion related research [Cowie *et al.*, 2001].

2.1.2.3 Emotion Perception

This section will briefly introduce a model of emotion perception proposed by Scherer [Scherer, 2003], which is an adaptation of the Brunwik's functional lens model of perception. The model, depicted in Figure 2.4, encodes emotional expressions of the speaker expressed by certain speech and voice characteristics. These characteristics are divided into two categories: distal and proximate cues, with regard to the distance between the vocal cues and the observer. Physiological changes affecting respiration, phonation and articulation are an effect of emotional arousal. These changes are present in the speech signal and are called distal cues (distant from the observer). As the signal is perceived by the auditory system of the observer, these cues are called proximal (close to the observer).

However, it is not always the case that emotional states produce reliable externalizations. The term ecological validity refers to the correlation between the internal state of the speaker and its externalization.

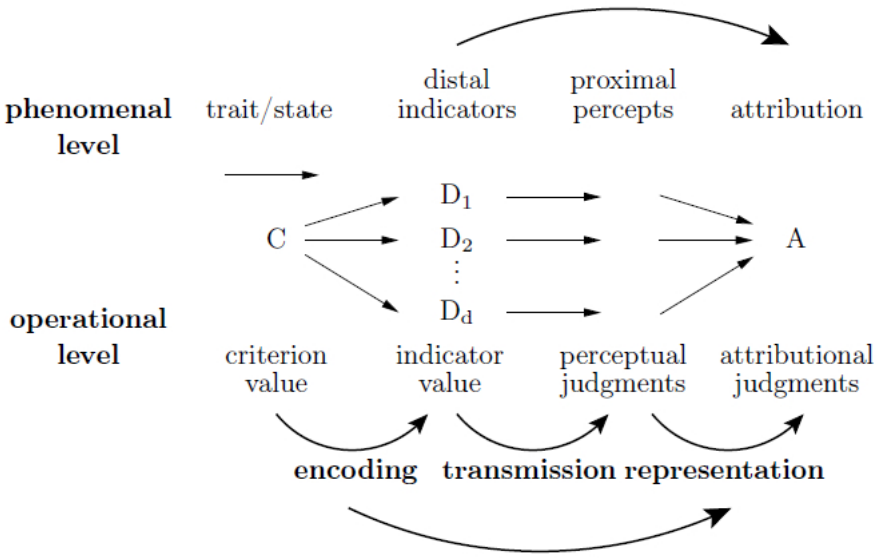


Figure 2.4: A Brunswikian lens model of the vocal communication of emotion [Scherer, 2003]

The perceptual model of emotion leads to different studies:

- Encoding studies: the search for acoustic patterns that are characteristic of certain emotions,
- Decoding studies: the analysis of the human capability of inferring emotions from speech,
- Inference studies: the examination of the underlying voice-emotion inference mechanism,
- Transmission studies: focusing on the transmission channel,
- Representation studies: the analysis of mental representation algorithms.

To conclude this section, we will mention that controversy does not seem to lack in modeling emotion. In speech emotion recognition however, the discrete approach is usually preferred, probably due to the ease of finding or developing databases with several fundamental emotions. The next section gives more details about emotional databases.

2.2 Emotional Speech Databases

Up to present, speech emotion recognition is based on using different classifiers on recorded data. Therefore, the availability of labeled speech emotional databases is a prerequisite for the design of such systems. The effectiveness as well as the perceived results of emotion recognizers are to a high extent dependent on the databases used for

training and evaluation. An overview of emotional speech databases up to 2006 can be found in [Ververidis & Kotropoulos, 2006] where a set of 64 databases are reviewed with regard to present emotions, data collection procedure, language, content, speech type (natural, simulated or elicited) and other physiological signals present in the recordings that can be used for emotion detection. Considering the fact that an overview was done over 64 databases, one might believe that the lack of speech data is in fact not true. But, despite the existence of these databases, they are not all publicly available.

There are three kinds of emotional databases with regard to the authenticity of emotion. Databases with acted speech include portrayals of emotions by professional or amateur actors. In general actors are asked to speak some given utterances while expressing a certain emotion and the recording is labeled as containing the specified desired emotion. Another kind of databases contain elicited emotions. This kind of emotion are neither real, nor simulated. They are induced to the participants by different modalities, like stories, movies, virtual reality or role playing. The last type are databases of spontaneous speech which contain real emotions.

Authentic emotions are of course the most desirable for emotional databases. However, making such recordings is very difficult since real emotions are very rare and their duration is very short. Also, real emotions can be deliberately hidden or changed by individuals, according to their perception of social rules and their will. Furthermore, certain emotions almost never occur, and due to sparse data classification cannot be performed. The spontaneous databases always contain a high amount of neutral utterances,and only a small set of emotional ones. Because of these difficulties of recording real data, most of the available databases contain acted speech, which means that emotions are not real but deliberately expressed.

Acted speech databases usually portray a small set of emotion classes like the basic emotions of Ekman or only subsets of these emotions. Due to the fact that in real data emotions occur rarely and also recording is not ethical for all types of emotions, sometimes coarse labels like positive or negative are used. There are also attempts to distinguish high level emotions like annoyance and frustration [Walker *et al.*, 2001][Ang *et al.*, 2002] or anxiety, irritation and resignation [Laukka *et al.*, 2008].

Several approaches were used for the collection of spontaneous emotion databases. On one hand the participants had human interlocutors and performed in interviews, phone conversations or were actively involved in meetings. On the other hand human machine interaction was used with the help of Wizard of Oz (WoZ) scenarios or computer based dialogue systems [Zeng *et al.*, 2007]. Recent scenarios for the recording of spontaneous databases are WoZ scenarios of children playing with Sony AIBO [Steidl *et al.*, 2008] [Batliner *et al.*, 2008], driving simulators [McNahon *et al.*, 2008], one-shooter computer games [Truong & Raaijmakers, 2008] or earthquake emulators [Ververidis *et al.*, 2008].

The performance of emotion recognizers is better on acted speech, especially when professional actors are used, since they perform in a manner that increases arousal and emotions are more obvious. Even though acted emotions are recognizable by humans, it appears that differences exist between real emotions and acted ones. Studies by [Audibert *et al.*, 2008] show that people are in most of the cases able to distinguish acted speech from natural speech.

The trend in speech emotion recognition is to turn from acted or elicited to spontaneous data. Several researchers encourage the use of real speech, for instance Douglas-Cowie et al in"Towards a new generation of databases" [Douglas-Cowie *et al.*, 2003]. However, there are also opinions that the real problems are not caused by acting itself, but by the protocols that are followed. In this direction, Busso and Narayanan [Busso & Narayanan, 2008] suggest that spontaneous data comes with a lot of problems with regards to preprocessing and actually capturing emotions, and propose a new protocol for acted speech recording which allows actors to improvise and get better emotionally involved in dialogues. Also, what differs from general acted database recording is the length of utterances, which they advise should be much longer than in present databases.

Besides the databases mentioned in [Ververidis & Kotropoulos, 2006] and [Zeng *et al.*, 2007] where an overview of audio-visual emotion recognition is given, we would like to mention some other recordings that captured out attention. An audio-visual database containing the 21 product emotions from [Desmet, 2002] was developed by Mathijs van Vulpen [van Vulpen, 2008], [Chitu *et al.*, 2008]. Participants portrayed the extensive range of emotions helped by an elicitation protocol. The spoken language is Dutch and the number of participant is so far 2, but a higher number is expected in the near future.

An extensive audio-visual emotional database was collected within the HUMAINE project [Douglas-Cowie *et al.*, 2007]. The HUMAINE database is actually comprised of more databases some with induced and some with natural emotions. The emotional content spans a high range of emotions, and has been labeled both globally and time aligned according more dimensions, like valence, intensity, genuineness and activation. Part of the database is publicly available for research.

Another audio-visual data collection that has been made available in 2008 is called *Vera am Mittag* [Grimm *et al.*, 2008] and contains 12 hours of recordings from the TV show with the same name. Spontaneous emotional colored speech (German) and behavior from guests of the show has been annotated in a continuous tri-dimensional space: valence, activation and dominance with 5 point scale precision by 17 human labelers. The FAU Aibo Emotion Corpus [Steidl *et al.*, 2008] [Batliner *et al.*, 2008] is another collection of German spontaneous speech. The participants are 51 children from two schools in Germany that interact with Sony dog-like robot Aibo. In a Wizard of Oz scenario, the children try to give spoken commands to Aibo which instead is remote controlled and does not always follow the orders, while the children experience different emotions. The database is preprocessed and labeled on more degrees of granularity and will be made publicly available.

Despite the high number of existing data collections, the problem of examining real emotions for accurate emotion classification is far from being solved. While some emotions are very easy to elicit, (e.g. laughter), this becomes a very challenging job for others (e.g. embarrassment). Psychology related studies provide knowledge for recording emotions that are difficult to elicit in laboratory environment. However, it is a matter of ethics to record and even elicit certain kinds of emotions [Zeng *et al.*, 2007]. An important factor concerning databases is their labeling. With acted speech, actors are asked to make an utterance expressing a certain emotion, and then the utterance is labeled

with the desired emotion. Whether the actor succeeded or not to express accurately the emotion is a factor of risk. The problem of labeling increases in difficulty when it comes to real data, since labeling becomes a subjective action and high amounts of data are discarded due to inter-labeler disagreement.

The dimensional approach can also be used for labeling. Two labeling tools, FeelTrace and Geneva Emotion Wheel, are reviewed in [Steidl *et al.*, 2008]. The labels are assigned in the activation- evaluation space and the control-valence space respectively.

Issues like privacy or copyright prevent several databases of being publicly available. Also, it is difficult to compare results of different researchers since the used databases have different technical characteristics, different kinds of emotions, different numbers of actors, or different labeling protocols. Directions for constructing databases that are appropriate for this task are formulated in [Chitu *et al.*, 2008].

After having described part of the available speech emotional databases and discussing some of the problems in this area, we will move on to describing relevant features for speech emotion recognition.

2.3 Speech Features for Emotion Extraction

Even though sound is a single channel, there are two types of features that can be extracted and analysed: acoustic features, also called paralinguistic, and linguistic features. A large set of acoustic features has been found so far to correlate with emotions. In general researchers use different subsets, and try to find the most suitable combinations. The acoustic features can be categorized in prosodic, spectral, and voice quality features. Our description will follow [Steidl *et al.*, 2008]. After giving an overview of acoustic features, linguistic features will also be discussed.

2.3.1 Prosodic Features

Prosody studies the rhythm, stress and intonations of speech and works on larger segments of speech, like syllables, words, phonemes or entire turns of a speaker. Pitch, loudness, speaking rate, durations, pause and rhythm are all perceived characteristics of prosody. Even though there are no unique acoustic correspondents for these characteristics, there are strong correlations between them.

The pitch signal is caused by the vibrations of the vocal folds. It carries information about emotion due to its relation with the tension of the vocal folds and the subglottal air pressure. The pitch frequency and the glottal air velocity at the opening time of vocal folds are the most used features related to pitch. The time between two openings of the vocal folds is called pitch period, and the vibration rate of the vocal folds is called fundamental frequency. Algorithms for estimating the pitch signal can be found in [Hess, 1992] and [Ververidis & Kotropoulos, 2006].

If features for an entire speech segment are to be analyzed, statistical functions like mean, median, minimum, maximum, standard deviation, or more seldom third or fourth standardized moment are applied to the F0 base contour. Sometimes, these functions are applied to the first or second derivative of the F0 contours, resulting in delta and delta delta features respectively.

Statistics of the pitch contour can also be misleading. Given the examination of an interrogative sentence, the pitch contour will usually be wider than for an affirmative one. However, this difference is only related to sentence nature, and not to emotion [Nogueiras *et al.*, 2001].

Sound energy is perceived as loudness and is related to emotional intensity. In general, the same statistical functions mentioned beforehand for pitch are used to describe the change in energy signal over time.

Duration features correlate with the speaking style, e.g. speaking rate, duration, pauses. The most used duration feature is the speaking rate, which equals to the ratio of the observed and the expected duration of a speech segment. The ratio of speech time to pause time, the duration of the longest pause, the number of pauses in a speech segment, and even the positions of F0 extrema are other used duration features. Prosodic features are the most popular in speech emotion recognition. Some examples of papers using prosodic features are: [Steidl *et al.*, 2008], [Ang *et al.*, 2002], [Batliner *et al.*, 2006], [Graciarena *et al.*, 2006], [Austermann *et al.*, 2005], [Devillers & Vidrascu, 2006], [Forbes-Riley & Litman, 2004], [Lee & Narayanan, 2005], [Schüller *et al.*, 2005b], [Truong & van Leeuwen, 2007] and [Krajewski & Kroger, 2008].

Table 2.1: Definition and acoustic measurement of voice cues in vocal affect expression [Juslin & Scherer, 2005]

Acoustic cues [1]	**Perceived correlate**	**Definition and measurement**
Pitch		
Fundamental frequency (F0) (59)	Pitch	F0 represents the rate at which the vocal folds open and close across the glottis. Acoustically, F0 is defined as the lowest periodic cycle component of the acoustic wave form, and is extracted by computerized tracking algorithms. Various measures: mean (M), standard deviation (SD), range (R), max, min, median, mode, and floor (i.e. the lower 5% of F0 values).
F0 contour (19)	Pitch contour	Sequence of F0 values across an utterance. Besides changes in pitch, the F0 contour also contains temporal information. The F0 contour is hard to operationalize and most studies report only qualitative classifications, the proportion of rising to falling F0 contours, or the range and gradient of F0 fall at the end of the sentence.
Jitter (13)	Pitch perturbations	Small-scale perturbations in F0 related to rapid and random fluctuations of the time of the opening and closing of the vocal folds from one vocal cycle to the next. Extracted by various computerized tracking algorithms.
Intensity		
Intensity (39)	Loudness of speech	Intensity is a measure of energy in the acoustic signal and it reflects the effort required to produce speech. It is usually measured from the amplitude acoustic wave form. The standard unit used to quantify intensity is a logarithmic transform of the intensity called the decibel (dB). Various measures: mean (M), standard deviation (SD), range (R), Max, Min, Median, Mode.
Attack (2)	Rapidity of voice onsets	The attack refers to the rise-time or rate of rise of amplitude for voiced speech segments. Usually measured from the amplitude of the acoustic wave form.
Shimmer (-)	Loudness perturbations	Refers to small regular or irregular variations of amplitude maxima in successive glottal cycles. Extracted by computerized tracking algorithms.
Temporal aspects		
Speech rate (41)	Velocity of speech	The rate can be measured as overall duration or as units per duration (e.g. words per minute). It may include either complete utterances or only the voiced segments of speech. Various measures: syllables per second, relative duration of voiced versus unvoiced segments, syllable duration, duration of accented vowels, total duration of utterance with or without pauses.
Pauses (14)	Amount of silence in speech	Refers to silent periods in an utterance and is usually measured in terms of absence of energy in the acoustic wave form . Various measures: relative number (Pn) and duration (Pd) of pauses (longer than 200-300ms) within or between selected units of analysis.
Rhythm (5)	Speech rhythm	There is yet no standardized measure of speech rhythm, but it has been suggested that the degree of regularity versus irregularity of speech may distinguish among positive and negative emotions.
Voice quality		

Continued . . .

Table 2.1: Definition and acoustic measurement of voice cues in vocal affect expression (continued)

Acoustic cues	**Perceived correlate**	**Definition and measurement**
High-frequency energy (24)	Voice quality	Refers to the relative proportion of total acoustic energy above, versus below, a certain cut-off frequency. As the amount of high-frequency energy in the spectrum increases, the voice sounds more 'sharp' and less 'soft'. Obtained by measuring the long term average spectrum (LTAS), which is the distribution of energy over a range of frequencies, averaged over an extended time period. Various measures: HF 500, HF 1000, spectral slope (linear regression of energy distribution in the frequency band above 1000Hz).
Formant frequencies (10)	Voice quality	Refers to frequency regions in which the amplitude of acoustic energy in the speech signal is high, reflecting natural resonances in the vocal tract. The first two formants largely determine vowel quality, whereas the higher formants may be speaker-dependent [Laver, 1980]. The mean frequency and the width of the spectral band containing significant formant energy are extracted from the acoustic wave form by computerized tracking algorithms. Various measures: mean (M) and bandwidth (bw) for F1, F2, F3, and F4.
Precision of articulation (8)	Auditory effort	The vowel quality tends to move towards the formant structure of the neutral schwa vowel (e.g. as in 'sofa') under strong emotional arousal. The precision of articulation can be measured as the deviation of the formant frequencies from the neutral formant frequencies, as reported in various sources. F1 (precision) is most commonly measured.
Glottal wave form (8)	Voice quality	The glottal flow wave form represents the time air is flowing between the vocal folds (abduction and adduction), and the time the glottis is closed, for each vibrational cycle. The shape of the wave form helps to determine the loudness of the sound generated and also its timbre. A 'jagged' wave form represents sudden changes in airflow that produce more high frequencies than a 'soft' wave form. The glottal wave form can be inferred from the acoustical signal using inverse filtering.

[1]Note: The values in parentheses indicate the number of studies that provided data points for each basic parameter in 104 studies of vocal expression of emotions and can be used as a rough index of the relative frequency with which each parameter has been measured previously.

In Table 2.1, a large set of acoustic cues are analyzed with respect to perceived correlate, definition, and measurements. Also, two lists of recommended features are provided [Juslin & Scherer, 2005]:

- Recommended minimum set of voice cues to index level of affective arousal: F0 (floor), F0(SD), voice intensity(M), speech rate (syllables per minute), and HF 500.

- Recommended minimum set of voice cues to discriminate different emotions: F0 (floor), F0(SD), F0 contour (up/down), jitter, voice intensity (M, SD), speech rate (syllables per minute), pauses (Pd), rhythmic regularity, HF 500, and F1 (M, precision).

2.3.2 Spectral Features

Besides the fundamental frequency, the speech signal contains other frequency related characteristics that are not considered prosodic, but spectral features. Harmonics are multiples (in general integer) of the fundamental frequency, and are characterized by amplitude and frequency. An interesting phenomenon is the capability of the ear to develop harmonics or even the fundamental frequency when they are missing from a sound, so that we will hear them even if they were not produced.

They are produced by the non-linearity of the air flow in the vocal tract. For example, during intense emotional states like anger or stress, the fast air flow develops additional excitation signals caused by vortices located near the false vocal folds. Apparently the harmonics generated by these excitation signals are more intense than the pitch signal. A method for finding the number of harmonics using the Teager energy operator is described in [Ververidis & Kotropoulos, 2006].

Formants are peaks in the frequency spectrum, caused by acoustic resonance in the vocal tract. In general the information needed for the recognition of vowels can be found in the lowest two formants. The main characteristics of formants are frequency and bandwidth. The bandwidth can be analyzed in order to detect slackened speech from the improved articulated one, which is a sign of stress or depression. The formant bandwidth is narrow with steep flanks in improved articulated speech, and becomes gradual for slackened speech. Linear prediction analysis can be successfully used for formant estimation [Ververidis & Kotropoulos, 2006].

Mel frequency cepstral coefficients (MFCC), which are generally used in speech recognition, are also spectral features successfully used for emotion detection. They provide a better representation of signals than equally spaced frequency bands, because the frequency bands on the mel scale provide an approximation of the human auditory system. A variant of MFCC which contains frequencies between 20 and 300 Hz was used to model pitch and proved to be more effective than pitch features [Neiberg *et al.*, 2006b] [Neiberg *et al.*, 2006a]. Also, it seems that in practice log-frequency power coefficients (LFPCs), which include pitch information, provide good results [Nwe *et al.*, 2003], [Ververidis & Kotropoulos, 2006]. Also, MFCCs can be successfully replaced by linear predictive cepstral coefficients (LPCC) or mel filter bank (MFB) features [Steidl *et al.*, 2008].

Another type of spectral features, performing similar to MFCCs, is RASTA-PLP, an acronym for Relative Spectral Transform - Perceptual Linear Prediction. They were successfully used in [Truong & van Leeuwen, 2007].

For examples of studies using spectral features, the reader can refer to: [Devillers & Vidrascu, 2006], [Matos *et al.*, 2006], [Neiberg *et al.*, 2006b], [Krajewski & Kroger, 2008], [Truong & van Leeuwen, 2007], [Zhang *et al.*, 2004] and [Batliner *et al.*, 2006].

2.3.3 Voice Quality Features

Studies on voice quality report that there is a strong correlation between voice quality features and emotions. Examples of voice qualities are neutral, whispery, breathy, creaky, and harsh or falsetto voice. The challenge is to estimate the glottal source. A technique used for cancelling the effects of the vocal tract and separating the source signal is inverse filtering.

Jitter and shimmer are an alternative to inverse filtering. Jitter measures cycle to cycle variation of period length while shimmer measures cycle to cycle variations of peak or average amplitude. Harmonics to noise ratio (HNR) measures the degree of periodicity in a sound and is another voice quality feature. Applications involving these features can be found in [Krajewski & Kroger, 2008], [Vlasenko *et al.*, 2007], [Hu *et al.*, 2007], [Devillers & Vidrascu, 2006] and [Razak *et al.*, 2005].

An important issue concerning features is the unit of analysis. In the case of acted databases, since actors are asked to speak utterances expressing different emotions, it is obvious that analysis at the utterance level gives the best results. Smaller units like frames, chunks determined using acoustic characteristics or segments between pauses have also been investigated [Hu *et al.*, 2007], [Vogt & Andre, 2005], [Vlasenko *et al.*, 2007], [Kim *et al.*, 2007] and [Shami & Kamel, 2005]. What appears from these papers is that smaller segments, like frames for example contain a lot of information, while segments of speech between pauses can be more suitable segments where emotion can actually be encountered. However, the recognition results are still better on the utterance level, but the combination of utterance and shorter segments level definitely improves the results. In [Datcu & Rothkrantz, 2006] the efficiency of using distinct numbers of frames per speech utterance is investigated. It appears that splitting the utterances in more segments can sometimes lead to improvements, but there is not a high difference compared to the results of analysing the entire utterance.

2.3.4 Linguistic Features

With regard to linguistic features, the affective states associated with words can be found using Whissell's dictionary of affect [Whissell, 1989] or Ortony's Affective Lexicon [Ortony *et al.*, 1988]. There are not many attempts of combining acoustic and linguistic cues, but in all cases the results are improved by using a fusion of the two modalities rather than either one of them.

One approach for emotion detection by linguistic cues is the bag-of-words approach which means that the text is represented as an unordered collection of words, disregarding grammar but keeping track of word frequencies. This probabilistic approach is based on

the estimation of the probability of one emotion giving a certain sequence of words, similar to the language models from speech recognition. The most frequently used are unigrams and bigrams.

However, most of the times sentences convey emotions by their underlying meaning, which cannot be captured easily. Advanced methods make use of statistical natural language processing, affect models based on complex psychological theories and common sense knowledge for a deeper understanding.

Linguistic information has been used for emotion recognition in [Ang *et al.*, 2002], [Chuang & Wu, 2004], [Devillers & Vidrascu, 2006], [Fragopanagos & Taylor, 2005], [Graciarena *et al.*, 2006], [Batliner *et al.*, 2003], [Lee *et al.*, 2004], [Lee & Narayanan, 2005], [Schüller *et al.*, 2004], [Schüller *et al.*, 2005a] and [Schüller *et al.*, 2005c].

2.3.5 Features and Emotions

	Pitch				Intensity		Timing	
	Mean	Range	Variance	Contour	Mean	Range	Speech rate	Transmission duration
Anger	>>	>	>>		$>>_M, >_F$	>	$<_M, >_F$	<
Disgust	<	$>_M, <_F$			<		$\ll_M, <_F$	
Fear	>>	>		↗	=>			<
Joy	>	>	>	↘	>	>		<
Sadness	<	<	<	↗	<	<	$>_M, <_F$	>

Explanation of symbols: >: increases, <: decreases, =: no change from neutral, ↗: inclines, ↘: declines. Double symbols indicate a change of increased predicted strength. The subscripts refer to gender information: M stands for males and F stands for females.

Figure 2.5: Effects of emotions on acoustic features [Ververidis & Kotropoulos, 2006]

Five of the basic emotions and their correlation with a set of features are discussed in [Ververidis & Kotropoulos, 2006]. The five emotions are: anger, disgust, fear, joy and sadness. The analyzed features are pitch, intensity and duration, for which several statistics have been extracted. The trends of these features are depicted in Figure 2.5.

The most intense emotion, with the highest energy and pitch level, is anger. Some differences between emotion expressions between genders appear: angry males generate speech with a higher level of energy than and with a lower speech rate than angry females. Also, it is interesting to note that males speak faster when they are sad than when they are angry or disgusted.

A more extensive study by Scherer, Johnstone and Klasmeyer [Scherer *et al.*, 2003] presents an overview of empirically identified major effects of emotion on vocal expression (see Table 2.2). The results are presented for arousal, happiness, anger, sadness, fear and boredom. The first column presents the acoustic parameters, while the next columns show empirically detected differences in parameter values from a normal value due to influences by the studied emotions. As the authors mention, for some of the parameters which are not often used, the results are based on very few studies, or even only one. Therefore, it is advisable to consider the table as a set of empirical expectations and not set of established results.

Table 2.2: Synthetic Review of the Empirical Findings Concerning the Effect of Emotion on Vocal Parameters [Scherer, 2003]

Acoustic Parameters	**Arousal/ Stress**	**Happiness/ Elation**	**Anger/ Rage**	**Sadness**	**Fear/ Panic**	**Boredom**
Speech Rate and Fluency						
Number of syllables per second	$>$	$\geq$	$\neq$	$<$	$>$	$<$
Syllable duration	$<$	$\leq$	$\neq$	$>$	$<$	$>$
Duration of accented vowels	$\geq$	$\geq$	$>$	$\geq$	$<$	$\geq$
Number and duration of pauses	$<$	$<$	$<$	$>$	$\neq$	$>$
Relative duration of voiced segments			$>$		$\neq$	
Relative duration of unvoiced segments			$<$		$\neq$	
Voice Source - F0 and Prosody						
F0 mean [3]	$>$	$>$	$>$	$<$	$>$	$\leq$
F0 5th percentile [3]	$>$	$>$	$=$	$\leq$	$>$	$\leq$
F0 deviation[3]	$>$	$>$	$>$	$<$	$>$	$<$
F0 range[3]	$>$	$>$	$>$	$<$	$\neq$	$\leq$
Frequency of accented syllables	$>$	$\geq$	$>$	$<$		
Gradient of F0 rising and falling [3 6]	$>$	$>$	$>$	$<$	$\neq$	$\leq$
F0 final fall: range and gradient[3 4 7]	$>$	$>$	$>$	$<$	$\neq$	$\leq$
Voice Source - Vocal Effort and Type of Phonation						
Intensity (dB) mean [5]	$>$	$\geq$	$>$	$\leq$		$\leq$
Intensity (dB) deviation [5]	$>$	$>$	$>$	$<$		$<$
Gradient of intensity rising and falling [2]	$>$	$\geq$	$>$	$<$		$\leq$
Relative spectral energy in higher bands [1]	$>$	$>$	$>$	$<$	$\neq$	$\leq$
Spectral slope [1]	$<$	$<$	$<$	$>$	$\neq$	$>$
Laryngealization		$=$	$=$	$>$	$>$	$=$
Jitter[3]		$\geq$	$\geq$		$>$	$=$
Shimmer[3]		$\geq$	$\geq$		$>$	$=$
Harmonics/Noise Ratio [1,3]		$>$	$>$	$<$	$<$	$\leq$
Articulation - Speed and Precision						
Formants - precision of location	?	$=$	$>$	$<$	$\leq$	$\leq$
Formant bandwidth	$<$		$<$	$>$		$\geq$

Other reviews of characteristics of specific emotion with regard to acoustic features can be found in [Nwe *et al.*, 2003], [Cowie *et al.*, 2001] and [Johnstone & Scherer, 1999]. Finding the most suitable set of features, that will yield the best performance and will include no redundancies, is still very challenging. It appears that a popular approach is extraction of a high number of features and then and then deciding for the best subset, e.g. [Schüller *et al.*, 2005a], [Vogt & Andre, 2005]. However, the obtained sets are database dependent, and high differences are observed between acted and natural speech datasets. Finding a set of features that is optimal and data independent is still an unsolved problem.

2.4 Classification Techniques

A large variety of machine classifiers are used for recognition of several emotional states from speech. This section will briefly describe the most popular ones, and will give references to papers that focused on these approaches.

2.4.1 Linear Discriminant Analysis

Linear classifiers are the simplest and the fastest classification methods, and their results are often comparable to the ones of more complex classifiers. A simple linear classifier is linear discriminant analysis (LDA). This method is successfully used in statistics and machine learning for finding linear combinations of features which can best separate more classes. The idea of LDA is to use a transform function that will change the coordinate system in such a way that will maximize the difference between classes, as shown in Figure 2.6 . Examples of research based on LDA are [Batliner *et al.*, 2003] [Krajewski & Kroger, 2008].

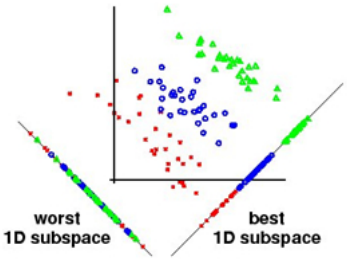


Figure 2.6: Dimensions for LDA (Image by Schwardt and du Preez)

[1]Depends on phoneme combination, articulation precision or tension in the vocal tract.
[2]Depends on prosodic features like accent realization, rhythm, etc.
[3]Depends on speaker-specific factors like age, gender, health, etc.
[4]Depends on sentence mode.
[5]Depends on microphone distance and amplification.
[6]For accented segments.
[7]For final portion of sentences.

2.4.2 K-Means and K-Nearest-Neighbors

K-means and k-nearest-neighbors (KNN) are linear classifiers as well. The k-means algorithm is based on Euclidian distance, and assigns a new sample to a class according to the distance between that sample and the mean of each class. K-nearest-neighbors on the other hand, classify a new item by a majority vote of its k nearest neighbors, so that the item will belong to the most common class amongst its neighbors.

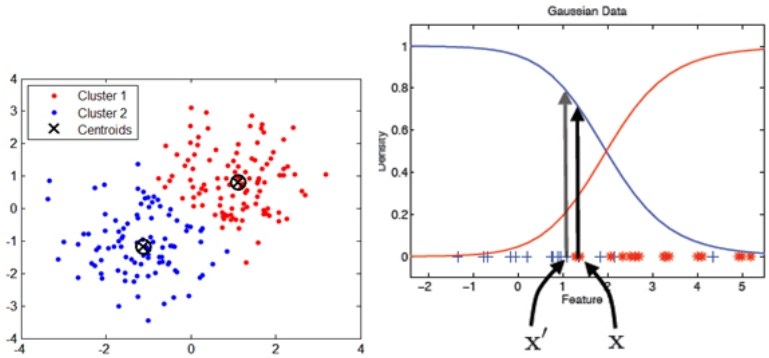


Figure 2.7: K-means (left) and 1NN (right)

Research conducted by [Pao & Chen, 2003], [Lee & Narayanan, 2005], [Petrushin, 1999] and [Schüller *et al.*, 2004] uses linear classification methods for speech emotion recognition.

2.4.3 Bayesian Networks

Bayesian networks (BN), also known as belief networks, are directed acyclic graphs. The nodes are related to state variables from a finite set of states. The edges between nodes are directed and express the conditional probabilities of nodes and their parent nodes. The joint probability distribution provides a complete representation of the conditional probabilities and of the network's structure.

Attempts to use dynamic Bayesian networks are present in the work of [Schüller *et al.*, 2005a] and [Ververidis *et al.*, 2004]. In [Schüller *et al.*, 2005a] the approach was to perform emotional key-phrase spotting by belief networks for the inclusion of context information from the utterance, such as negations.

2.4.4 Hidden Markov Models

Hidden Markov models can be regarded as the simplest dynamic Bayesian networks (DBN). They have a long tradition in speech recognition based on the idea that the statistics of voice are not stationary. The use of HMM and their capability to model the temporal behavior of speech as opposed to the global statistics approach has more advantages. For example it can be useful for dealing with phenomena such as the rising pitch in interrogatory sentences which have no emotional connotations. Also preprocessing does not have to wait until the entire utterance has been pronounced, providing therefore capabilities for real-time application [Nogueiras *et al.*, 2001].

HMMs are successfully used in emotion recognition from speech signals. [Lee *et al.*, 2004], [Nwe *et al.*, 2003], [Schüller *et al.*, 2003], [Matos *et al.*, 2006], [Lin & Wei, 2005], [Nogueiras *et al.*, 2001] and [Vlasenko *et al.*, 2007].

2.4.5 Gaussian Mixture Models

Gaussian mixture models (GMM) are among the complex statistical methods. The original feature probability density function (pdf) is approximated with a set of weighted Gaussians. GMMs are used in [Neiberg *et al.*, 2006b], [Graciarena *et al.*, 2006], [Truong & van Leeuwen, 2007] and [El Ayadi *et al.*, 2007]. GMMs can be used as the state dependent probability distribution functions of HMMs. Therefore, a GMM is equivalent to a hidden Markov Model with just one state.

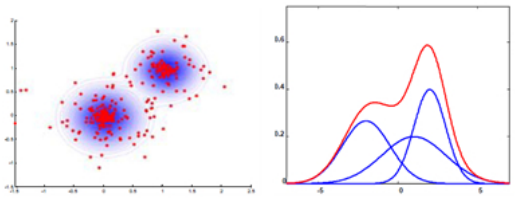


Figure 2.8: Examples of GMMs

2.4.6 Artificial Neural Networks

Artificial Neural Networks (ANN) are another example of more sophisticated classifiers. They are made up of a number of interconnected artificial neurons, which mimic the behavior of biological neurons. Several architectures with different numbers of neurons and layers can be used. A multi-layer perceptron is a feedforward neural network with three or more layers of neurons. They use non linear activation functions and are able to classify data that is not linearly separable or separable by a hyperplane. A challenge is to find the best architecture of the network, that will assure the best performance for the given datasets.

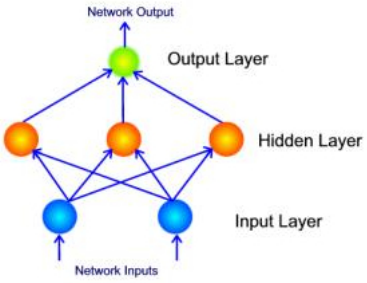


Figure 2.9: Neural network arhitecture

Examples of emotional speech recognition with the use of neural networks are [Fragopanagos & Taylor, 2005], [Batliner *et al.*, 2003], [Razak *et al.*, 2005],

[Nicholson *et al.*, 1999], [Petrushin, 1999] and [Krajewski & Kroger, 2008].

2.4.7 Support Vector Machines

Support Vector Machines (SVM) are among the most popular classifiers in speech emotion recognition, due to their high generalization capability. Given the separation problem of two classes, a support vector machine will try to determine a hyperplane that can completely distinguish these two classes. The idea is to find a hyperplane that maximizes the margin between the two datasets, and the samples that lie on the margin are called support vectors. The chosen hyperplane has the largest distance to the neighboring samples from both classes. As opposed to traditional SVMs that can only construct hard decision boundaries with no probability outputs, [Chuang & Wu, 2004] propose using SVMs with continuous probability outputs.

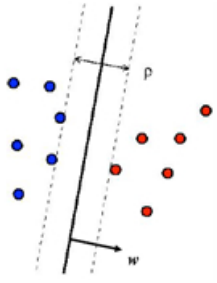


Figure 2.10: SVM classifier

SVMs are used for emotion recognition from speech signals in [Chuang & Wu, 2004], [Schüller *et al.*, 2004], [Lee *et al.*, 2004], [Devillers & Vidrascu, 2006], [Graciarena *et al.*, 2006], [Truong & van Leeuwen, 2007], [Schüller *et al.*, 2005a], [Lin & Wei, 2005] and [Hu *et al.*, 2007].

2.4.8 Fuzzy Rules

In the same way as Bayesian networks are used for dealing with uncertainty, fuzzy rules are used for approximate reasoning based on the idea of fuzzy sets. In classical set theory, membership of an element in a set is binary assessed. In contrast, for fuzzy sets which are a generalization of classical sets, allow gradual assessment of membership by the use of membership function with values in the interval [0,1].

In [Razak *et al.*, 2005], emotion recognition is performed both with neural networks and fuzzy sets. The results show that fuzzy classification works well when the training set is small, while neural networks give better results for a larger training set. These results can be explained by the way in which the classifiers perform recognition. For NN, the more data is used for training, the better trained the network is, and the better the results. On the other hand, for the fuzzy model, larger training data leads to more variation in the data and larger variation in the fuzzy sets, so the mean and variances stored in the knowledge base become less accurate. Fuzzy classification is used also for emotion recognition by the robot head MEXI [Austermann *et al.*, 2005].

2.4.9 Decision Trees

Decision trees are predictive models that map from observations about an item to conclusions about its target value. In [Forbes-Riley & Litman, 2004], decision trees were boosted with AdaBoost. [Ang *et al.*, 2002] used a bruteforce iterative feature selection algorithm to find a minimal set of useful features and avoided this way the greedy search problem, and they arranged the most used features higher in the tree. Binary decision trees have been sucessfully used also by [Cichosz & Slot, 2007].

2.4.10 Ensembles

As the performance of certain classifiers seems to be bound at certain limits, a new direction is to combine different classifiers and build a meta-classifier, in order to achieve better results. A rather simple method is the unweighted vote within the base classifiers. The class predictions of the base classifiers are summed, and the result of the meta-classifier is the class with the highest number of votes. In [Morrison *et al.*, 2007], unweighted vote shows better performance than each base classifier in turn (random forests, K* instance-based learner, KNN, MLP) except SVM.

Stacked generalization is another approach of combining results from more classifiers. A meta-classifier is trained with the target classes of several base classifiers. It uses the results of the base classifiers' predictions and the target classes in order to determine which classifiers are correct and which not and generates a higher level prediction. StackingC, an improvement of the original method, uses only target predictions which were associated with the target class during training and testing, reducing therefore the dimension of the meta-learning phase. StackingC outperforms the base classifiers in [Morrison *et al.*, 2007], [Schüller *et al.*, 2005c] and [Schüller *et al.*, 2005b].

A high number of feature and a high number of classifiers are combined in Combining efforts for Improving Automatic Classification of Emotional User States, [Batliner *et al.*, 2006] using the ROVER framework for speech recognition.

2.5 Conclusion

After presenting the state of the art in emotion recognition from speech, it is time to come back to the main problems we want to discuss, which were stated in the first chapter. As can be noted so far, there is no clear recipe on how to develop a successful emotion recognizer. For almost all decisions that one needs to make there are several options, each with pros and cons. As most of the results are depending on the chosen database, it becomes difficult to compare results stated by different researchers and to decide which is the best approach. However, there are already studies investigating the differences between some approaches.

The first step in designing an automatic speech emotion recognizer is deciding upon the emotion model. The current research is governed by the discrete and the dimensional approaches. In most of the cases a basic set of emotions is used e.g. the six basic emotions of Ekman. This is mostly the case with acted databases. The trend is to use more high level emotions in classification. The dimensional approach is less used even though there

are strong opinions encouraging it, coming mainly from psychologists. Recent research proposes new models, as similar as possible to human cognitive perception. They argue that humans perceive emotions in a continuous space, but they have the capability to categorize emotions and form different clusters for which they recognize certain names.

The problem of emotion models is strongly related to the one of emotional databases. The databases contain recordings of acted, elicited or natural speech, and what is of high importance for emotion recognition, they are annotated in general with a discrete set of emotional labels. In the case of natural speech, for example in applications meant to improve call center applications, coarse labels like positive and negative are also sometimes assigned. The annotation of databases can be application dependent, but in most of the cases a discrete set of emotion labels like the six basic emotions or other more high level ones are assigned.

Up to a certain point in time, only acted databases were used. Actors, not always professionals, are asked to pronounce some previously defined utterances expressing a certain emotion. In a later stage, human judges can be asked to check whether the actor succeeded in conveying that emotion. Nowadays, many opinions are against acted databases, mainly because of the tendency to exaggerate that actors often have. Databases with elicited emotions are considered a step forward, since the emotional states are induced to the participants by the use of stories or music and more natural emotional utterances are expected.

Of course, the best are databases with spontaneous speech, where all issues about lack of genuineness are no longer valid. However, real data comes with a lot of extra problems. First of all, recording of natural speech are rare, and most of the time not available due to privacy issues. If acted databases contained full blown emotions, and there was a balance between the numbers of utterances for each emotion, this is no longer the case for spontaneous data. For any natural recording, most of the speech is neutral, and when emotions appear, they are usually not very intense and they are difficult to recognize even by humans. The sparse data problem leads to difficulties in grouping together segments of the same emotion, and having enough data for each group so that a classifier could be trained. Another problem is the annotation of spontaneous data, which can be very difficult and even between experienced labelers there are lots of disagreements, in which case those data should be discarded. The high quality recording from acted databases are replaced with more noisy sound, which has been captured in most cases in a Wizard of Oz scenario, in a virtual reality set-up, or has been extracted from a call-center.

There are still researchers that argue for the use of acted data to avoid all the problems caused by natural speech. They claim that the problems lie in the protocol that has been used for recordings, and not in the acted emotion itself. They encourage longer dialogues, allowing actors to improvise and believe that emotion will grow more real and still all disadvantages of real data will be avoided.

Two kinds of features can be extracted from speech: acoustic features, and linguistic features. Both of them can be used for emotion detection, either together or individually. All experiments showed that the combination leads to better results than each of the individual ways, and acoustic features perform better than linguistic ones. Early fusion and late fusion experiments, as well as fusion using a neural network were reported. The

improvement caused by adding linguistic features to the acoustic results is in average 5% absolute. However, the amount of research in this direction is still small, and probably a new fusion modality and more accurate linguistic classification will lead to better results.

There are many indications from psychologists of the correlation between acoustic changes and the emotional states. Most of these indications address the problem of the six basic emotions, so we can find information and empirical studies showing how the features change over time according to emotions. In studies on emotion recognition, pitch, energy, voice quality and spectral features are the most popular. The individual values of these features are not very useful by themselves, because most information lies in the way these values change over time.

A perfect set of acoustic features that will improve recognition and that can be effective on any database has not been found yet. Recent studies employ a large amount of features, sometimes even thousands, and then optimize this set using principal component analysis, genetic algorithm or other methods to reduce redundancy and increase representativeness. From these experiments it appears that feature sets are database dependent.

Some studies investigated in more depth the differences between acted and spontaneous data. Their findings show that different features are useful for acted and for spontaneous data, and also different units of analysis. In the case of acted data, actors are asked to express emotions while saying an utterance. The same emotion is expected during the entire utterance, and therefore in the case of acted data, utterance level analysis is the most efficient. Things change when it comes to natural speech. The length of emotional speech within an utterance can be longer or shorter. There is no recipe for finding it, and analyzing exactly that area. However, utterance level analysis still yields the best results. Using utterance level, a lot of information that lies in lower levels, like frame, phoneme, or segments between pauses, is lost. Studies show that the best results are obtained using a combination of utterance level and a lower level, which are better than each individual result.

After the feature sets are obtained, different kinds of classifiers are used for training and testing. Experiments using linear classifiers, Gaussian mixture models, neural networks, hidden Markov models, Bayesian networks, random forests, support vector machines and others are reported. In general it is difficult to compare the results within different experiment, because different general conditions are applied: different databases, units of analysis, feature sets, some report speaker dependent, some speaker independent results, etc. There are also studies investigating the capabilities of such classifiers. It appears that support vector machines are among the most successful.

Recent approaches combine the results of different weak classifiers in ensembles, and their combination is a stronger classifier, which in almost all cases gives better results than each individual classifier. However, sometimes the results can be very close to the ones of the best performing classifier, and it might not worth using the entire ensemble. Among the reported classifiers, hidden Markov models have the ability to model temporal behavior, and can be used for coping with pitch changes that appear for example in interrogatory sentences, that can be easily mistaken for emotional changes. Also, this type of classification is useful for real time application, because preprocessing does not have to wait until the entire utterance has been pronounced.

Linguistic features on the other hand have been less investigated. There are several approaches for detecting emotion using textual cues, like keyword spotting, lexical affinity, statistical natural language processing and hand-crafted models. Most studies use keyword spotting and the performance is rather good. Better results are expected for more complex approaches like the last two mentioned.

A lot of research has been done in the field of emotion recognition. The best results are obtained using multimodal approaches, including emotion extraction from visual cues as well, but they are not always available. However, speech itself contains two modalities, the acoustics and the meaning of words. Their combination leads to better results than each of them taken separately.

For the future, it would be good if some standards for emotional databases will be considered and followed, which will enhance the comparison between different results. There is still place for feature set optimization, and finding something more general that would work for any database. An idea that proved to be successful was the use of more corpora for training and testing, but this option has also disadvantages. More research on fusion between linguistic and acoustic classification will also be beneficial.

3 Resources

As opposed to the previous chapter, which gives some general lines about the databases, features and recognition approaches in the emotion recognition area, this chapter gives an overview of the specific resources used for our research including motivations for our choices and theoretical details.

First, the emotional databases which were used as part of our experiments are described in detail. Using more databases is one of the prerequisites for building a more general system, and the characteristics of the database have a strong influence on the success.

The chapter continues with a closer look on the feature sets used for our project. There are two important types of features of the speech signal: prosody features, which are captured at the utterance level, and spectral features, which are extracted frame-wise. If for the spectral features there are some standard choices, this is not the case for the prosody features (see section 2.3). Our choice for the utterance-level features is justified in this section. Also, the modalities and the tools we have employed for feature extraction are described.

In the final part we introduce the methods for classification which we have used in our research: support vector machines, Gaussian mixture models, modalities for fusion and calibration of the results from more classifiers. For each method we also describe the tools which we have used.

3.1 Emotional Databases

3.1.1 The German Database of Emotional Speech

The German database [Burkhardt *et al.*, 2005] is probably the most often used database in the context of emotion recognition from speech, and also one of the few for which some results can be compared. We will refer to it further on simply as *Berlin*, for ease of communication. It is one of the databases with acted emotional content.

It contains audio recordings of ten actors, five male and five female. The actors were found through newspaper advertisements and the purpose was to find non-professional actor that would not perform in an exaggerate manner. Three professional listeners selected 10 out of the 40 participants based on the recognizability and naturalness of their performance. It is interesting to notice that all but one had indeed some acting education.

The actors had to portray emotions from the following set: anger, disgust, fear happiness, sadness, surprise and neutral. In order to facilitate their ability to perform naturally, they were asked to induce themselves a specific state by remembering events that have caused them such emotions. This technique is known as the Stanislavski

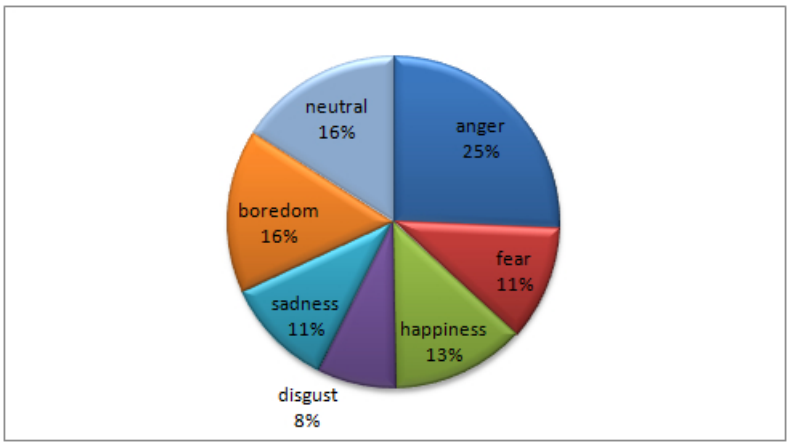


Figure 3.1: Amount of recordings from each emotion for the Berlin database

method.

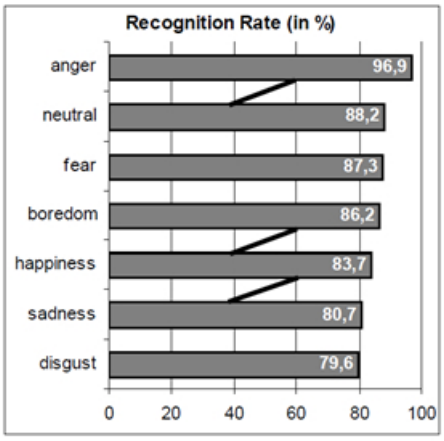


Figure 3.2: Human recognition rates and significant differences between emotions

A total of approximately 800 sentences were recorded. After a perception test carried out by 20 participants, the amount was reduced to around 500 samples. The selected utterances have a human recognition rate better than 80% and naturalness scores of more than 60%. The recognition rates per emotion are presented in Figure 3.1.1, where the connection lines should be interpreted as significant differences between emotions.

The textual content of the utterances is predefined. The choice was made in favour of everyday life utterances that have no emotional content, because they contain no emotional bias and are easy to remember and can be said naturally by participants.

The recordings were taken in an anechoic chamber with high quality technology. The initial sampling frequency was 48kHz. But later it was downsampled to 16kHz. As can be seen from the histogram in Figure 3.3, the utterances' lengths are mostly short, around two seconds. They are ranging from 1.22 to 8.97 seconds, with a mean length of 2.76. A total of 22.80 minutes of recording are available. The recordings are also phonetically

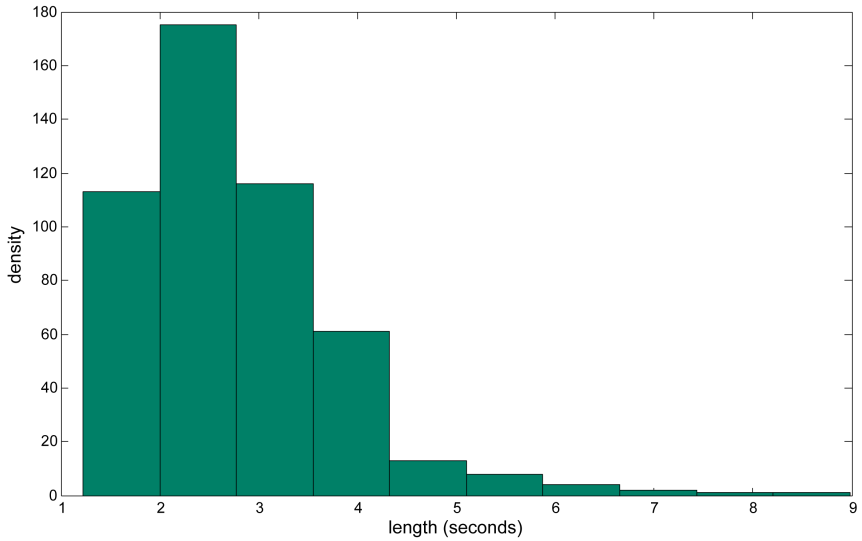


Figure 3.3: Histogram of utterances' lengths in seconds for Berlin database

labeled with some markers for voice-quality, phonatory and articulatory settings and articulatory features.

3.1.2 The Danish Emotional Speech Database

The Danish database [Engberg & Hansen, 1996], also known as DES, was developed within the framework of the VAESS project, which aims at improving the quality and range of synthetic voices by showing a range of emotions. The recordings took place in an acoustically damped sound studio, using a high quality microphone.

The Danish database is also part of the databases with acted emotions. Four actors of radio theatre (two male and two female) were employed to portray the following emotions: neutral, surprise, happiness, sadness and anger. The textual content of the utterances was chosen based on the assumption that listeners that would attempt to recognize the emotions would first use the semantic meaning of the utterance. Therefore, they decided to use semantically neutral text that would be prompted to the actors.

Each actor was asked to express the previously mentioned emotions. The prompting text for each speaker contained 2 single words, 9 sentences and 2 passages of fluent speech Furthermore, there are 8 passages and 10 sentences for target voices. It can be seen from the previous image that the amount of recordings is the same for each speaker, so in the case of this database the data is more balanced than for Berlin.

In order to test the ambiguity of the recordings, a listening test was performed. 20 normal-hearing listeners were employed for to judge the emotional content for each actor and for each utterance. They were also asked to rate the difficulty of the task. It appears that the listeners have the ability to adapt to a speakers voice, since the recognition rate proved to be higher for the second part of the recordings of the same speaker. Also, there were differences in the recognition rates between genders: female participants were able to recognize the correct emotion in 69% of the cases, while men in 66% of the cases.

Figure 3.4: Amount of recordings from each emotion for the DES database

Listeners ? / Actors ?		RESPONSE in %				
		Neu.	Sur.	Hap.	Sad.	Ang.
STIMULUS	Neu.	**60,8** 57,8-63,7	**2,6** 1,8-3,8	**0,1** 0,0-0,6	**31,7** 29,0-34,6	**4,8** 03,7-06,3
	Sur.	**10,0** 8,3-12,0	**59,1** 56,1-62,1	**28,7** 26,0-31,5	**1,0** 0,5-1,8	**1,3** 0,7-2,1
	Hap.	**8,3** 6,8-10,1	**29,8** 27,1-32,7	**56,4** 53,4-59,4	**1,7** 1,1-2,7	**3,8** 2,8-5,1
	Sad.	**12,6** 10,7-14,8	**1,8** 1,2-2,8	**0,1** 0,02-0,6	**85,2** 82,9-87,2	**0,3** 0,1-0,9
	Ang.	**10,2** 8,5-12,2	**8,5** 6,9-10,3	**4,5** 3,4-6,0	**1,7** 1,1-2,7	**75,1** 72,4-77,6
Total		**20,4** 19,3-21,5	**20,4** 19,3-21,5	**18,0** 16,9-19,0	**24,3** 23,1-25,5	**17,0** 16,0-18,1

Figure 3.5: Confusion matrix for the listening test (DES)

Figure 3.5 shows the confusion matrix of the listening test: the percentages from each emotion that were recognized as any emotion. It can be noticed that for some emotions the recognition rates are not very high, e.g., happiness 56.4%, which was extensively confused with surprise. Also, it was noticed that the longer passages, which contained more prove of emotion were easier to detect that the shorter utterances. The results of the questionnaire show that 75% of the listeners found it difficult or neither easy nor difficult to recognize the emotions.

In total the database contains 30.68 minutes of recordings, ranging from 0.52 seconds to 28.43 second and a mean length of 5.46 seconds. The distribution of the lengths is depicted in the histogram from Figure 3.6.

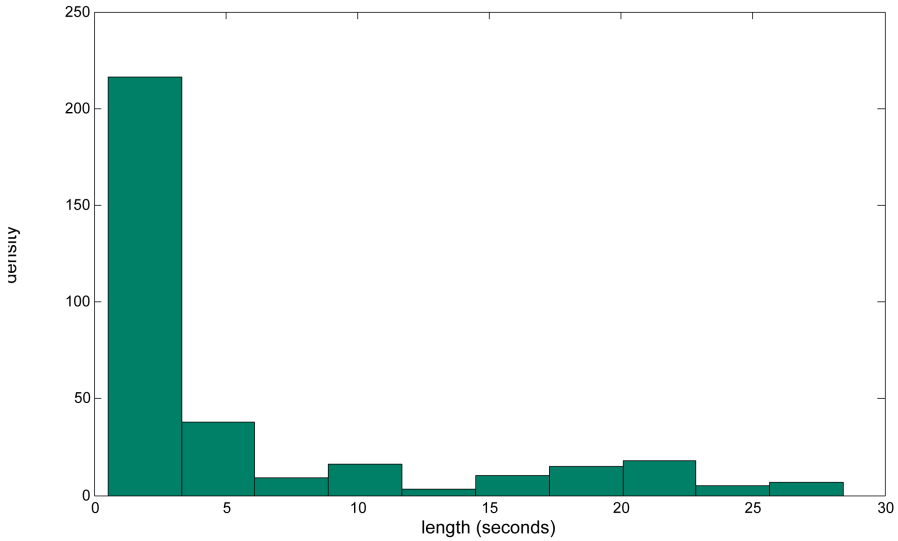


Figure 3.6: Histogram of utterances' lengths in seconds for DES database

3.1.3 The eNTERFACE'05 Audio-Visual Emotional Database

The eNTERFACE'05 database [Martin *et al.*, 2006] was designed with the aim to fulfil the need of a common database for multimodal emotion recognition. We will use the name ENT to denote this database is the rest of this report.

For the purpose of the recordings, 46 persons from 14 different nationalities were asked to react in six given situations. This methodology was chosen in order to elicit different emotions: happiness, anger, sadness, surprise, disgust and fear. A few screen shots from the video recordings are depicted in Figure 3.7. Among the participants there were no professional actors.

The initial idea of letting the subjects to react in their own language was discarded because each language has specific prosodic features, and the purpose was to have the variations in prosody exclusively caused by emotion. The second idea was to let the actors express themselves freely given the presented situation, but since the participants were not English native speakers they had sometimes problems in expressing themselves and finding the right words, and therefore were not spontaneous. The final approach involved a predefined set of answers.

Figure 3.8 shows the distribution of recordings for each emotion. The database is balanced with regards to amount of recordings per emotion.

In order to ensure the quality of the emotional content of the database, two experts decided to remove 4 subjects whose performance was not satisfactory. From the remaining 42 people participating to the recordings, 25 performed satisfactory in all 6 emotions, while the rest made also recordings for certain emotions that had to be discarded.

However, on the database's website a larger number of recordings is available, so we believe that the downloadable material contains all recordings, not only the selected ones, which can of course lead to difficulties in classification. While listening to the samples from the database, we noticed that for one speaker, namely speaker number 6,

Figure 3.7: Examples of actor expressing different emotions from the eNTERFACE'05 database

Figure 3.8: Amount of recordings from each emotion for the eNTERFACE'05 database

the recordings were much longer and contained the portrayal of more than one emotion in one file, including chatting with the staff. We decided to exclude these files from our

analysis, and also from the plots.

Figure 3.9 gives an insight into the lengths of the utterances. The total of 59.06 minutes of recordings vary in length from 1.12 to 6.84 seconds, with an average of 2.81 seconds.

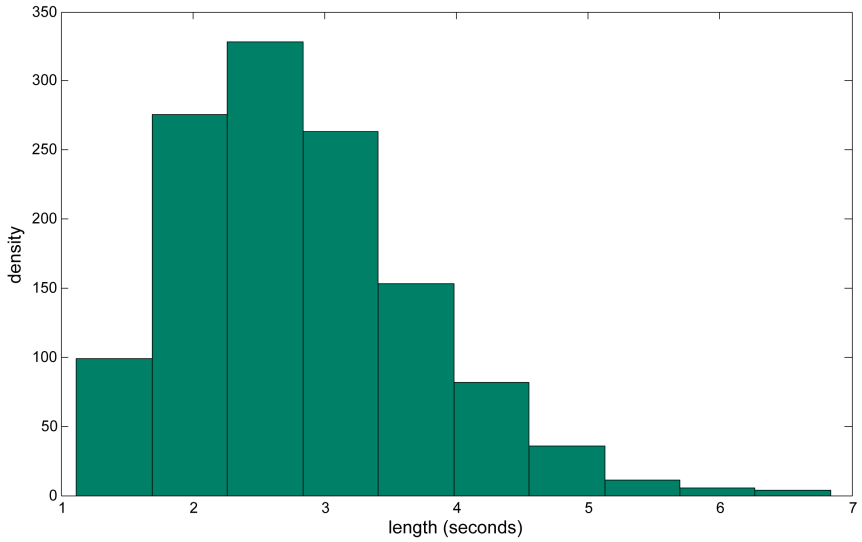


Figure 3.9: Histogram of utterances' lengths in seconds for ENT database

The database was recorded using a standard mini-DV digital video camera and a high quality microphone. The resolution of the camera was 800.000 pixels.

3.1.4 The HUMAINE Database

The HUMAINE database [Douglas-Cowie *et al.*, 2007], developed within the Human-Machine Interaction Network on Emotions (HUMAINE) project, focused on gathering data as naturalistic as possible. The purpose was to provide data reflecting a broad range of feelings, action tendencies and forms of expression that are present in human life.

The database was put together from a large number of smaller audio-visual databases, some of them naturalistic, some containing induced emotions. A total of 48 clips are freely available for research. Table 3.4.1 shows the amount of samples from each of the original databases. Part of the naturalistic clips were extracted from the Belfast Naturalistic Database which contains recordings from TV chat shows and religious programs and discussions between old acquaintances. The other part was extracted from the Castaway Reality Television Database which contains recordings of participants in a competition that had to perform a range of activities (e.g., lighting fire, feeling snakes) on a remote island. The emotional content of rest of the database was induced using different modalities:

- interactions with avatars with different personalities that tend to attract the user in their own state (Sensitive Artificial Listener)

Table 3.1: Amounts and types of clips included in the HUMAINE database

Resource Dataset	**No clips selected**	**Naturalistic / induced**
Belfast Naturalistic Database	10	Naturalistic
Castaway Reality Television Database	10	Naturalistic
Sensitive Artificial Listener (Belfast recordings in English)	12	Induced
Sensitive Artificial Listener (Tel Aviv recordings in Hebrew)	1	Induced
Activity Data/Spaghetti Data	7	Induced
Green Persuasive Dataset	4	Induced
EmoTABOO	2	Induced
DRIVAWORK (Driving under Varying Workload) corpus	1	Induced
GEMEP	1	Induced

- the participants took part in outdoor activities like mountain bike racing or were supposed to feel in boxes that contained objects like spaghetti and buzzers (Activity Data/Spaghetti Data)

- asking participants to play a game where they have to explain a 'taboo' word using gestures and body movement (EmoTABOO)

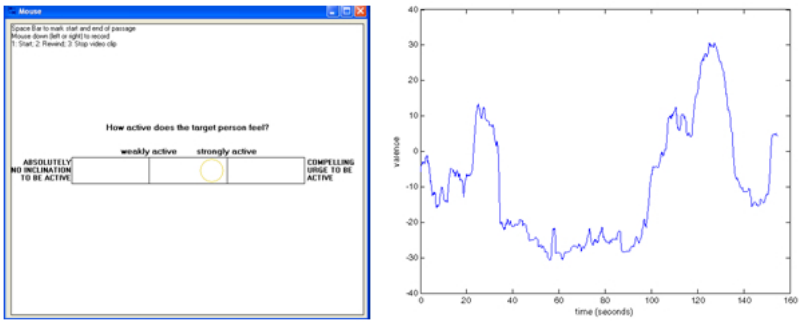
- engaging the participants in persuasive discussion over topics with emotional overtones (Green Persuasive Dataset)

- using a driving simulator and recording the participants in relaxed, normal or stressful conditions, where additional workload is demanded, e.g. mental arithmetic (DRIVAWORK).

Besides the previously mentioned databases, the HUMAINE database contains one additional file from the GEMEP which is an acted database.

The 48 clips were selected carefully is such a way that a broad range of emotions and combinations of emotions are covered in different contexts. Therefore, a more specific style of labeling was adopted. The first type of annotation consists of labels over each clip as a whole. For this purpose a large set of descriptors was used, including emotion-related states, combination types, context labels, key events, everyday emotion words and appraisal categories.

The second type of annotation was done using trace programs and it follows the evolution of different concepts over time. For this purpose, the labelers were annotating a clip by moving a cursor on a range for different dimensions, one at the time. Examples of labeled dimensions are: intensity, activation, masking and valence. Figure 3.10(a) illustrates the interface used by the labelers, and Figure 3.10(b) shows the plot of an resulting annotation file.



(a) Screenshot from trace-annotation program

(b) Result of trace annotation

Figure 3.10: Trace annotation

Using the ANVIL software, annotations on several dimensions can be seen, as shown in Figure 3.11.

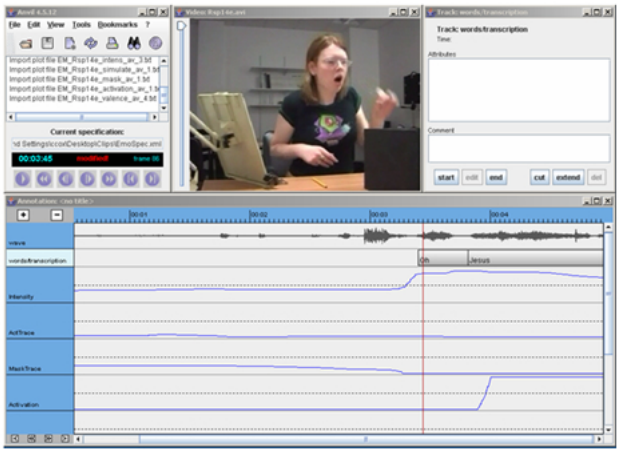


Figure 3.11: Screenshot from ANVIL with more traces visible

Besides the emotional annotation, the database is labeled for speech and language descriptors, including the transcription of words on a time line, gesture descriptors, face descriptors and physiological descriptors. For more details please refer to [Douglas-Cowie *et al.*, 2007].

From the HUMAINE database, we use two types of labeling, both of them time aligned: valence and activation. Also, we used the word transcription in order to look for silences. We were not provided with a final result of the annotation, but with the

Table 3.2: Cronbach's Alpha for labelers of HUMAINE database

Dimension	Mean (all files)	Mean (files with positive Alpha
Activation	0.44	0.68 (-8 files)
Valence	0.57	0.72 (-5 files)

files resulted from each individual labeler. As each labeler moved the cursor at different times, we needed to interpolate the results and resample at equal intervals. Afterwards, the mean of the annotation was calculated.

To inspect the agreement between labelers, Cronbach's Alpha was computed. Alpha is negative whenever the average covariance among the items is negative. Cronbach's Alpha measures how well a set of variables or items measures a single, uni-dimensional construct. If the average covariance among the items is negative, then Alpha is negative.

$$\alpha = (N * \bar{c})/(\bar{v} + (N-1) * \bar{c}),$$

where N is the number of components (items or testlets), $\bar{v}$ equals the average variance and $\bar{c}$ is the average of all covariances between the components.

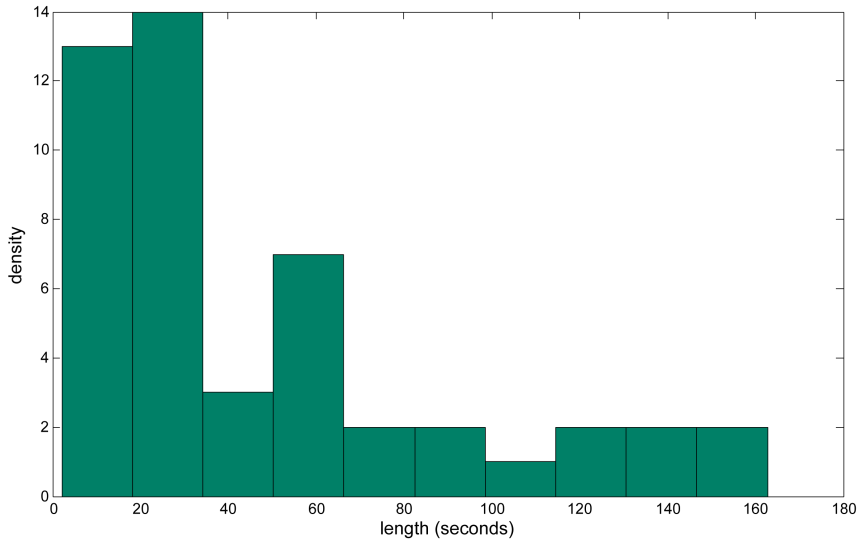


Figure 3.12: Histogram of utterances' lengths in seconds for HUMAINE database

The duration of the clips ranges from 2.12 seconds to 162.74 seconds, with a mean of 48.37 seconds. The total duration of the recordings is 39.50 minutes.

For our analysis, we will use just the part of the clips for which the word transcriptions are provided, due to our interest in the lexical information as well. Therefore, the files in French, German, and Hebrew, as well as the the files containing no words were discarded.

3.1.5 The South-African Database

The South-African Database, to which we will also refer to as *SADB*, was provided by TNO Human Factors. Is very important in our study because is contains genuine recording from call-centers. The data has been labeled according to two classes: *emotional* and *english*. In most of the cases the emotional utterances show anger, dissatisfaction or frustration, while the english ones are just emotionally neutral sentences. The proportion of english and emotional samples from the database is depicted in Figure 3.13.

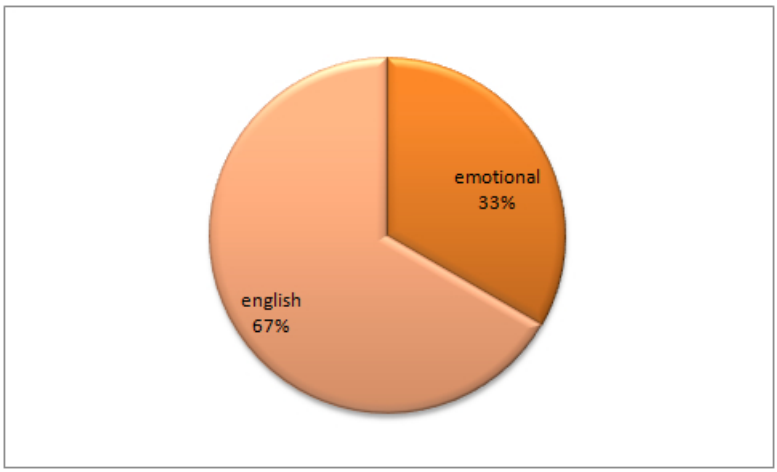


Figure 3.13: Amount of recordings from each emotion for the South-African database

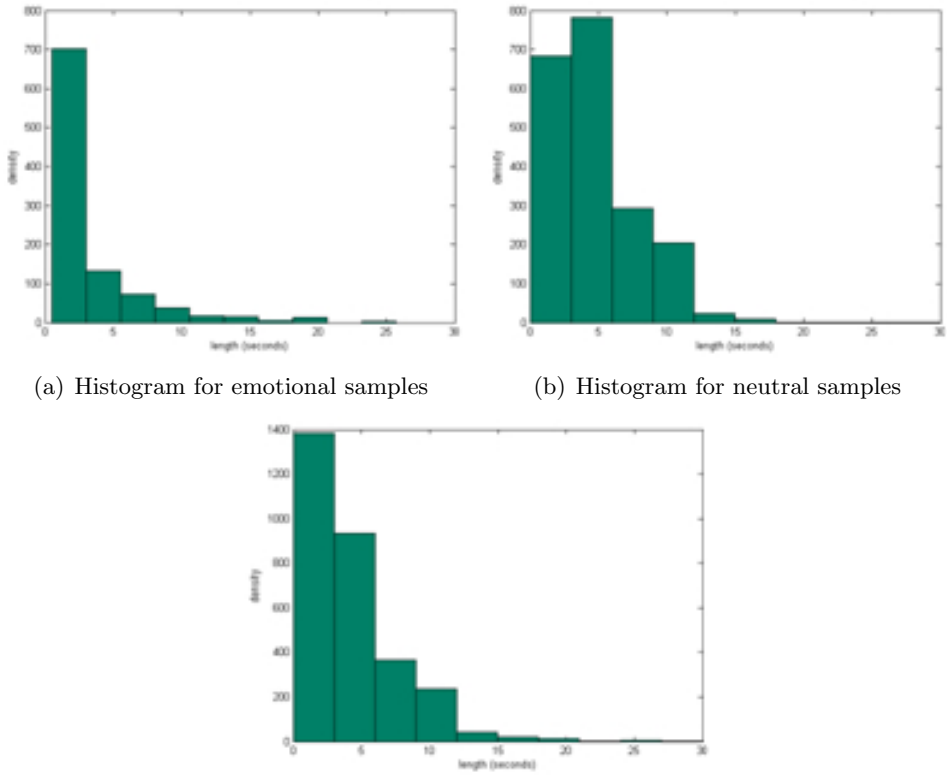
The recorded utterances are mainly short, ranging from 0.006 seconds to a maximum of 29.95 seconds. The average length of the recordings is 4.30 seconds. In total, 215.02 minutes of speech were recorded.

3.2 Speech Features

3.2.1 Prosodic features

3.2.1.1 Finding a General Set of Features

Section 2.3 gave an overview of the most important features used so far for emotion recognition in speech. One can easily notice that choosing an optimal feature set is an open problem. Table 2.1 shows a large number of features and their perceived correlates. Table 2.2 provides an overview of the effects that arise in the acoustic features when different emotional stimuli arise. The choice of the feature set has a great importance in solving the classification problem. Even though there are features for which research has proven their changed behavior when influenced by emotions, it is important to decide which ones to use and also which statistics would comprise the most relevant information. A high number of features and statistics contains a lot of redundancy and also the process becomes very expensive. On the other hand, too little features might miss the important information about emotion. Therefore, a compromise should made be for choosing the



(a) Histogram for emotional samples

(b) Histogram for neutral samples

(c) Histogram for all samples

Figure 3.14: Histogram of utterances' lengths in seconds for the South-African database

right set of features. Please note that this section addresses the selection of features over whole utterances.

It seems to be the trend to generate an increasingly large set of features and statistics of these features, and use a special algorithm like PCA to select the most appropriate ones. For instance in [Schüller *et al.*, 2004] an initial set of more than 200 pitch and energy related features is considered. The features are ranked using LDA and a final set of 33 features is obtained and further used.

In [Schüller *et al.*, 2005c] a large set of 276 features was chosen with the purpose to become mostly independent of the speaker as well as of the spoken content. The features are derived from raw contours of zero crossing rate (ZCR), pitch, first seven formants, energy, spectral development and harmonics to noise ratio (HNR). A distribution of the features is provided in the following table. Frames of 20 ms are analyzed every 10 ms using a Hamming window function. 30 features with the best information gain ratio by SVM-SFFS (Sequential Forward Feature Selection algorithm) were chosen.

In a common initiative of more sites described in [Batliner *et al.*, 2006], 4024 features including linguistic features were extracted: prosodic, spectral, MFCC, part-of-speech, lexical and genetic search features generated automatically based on evolutionary alteration and combination. A final set of 381 features was used for classification.

Another impressive contribution is presented in [Schüller *et al.*, 2007] where a set of 4244 features are extracted, and a final set of 150 features with the higher information gain ratio were selected.

Even though there is an improvement in the results achieved using the reduced feature set, this approach generates models which are very well adapted to the specified database, but the capabilities for generalization are decreased. A strong prove is given in [Vogt & Andre, 2005] where feature sets for acted and spontaneous databases are compared. The Berlin emotional database was chosen as an acted database, and the SmartKom corpus which contains recordings in Wizard-of-Oz (WOZ) scenarios of people interacting with a dialogue system was considered as a spontaneous database.

The feature set is based on pitch, energy and MFCC time series, from which a high number of statistics are derived. Mean, maximum, minimum, range between minimum and maximum, variance, median, first and third quartile and interquartile range of a segment for each of the series earlier mentioned, form the 1280 dimensional feature vector.

The best feature sets were determined using Weka with correlation based feature selection (CFS) and Best-First search. A set is reduced to 90-160 features. The selected features are different for acted and for spontaneous speech. It appears that emotions in acted speech can be well recognized using the pitch values, while for the spontaneous case, MFCCs perform better, and especially the low coefficients were selected and the first derivatives.

Based on these findings, we decide to use a set of features that has a more general character and is expected to have reasonable results on more datasets. The feature set should be easily extracted automatically and it should contain functionals over the whole utterance.

The only hint we found in the literature about a minimum required set of features was in [Juslin & Scherer, 2005]: recommended minimum set of voice cues to discriminate different emotions: F0 (floor), F0(SD), F0 contour (up/down), jitter, voice intensity (M, SD), speech rate (syllables per minute), pauses, rhythmic regularity, HF 500, and F1 (M, precision). However for some of these features an automatic extraction can be more difficult.

A more general approach was followed in [Truong & Raaijmakers, 2008]. The chosen feature set contained the following: mean, standard deviation, range (max-min) and averaged slope were extracted for pitch and intensity. Spectral features like the averaged spectrum, the center of gravity, skewness and Hammarberg index were also measured.

Based on the previous findings we decided to experiment with the feature set presented in Table 3.3.

Because these features were extracted using the program Praat described in the next section, we will sometimes refer to them as 'Praat features'.

3.2.1.2 Praat

Praat [Boersma & Weenink, 2009] is a free standalone program available for most operation systems, that enables automatic analysis of speech signals. Besides speech analysis, Praat offers also solutions for speech synthesis, speech manipulation, labelling and seg-

Table 3.3: The final feature set used in our experiments

Feature type	**Functionals**
Pitch	Mean Standard deviation Range Absolute slope (without octave jumps) Jitter
Intensity	Mean Standard deviation Range Absolute slope Shimmer
Formants	Mean F1 Mean F2 Mean F3 Mean F4
Long term averaged spectrum	Slope Center of gravity Skewness Hammarberg index High energy

mentation, statistics and learning algorithms. There are more modalities for using Praat: through its graphical interface, through scripting, and even by sending commands from other programs using the *sendpraat* program.

For our research Praat was used for feature extraction of a number of pitch, intensity, formants and long term spectrum features. All these features were extracted by means of a script. Even though Praat provides feature extraction capabilities at both frame level and utterance level, we decided to use the utterance level features since their performance looks promising [Vlasenko *et al.*, 2007]. As a preprocessing step, the unvoiced segments from the speech sample were eliminated using the voiced-unvoiced algorithm of Praat. An example of a sound file analyzed with Praat is depicted in Figure 3.15. The middle part of the image shows the spectrum, the pitch (blue), the intensity (yellow) and the formants (red). In the lower part of the image, labels for voiced (V) and unvoiced (U) segments of speech are provided.

As part of our real-time system, which was developed in Matlab, we also used communication with Praat by means of the Sendpraat programs. This program alows us to give commands to Praat from another program. However, it is not possible to send the output from Praat directly to this other program, in our case Matlab. Therefore the communication was attained with the help of an external file.

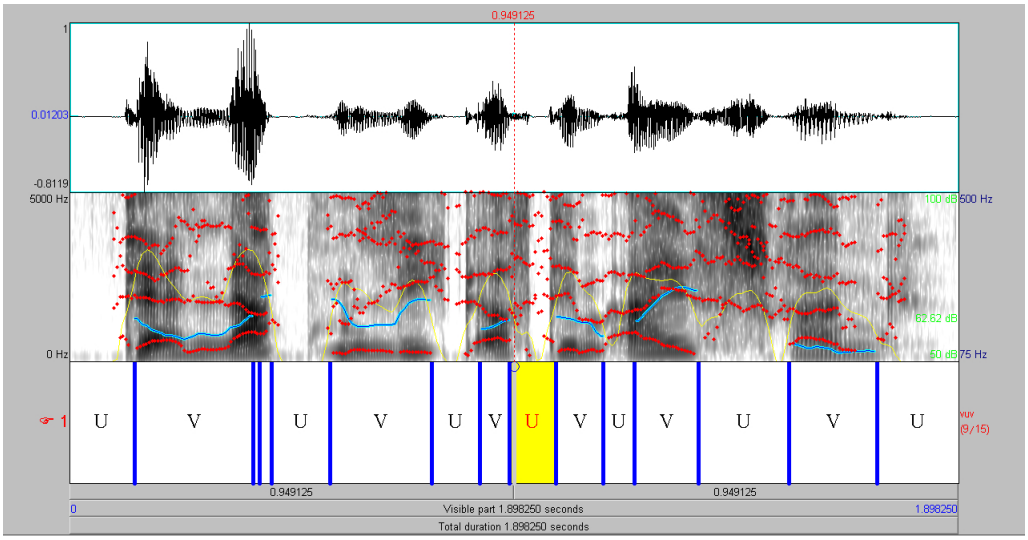


Figure 3.15: Screenshot of Praat voiced-unvoiced detection

3.2.2 Spectral Features

3.2.2.1 Relative Spectral Perceptual Linear Predictive (RPLP) features

Perceptual Linear Predictive(PLP) coding of speech introduced by [Hermansky, 1990] is a method to model speech based on the short term spectrum. The auditory spectrum is approximated by three concepts from the psychophisics of hearing: the critical band spectral resolution, the equal loudness curve and the intensity-loudness power law. The PLP technique is vulnerable to the changes of the short term spectral values due to the frequency response of the communication channel.

A technique called RelAtive SpecTrAl (RASTA) has been developed by [Hermansky *et al.*, 1992] which makes the PLP more robust against convolutional noise. For this purpose, a band-pass filter is applied to the energy in each frequency subband. The result is that short term noise variations are smoothed and the offsets from the static spectral colouration of the speech channel that could appear for instance from a telephone line are removed. We will further refer to these features as RPLP. The most applications of RPLP are in speaker recognition.

In order to extract the features from the sound signal, we have used a tool provided by TNO. The implementation uses the RASTA program developed by the International Computer Science Institute (ICSI). Voice activity detection is done based on energy levels. Every 16 ms, 26 coefficients are extracted for a frame of 32 ms : 12 RPLP coefficients, one energy feature and their deltas.

3.2.2.2 Mel-Frequency Cepstral Coefficients (MFCC) features

The MFCC are coefficients that make up the mel-frequency cepstrum, which is a representation of the short-term power spectrum of sound. This representation is based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

The frequency bands of the MFC are equally spaced on the mel scale, which leads to a good approximation of the human auditory system. Most applications of MFCC are in speech recognition.

The Hidden Markov Model Toolkit (HTK) is a portable toolkit for building and manipulating hidden Markov models originally developed at the Machine Intelligence Laboratory (formerly known as the Speech Vision and Robotics Group) of the Cambridge University Engineering Department (CUED). Its main usage is in speech recognition research, although it has been used for numerous other applications including research into speech synthesis. It consists of a set of library modules and tools available in C source form. The tools provide sophisticated facilities for speech analysis, HMM training, testing and results analysis. As part of our research we have used the HTK tools HCopy and HList for extraction of MFCC features.

3.3 Classification Techniques

From the classifiers we have read about, our first choice falls on support vector machines (SVM) since it appears that they give very good results, many times the best. Gaussian mixture models (GMM) are also chosen, since they are based on a more easy to understand concept, and seem like a good way to start experimenting. Also research shows that a combination between GMM and SVM can lead to improved results, so this is one more reason for choosing the two.

These classification approaches are expected to work fine for predicting classes given speech. Because we are interested in valence and activation values given speech, we also need to look at emotion recognition as at a problem in a continuous space. We chose support vector regression (SVR) as a very handy solution given the previous experience with SVM.

3.3.1 Support Vector Machines

SVM [Cortes & Vapnik, 1995], also mentioned in section 2.4.7, represent a classification method which yields very good results in practice. A general classification setting includes training and test data, and each sample has a list of attributes and a desired label or target value. The training samples are attribute-label pairs $(x_i; y_i)$ where x_i is the attribute vector (or feature vector) and y_i is the label for sample i. The SVM learns relevant information of the patterns in the data, and is used for prediction of labels for new sequences of attributes. This is done in fact by means of an optimization problem:

$$\min_{w,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0.$$

The feature vectors x_i are mapped onto a higher dimensional space using the function ϕ, and the SVM builds a hyperplane that is separating the classes in such a way that

the margin is maximized. In the previous equation, C is the cost of an error.
The first formalization of SVM (1963) was for linearly separable data. Nowadays different kernel functions are used in order to enhance the SVM separation capabilities to non-linear problems. Even though in literature more versions can be found, there are four kernels which are most widely known:

- linear: $K(x_i, x_j) = x_i^T | x_j$
- polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma \geq 0$
- radial basis function: $K(x_i, x_j) = exp(-y||x_i - x_j||^2)$
- sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$,

where γ, r and d are the kernel parameters. For more information please refer to [Boser *et al.*, 1992].

SVM is a solution for two class problems. However, there are several approaches to extend this technique for multi-class problems. The most popular are *one versus one* and *one versus the rest*. If there are n classes, the *one versus one* approach will build $n(n-1)$ classifiers, to make a decision between each pair of classes. The class with the highest number of votes wins. In contrast, the *one versus the rest* approach builds n classifiers, each of them being supposed to discriminate between one class and all the other classes with a 'winner-takes-all' strategy.

Besides classification problems, SVMs are also used for regression. For more details see [Drucker *et al.*, 1996].

The end of the section describes the tools we have used for SVM classification and regression.

3.3.1.1 SVMTorch II

SVMTorch II [Collobert & Bengio, 2001] is an open software tool which implements Vapnik's support vector machines available from the IDIAP Research Institute. It has good capabilities for dealing with large amounts of data and it is very fast. A simple extension to multi-class problems using the one-over-all mechanism is also provided. The tool is relatively easy to use as a console application. As a preliminary step, the data has to be prepared in the right format:

<Number n of training/testing samples> <Dimension d of each sample+1>
$< a_{11} > < a_{12} > < a_{13} > .... < a_{1d} > < a_{1\_out} >$
.
.
.
$< a_{n1} > < a_{n2} > < a_{n3} > .... < a_{nd} > < a_{n\_out} >$,
where $< a_{ij} >$ is an ASCII floating point number corresponding to the j-th value of the i-th example and <ai_out> is the i-th desired output (in classification, it should be ± 1 or class number for multi-class problems).

Afterwards, training and testing run with different options for kernels and setting other parameters. The following kernels are available: linear kernel, polynomial kernel, Gaussian kernel, sigmoidal kernel as well as user defied kernels.

The training output is a model file against which the test files will be tested. In the case of n-class classification, the program will create n model files, each of them with a general file plus extension i, where i corresponds to the model for class i (+1) against all other classes (-1). The result of testing can be either a file with a class number for each sample from the testset, or an accuracy value with false positives and false negatives.

3.3.1.2 LIBSVM

Another SVM implementation is LIBSVM [Chang & Lin, 2001]. The main functionalities of SVMTorch are also present in the case of LIBSVM. Several kernels are provided: linear, polynomial,radial basis function, sigmoid, and pre-computed kernel (kernel values in training set file).Furthermore, there are more SVM types available: C-SVC, nu-SVC, one-class SVM, epsilon SVR and nu-SVR.

The tool can also be used as a command line application, from different environments, and a small GUI for demonstration is also available. Furthermore, LIBSVM provides a mechanism for scaling the data, for choosing the right parameters and for cross-validation. The same as with SVMTorch, a preprocessing step is needed to prepare the data in the following format:

<label> <index1>:<value1> <index2>:<value2> ... . .

In the case of classification, <label> is an integer indicating the class label. For regression, <label> is the target value which can be any real number. The pairs <index>:<value> correspond the the features. Multi-class classification is supported by means of the one against one approach.

What made LIBSVM particularly interesting for our project was the availability of libraries for different programming languages, including a library for Matlab which is the environment in which our real-time system was developed. Furthermore, the results can be outputted as probabilities of the sample being recognized as each class, which is more intuitive.

3.3.2 Gaussian Mixture Models

As opposed to SVMs which are used for supervised learning, GMMs are probabilistic models used for density estimation and can be regarded as unsupervised learning or clustering techniques. For the purpose of classification, the entire data from one class should be modelled with such a mixture of Gaussians.

Supposing there are K clusters parameterized by μ and Σ and that the data is denoted by X, where $X \in R^d$. The density of component k is:

$$f_k(x) = \phi(x|\mu_k, \Sigma_k).$$

The prior probability (weight) of component k is denoted by a_k. Therefore, the mixture density is:

$$f(x) = \sum_{k=1}^{K} a_k f_k(x) = \sum_{k=1}^{K} a_k \phi(X|\mu_k, \Sigma_k).$$

Modeling the data using mixture models with high likelihoods has the advantage that data within the clusters are tight (the distribution has high peaks) and the mixture is a good representation of the data that captures the dominant patterns.

The parameters of a GMM are the priors (or weights) of each Gaussian (a_k), the means (μ_k) and the covariances (Σ_k). The parameters are estimated using the maximum likelihood (ML) criterion. This is done using the expectation-maximization (EM) algorithm [Dempster *et al.*, 1977]. The EM algorithm can be briefly described by the following four steps, where the parameters at the r^{th} iteration are denoted by the superscript r:

1. Initialize parameters
2. Expectation step: Compute the posterior probabilities for all i = 1, ..., n, k = 1, ...,K.
$$p_{i,k} = \frac{a_k^{(r)} \phi(x_i|\mu_k^{(r)}, \Sigma_k^{(r)})}{\sum_{k=1}^{K} a_k^{(r)} \phi(x_i|\mu_k^{(r)}, \Sigma_k^{(r)})}$$
3. Maximization step:
$$a_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{i,k}}{n}$$
$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{i,k} x_i}{\sum_{i=1}^{n} p_{i,k}}$$
$$\Sigma_k^{(r+1)} = \frac{\sum_{i=1}^{n} p_{i,k} (x_i - \mu_k^{(r+1)})(x_i - \mu_k^{(r+1)})^t}{\sum_{i=1}^{n} p_{i,k}}$$
4. Repeat steps 2) and 3) until convergence.

A decision for new data is taken using a soft decision value, which is the log-likelihood ration of the data given the GMM of each class.

When using GMM, one important decision that needs to be made is the number of Gaussians. A rule of thumb says that we are supposed to use between 10 and 50 frames per parameter of the model. We denote the feature dimension D, the number of Gaussians G, the training time T and the step size S. The parameters of the model are the mean, the diagonal covariances and the weights. In total, there are $(2D + 1)G$ parameters. The number of frames is $Nf = T/S$. Therefore, the number of frames per parameter of the model is: $\alpha = \frac{T}{S(2D+1)G}$. Therefore, we can extract $G = \frac{T}{S(2d+1)\alpha}$ and we can vary α between 10 and 50 to have a guidance of the number of Gaussians that would be suitable.

For the purpose of our experiments we have used an implementation for training GMM and computing log-likelihoods provided by TNO Human Factors and implemented

by D. van Leeuwen. The modeling tool can also be used for training an Universal Background Model (UMB). This can be regarded as a GMM of the entire data from all classes. After having such a universal model, separate models for each class can be adapted. This results in faster training and performance comparable to the one of the GMMs obtained from the different classes. Tool for GMM adaptation from UBM are also provided by TNO.

3.3.3 Dot Scoring

Recent work [Brümmer, 2009] in language recognition focuses on improvements on the GMM approach. Instead of using the variable length feature vector that is usually inputted to GMM, some fixed size sufficient statistics are extracted and further used (zero and first order statistics). The analysis is made for every 10 ms of speech, after the pauses are eliminated. It is assumed that a different GMM is used to generate each speech segment. All these GMMs have the weights and precisions set equal to those of a UBM obtained from all the available data, and only the means are different. The method includes channel compensation, meaning that the impact of the communication medium is reduced. For a detailed overview of this approach please see [Brümmer, 2009].

3.3.4 Fusion and Calibration

Fusion of the results of more classifiers appears to be an effective method for increasing performance. As part of our experiments, we have performed fusion based on a linear combination of the t-normalized scores of our classifiers. For this purpose we have used equal weights for the classifiers.

Besides the fusion done using a linear combination of the t-normalized scores, linear linguistic regression fusion was performed.

Given that N classifiers are to be combined, and their output scores are s_{1j}, s_{2j}, .., s_{Nj} given a detection trial j, the fusion creates a score based on the following linear combination of them:

$$f_j = a_0 + a_1 s_{1j} + a_2 s_{2j} + \ + a_N s_{Nj}.$$

The constant a_0 is added to the formula in order to improve the calibration, and it has by itself no discriminating power. The weights a_1 .. a_N can be interpreted as the contributions of each classifier to the final score.

This approach is implemented using the *FoCal* tool developed by N. Brümmer. The tool provides simultaneous fusion and calibration. The fusion is linear, and a specific weight is assigned to each component in such a way to make the fusion optimal. For more details please refer to [Brümmer & du Preez, 2006].

What makes the results of the logistic regression training more powerful is that the fused score tends to be a well-calibrated detection log-likelihood-ratio.

Models for Emotion Recognition from Speech

4

In section 2.1.1 we have discussed psychological models underlying emotion recognition by humans. All these models differ from each other and there is not a unique model accepted by all psychological researchers. Nevertheless, in Figure 4.1 we present a model that combines most of the available models, and can be regarded as an average result.

4.1 Human Models

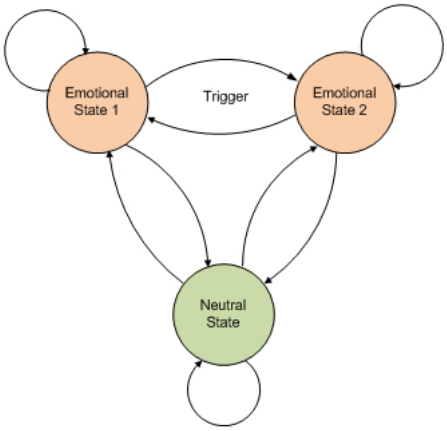


Figure 4.1: Emotion recognition model based on finite state machine

We can regard the human perception of emotion as a finite state machine, where external triggers determine transitions to new states or to the same state. This model is obviously based on the assumption of a discrete model of emotion. The triggers for emotion recognition from speech can be generated by the perception of the acoustic signal, by the perception of the spoken content, and most of the times a combination of both. A general model of emotion recognition by humans would of course include other types of triggers, like visual triggers from facial expressions and gestures, and also the perception of the context is of main importance.

Another view on emotion recognition coming from the tower of Hanoi idea. The different emotional states are regarded in a competitive manner. The triggers generated by speech (either acoustic or linguistic), contribute to one ore more states. In the case of this model we can imagine that discs are added to one or more rods, and at each moment, the emotion corresponding to the rod with the highest number of rods is winning, as shown in Figure 4.2. Furthermore, the model accounts for the fact that the triggers have

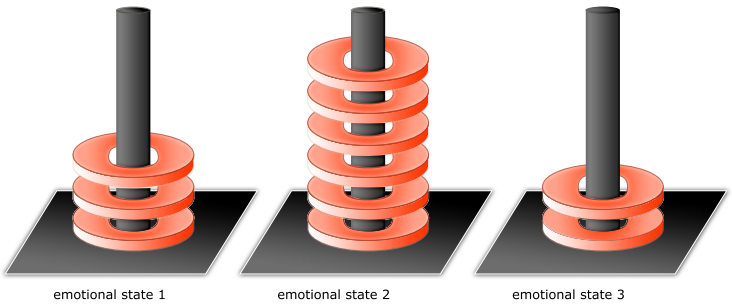


Figure 4.2: Model for emotion recognition based on the tower of Hanoi

a temporary effect, and when no discs are added and the old ones are removed time after time, the emotional state is less likely.

In the remainder of this chapter we propose three models for emotion recognition that could be used in an automatic application. We analyse the advantages and disadvantages of these models and based on them we come up with a decision for the model we will use in our research.

4.2 Model I

This model is based on the perception of words and we regard emotions as points projected in a two dimensional space. As we mentioned in 2.1.2.2, there is the availability of a dictionary proposed by Whissell: the Dictionary of Affect in Language (DAL) [Whissell, 1989]. Here, a large amount of words are included, and for each of them values for valence and arousal are provided. This means that by looking up the spoken words in this dictionary we can already have an insight in where the words are located in the two dimensional space of valence and arousal.

Furthermore, given a sequence of words, $\omega_1, \omega_2, ..., \omega_n$, we can come up with algorithms for finding at least some indications about the location on the 2-dimensional emotional space where the emotion can be located. As shown in Figure 4.3, one option would be to calculate the mean of such vectors that indicate valence and arousal locations in the 2D-space.

The main advantage of this method is that it is very fast and simple. Of course, the solution could be additionally improved by adding rules for dealing with negation, exclamations or other possible modifiers of the emotion [Fitrianie & Rothkrantz, 2006]. One important disadvantage is that people do not always express into words their emotions. Therefore, in most of the cases such a model would be unable to capture the emotional content of an utterance, unless the speaker gives clear indication of his or her state. It is true that DAL contains valence and arousal scores for many words, but intuitively most of these words gain an emotional meaning only from the context. Searching for emotional keywords is expected to give better performance, but it is bounded to strictly depend on somebody expressing his emotional state with words. Another disadvantage is that the entire focus is on words, and information from the acoustic signal is ignored.

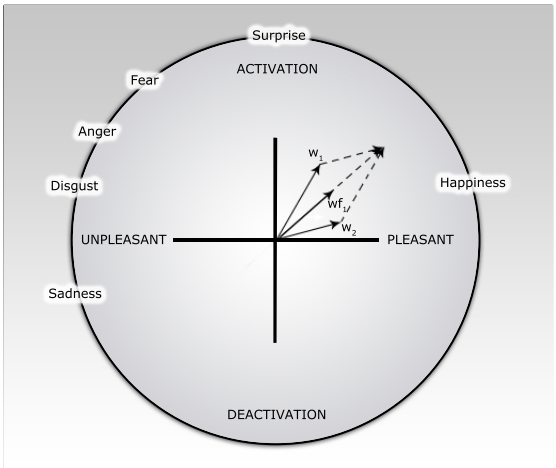


Figure 4.3: Model for emotion recognition (1)

In this case segments of speech uttered in an ironic way for instance, stand no chance as being recognized for what they really are. Furthermore, in the case of a real-time application, the success of this model is strongly influenced by the accuracy of a speech recognizer that would give the word transcriptions.

Research in natural language processing focuses on building good parsers and acquiring as much as possible information about what is being uttered, looking not only at words but also aiming to understand the semantics. Advanced NLP is beyond the scope of this thesis, but we do believe that a good implementation of an emotion recognizer based on spoken text will increase the ability to transmit the right emotion to machines. Recently a NLP tool was developed by [van Willigen, 2009].

4.3 Model II

The second model is based on the assumption that data labeled according to valence and arousal is available. In this case we are not looking at the words, but at the speech signal, or even both. Assigning valence and arousal scores for an utterance is a very difficult task, and probably almost impossible to do in an consistent way. This is of course when we are regarding the problem in a continuous space. The main reason for this is that our minds are not trained to give scores for such aspects as valence and arousal. Besides some set landmarks, mainly for the extreme and medium situations, people look at valence and arousal as ordinal scales. One proof is that when people are asked to perform such an annotation, as it is the case with the HUMAINE database (see section 3.1.4, they almost never use the same values. The just give a direction in their annotation, and a displacement based on how intense they perceive the modification. Basically this model can be used for automatic annotation, since the data is not sufficient for training.

One solution for using the data from such an annotation is looking at the gradient. This way we can perceive the direction and amplitude of the modifications. However,

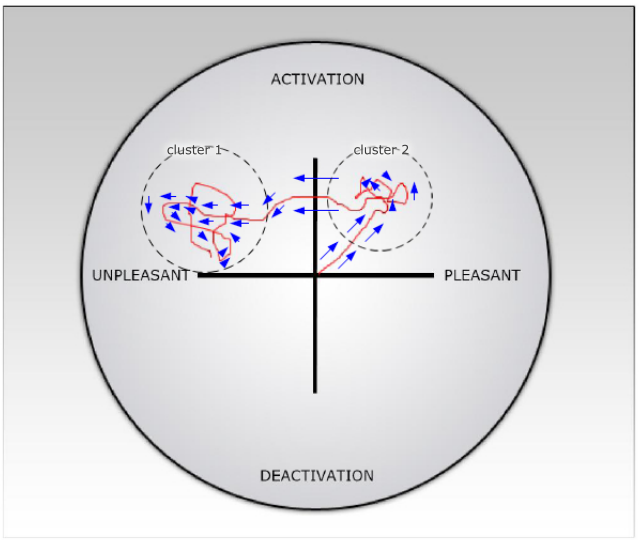


Figure 4.4: Model for emotion recognition (2)

this data is sufficient for learning about the direction of change, but the problem of amount of chance is still open. By having some knowledge about an initial position in the 2-dimensional space, and using the gradient, we can follow that direction and end up on the circle, in a point corresponding to a particular emotion (pure or blended). The disadvantage is that it is very difficult to decide, given an initial emotional state and this kind of annotation, whether to move to another cluster or keep staying in the same cluster.

A possible approach for using this model would be to use regression for learning the behaviour of the valence and arousal coordinates. This means that the result could be a set of directions which give insight in the evolution of emotions. The advantage of such a model is that it is expected to cope with data that otherwise could not be used because of the inconsistencies in labeling and the way in which people asses emotions in a continuous space. Also, the dynamic of emotion is an issue of high importance, which could lead to promising applications. For example, one might use such an emotion recognizer to observe the change in emotional state based on the presence of certain stimuli.

4.4 Model III

The third model is built based on the fact that we can choose areas in the valence and arousal space which we expect to correspond to emotions. This is actually the idea behind the discrete model of emotions. How to exactly define such areas is still an open question. From the point of view of valence and arousal, they might be separate, but from the point of view of their acoustic correlates and their evolution when emotional stimuli are triggered many of them overlap.

The performance of such a system, as depicted in Figure 4.5 is determined among others by the choice of clusters, which will be in the end the classes. By this we mean

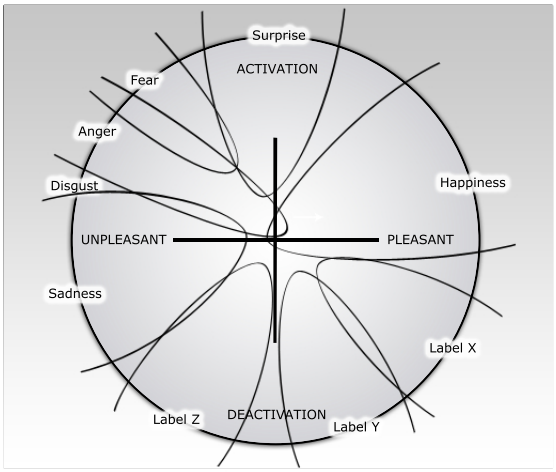


Figure 4.5: Model for emotion recognition (3)

both the areas covered by the classes, as well as the number of classes. These choices can be done differently according to each application. As visible from the figure, there are places in the 2-dimensional space that are uncovered by the classes. This means that only a closed set of cases are covered. Furthermore, it is very important how we delimit these classes. We might have the borders very close to the outer circle, and look in this way for pure emotions, or we can come closer to the origin and also allow the borders of the classes to intersect. In an ideal situation, the classes are disjoint. These provides the right setting for using classifiers and expecting them to learn from examples and later to recognize sample belonging to different classes. A disadvantage of this approach is that setting the borders between emotions is not trivial. Even in the case of humans, many times we are not able to detect precisely which emotion is expressed. This inclination to confusion has its roots in the fact that different emotions have similar characteristics (see section 2.3 and Table 2.1).

4.5 Discussion

Each of the three models presented in this section have their strong points, and of course their weaknesses. Probably the best results can be achieved by making a more complex model, based on both acoustic and linguistic information. Combining these channels of information in an efficient way is of course very challenging and need further research.

For the purpose of our thesis, we will use the third model. This approach is consistent with the databases chosen to experiment with, except the HUMAINE database. The setting created by these databases and the model is very suitable for the use of classifiers, since the databases are built based on emotion classes.

In the case of the HUMAINE database we believe the second model as well as the first model could be used. Since our research focuses on analysing emotion based on the acoustics of speech, we are inclined to use the second model. A possible model made

by combining the information gathered by analysing the database based on the first two models would probably improve the system.

5 Research Topics

This chapter is an overview of our experiments and the methodology that we have used. The experiments are supposed to highlight the beneficial actions for developing an emotion recognizer based on speech. As we already mentioned in section 2.5, the choices are straightforward. To begin with, we introduce the methodology followed for the experiments. Afterwards, each experiment is described in turn and observations are made.

The first experiment aims at testing the discrimination capabilities of the prosodic feature set described in section 3.2.1.1 using SVM as a classifier. The second experiment focuses on investigating the generalization capabilities of an emotion recognition system based on an extended corpus (a combination of more databases). After acquiring knowledge about these more general aspects, we present an experiment in which several feature sets and classifiers are used separately and afterwards fused, in order to increase the detection performance. In contrast to the studies mentioned so far which are based on the discrete model of emotion, meaning the recognizer has to choose between several classes, the fourth experiment deals with emotions regarded in a continuous space. Based on this we will try to predict levels of valence and activation for a speech sample for different moments in time.

The chapter ends with some conclusions about the relevant information which came out of the results of the experiments.

5.1 Methodology

This section describes the manner in which the classification experiments were developed. Figure 5.1 shows the action flow of the classification process.

The idea behind classification is learning by examples. This means that the classifier will be trained on some data, and that it is expected to apply the models it builds to new data. In our case, we use several databases of emotional speech.

5.1.1 Speaker Independence and Cross-Validation

The main requirement is that the database needs to be split into a training set and a evaluation set. The choice of training and test sets should not be made at chance because this will not give accurate results of the performance of the model. Our classification experiments were made in a *cross-validation* setting. This means that the entire data is split into n parts, a procedure called *jack-knifing*. In the case of n-fold cross-validation, each of the n samples will be used as evaluation data once, while the other $n-1$ samples constitute the training set. The procedure is repeated n times. The results of all the n

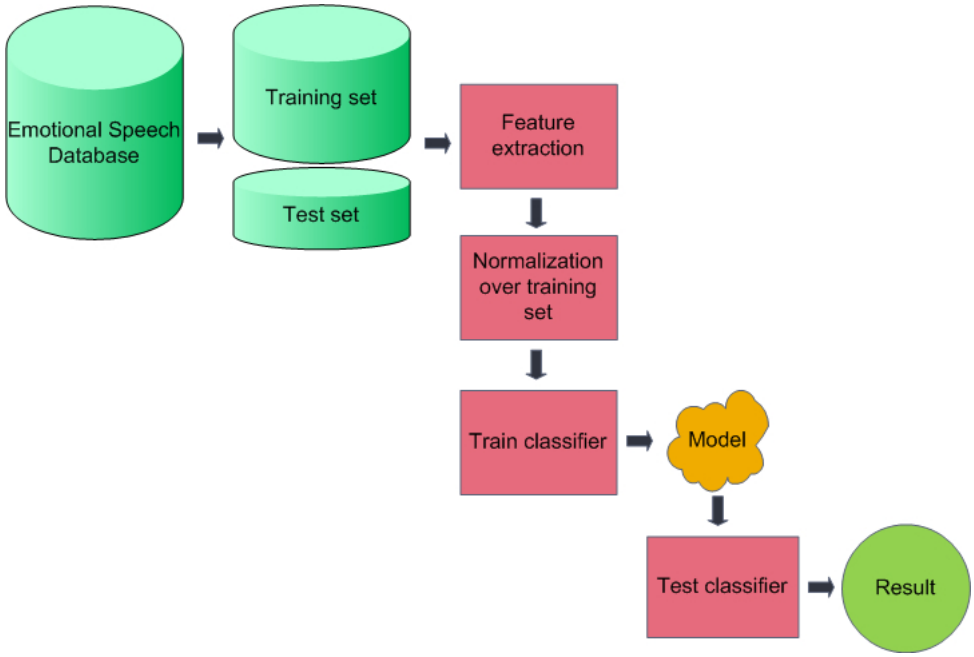


Figure 5.1: Classification flow

folds are then averaged, and the result can be considered a good approximation of the recognition performance on new data.

A version of n-fold cross-validation that we have used is *leave-one-out* (in our case leave-one-speaker-out) cross-validation (LOSO). This implies that for each fold the test set contains only samples from one speaker, and that the training set does not contain any sample from this speaker.

The LOSO-cross-validation is a good way for building *speaker-independent* models, which is an important feature of a robust emotion recognizer. Speaker independence can also be achieved in a n-fold cross-validation setting, by taking care that the samples in each fold do not contain data from speakers that are present in all the other folds. In the case of n-fold cross-validation speaker independence must be provided by splitting the data in jacks in such a way that speaker from one jack is not found in any other jack. The main heuristic used for creating speaker independent jacks is looking at the names of the files. In the case of most databases, the file name contains and indication of the speaker's identity.

In some cases like the fusion of more classifiers, a development set is necessary besides the training and evaluation sets. This procedure is called *double-cross-validation*. The extra cross-validation is used for estimating fusion and calibration parameters.

5.1.2 Preprocessing

After the training and test sets are prepared, a few processing stages are necessary. In the case of utterance level features, they are normalized over the training set, based on the

fair assumption that there is no information available of the test set. The normalization is done per feature type in order to achieve $\mu = 0$ and $\sigma = 1$, process also known as z-normalization.

In the case of double cross-validation, the normalization parameters were not computed over the train set but over the development set, in our case of the non-target samples. This process is a variant of t-norming. After the features are normalized, it is also necessary that the training and evaluation files are in the format specified by the classifier (see section 3.3.1).

5.1.3 Evaluation Measures

There are two problems we are dealing with within our experiments: detection and classification. Classification is a problem aiming to give a label to a sample based on its features and a model. For classification we will look at the results from two perspectives:

- classification accuracies, which means the percentage of samples which were correctly classified,
- confusion matrices, which show the types of mistakes between any pair of classes.

Detection aims at distinguishing one class from all the other ones; in short, by detection we look at a sample and we want to know if it is class X or not. It is a less difficult problem than classification, because we are only interested in detecting target samples from non-targets. For instance, given a call center application, we want to detect the angry customers. Therefore, we will consider the speech samples from angry customers our target, and all the other samples as non-target. Given that the output of a classifier can be seen as a score which is higher for targets and lower for non-targets, the distribution of such scores can be pictured as in Figure 5.2.

There are two types of error that we can make in detection:

- *false alarms*: classifying a sample as target when it is in fact a non-target, and
- *misses*: classifying a sample as non-target when it is in fact a target.

The well-known approach of determining the accuracy based on the number of errors made divided by the number can sometimes give misleading insight in the performance of a recognition system. This is especially the case when the amount of target and non-target data is unbalanced. For example, if the number of target samples is very small compared to the number of non-target samples, we can miss almost all targets and still have a high accuracy, which is definitely not a relevant performance measure. In this case it is important to look at both error types: misses and false alarms.

The separation between the two classes depicted in Figure 5.2 cannot be perfect because of the overlap between the classes. A decision threshold is used for determining a border between the classes. The threshold can be located in different manners: for example allowing the targets to be detected perfectly with the expense of some non-targets being detected as targets or the other way around. However, when there is a trade-off between more error types involved (as the decrease of one error type determined

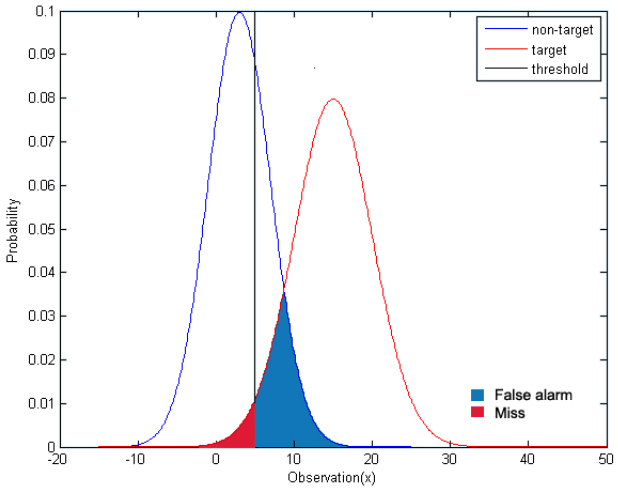


Figure 5.2: Distributions and error types for two classes

the increase of the other error type), the performance of a system is better described using a curve.

The DET (Detection Error Tradeoff) curve is a plot of the false alarm probability on the horizontal axis and miss probability on the vertical axis. It is a variant of the traditional ROC curve (Receiver Operating Characteristic or Relative Operating characteristic) [Egan, 1975] which presents false alarm probability on the horizontal axis and the correct detection rate on the vertical axis. In the case of the DET curve both error types are treated uniformly and the resulting plots are close to linear. The best performing systems lead to DET curves very close to the origin and ROC curves close to the higher left corner. The DET curve provides a better visualization of the systems contrast, for more details please refer to [Martin *et al.*, 1997].

A value that summarises the DET curve is called the equal error rate (EER). It is the value of the joint error rates given that they are equal.

DET curves and equal error rates are very good measures of the discrimination capability of a recognition system. However, for a real application where a decision needs to be made a specific threshold has to be set beforehand. The process of setting a threshold is called calibration.

The Detection Cost Function is a simultaneous measure of the discrimination capabilities and of calibration. It is based on computing a cost based on the actual costs of misses (C_{miss}) and false alarms (C_{fa}), as well as of the probability of a target (P_{tar}). Based on these parameters, the detection cost function is calculated as follows:

$$C_{det}(P_{miss}, P_{fa}) = C_{miss} P_{miss} P_{tar} + C_{fa} P_{fa} (1 - P_{tar}).$$

After calculating the detection function, the natural approach is to choose a threshold that minimizes the cost function. However, choosing such a threshold that is expected to have a proper behavior on unseen data is a challenging task. A new metric called C_{llr} encapsulates information about the application independent performance of a recognition

system. For a more detailed description please refer to [Brümmer & du Preez, 2006] or [van Leeuwen & Brümmer, 2007].

For the purpose of this thesis, the results will be interpreted in terms of accuracies and confusion matrices for classification, and DET curves with equal error rates for detection. Therefore, in most cases we are concerned with the detection performance and not so much with calibration. Only when using the fusion and calibration tool described in section 3.3.4 we will also look at the cost function and C_{llr}, since this tool is bound to give well calibrated log-likelihood scores.

5.2 Experiment 1 - Testing the General Features Set

5.2.1 Experiment Setup

This experiment focuses on determining the capabilities of the utterance level feature set presented in 3.2.1.1. Praat was used for feature extraction. The classification is done using LIBSVM (see section 3.3.1.2). The experiment is done on the following databases: Berlin, DES and ENT, for more details please refer to section 3.1.

As we decided to build a speaker independent model, we had to consider the number of speakers for each database. The Berlin database contains utterances from 10 speakers, DES from 4 speakers, and ENT from more than 40 speakers. Therefore, our option was to use a leave-one-speaker-out cross-validation approach. In order to cope with speaker differences, z-normalization of the features takes place (see section 5.1.2).

As mentioned in the previous section, the features were extracted using the program Praat. In a preliminary step to feature extraction, the voiced-unvoiced algorithm of Praat is used in order remove the unvoiced sound segments and a new sound object is created by the concatenation of the voiced parts. This is an easy to use method for removing irrelevant segments like silences. The new sound object was later used for feature extraction.

The feature set has every time the same length because a specified number of features are extracted from each utterance. For doing SVM classification, training and test files should be created and converted to the required format, for instance the format required by LIBSVM described in section 3.3.1.2. The normalization and conversion to LIBSVM format were achieved by means of Octave scripts. Data organization into folds and the the whole process of cross-validation was performed using Bash scripts.

SVM theory suggests the use of RBF kernel in the case of short feature sets, which is also our case. The results from the next section were obtained using LIBSVM with RBF kernel and the probability estimates mode.

5.2.2 Results and Interpretation

The results of this experiment are presented as both classification and detection results. In the case of detection, two types of preliminary results were used: the multi-class results outputted by LIBSVM using the 'one versus one' approach, and the results from a 'one versus the rest' approach for which we have split the data manually into target and

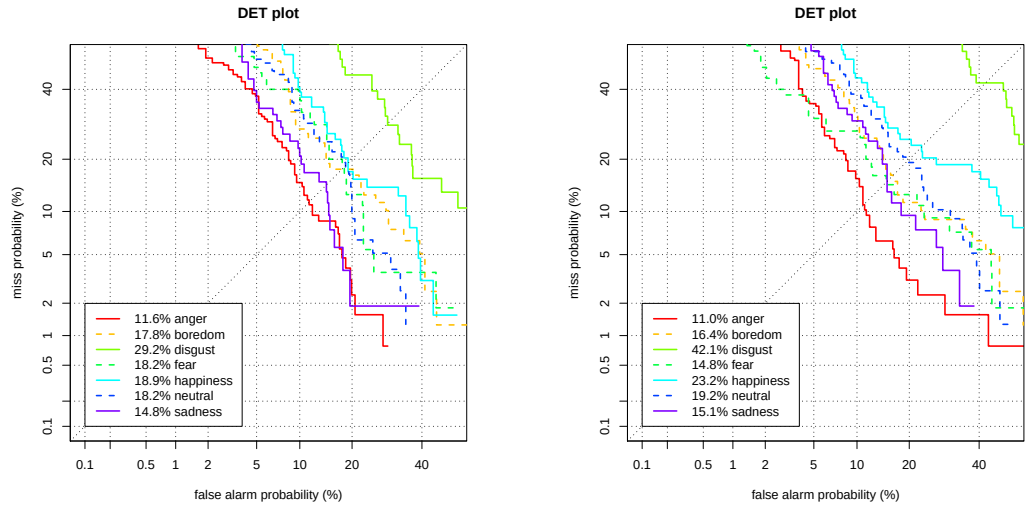
non-target classes. For detection, equal error rates are analysed as well as DET curves (see section 5.1.3). For classification the results of the confusion matrices are analysed.

The SVM parameters are the following:

- SVM type - C-SVC
- Kernel: RBF $exp(-gamma * |u - v|^2)$
- Cost 1
- Gamma 1/k where k is the number of attributes in the input data
- Using probability estimates
- Weight one.

Different kernels were tested but the results were never better than the one of the RBF kernel, which is also the one sustained by theory. A parameter optimization procedure provided by LIBSVM for tuning the parameters C and Gamma was performed, but again there were no improvements in the results.

5.2.2.1 Results on the Berlin Database



(a) One vs. one SVM

(b) One vs. all SVM

Figure 5.3: DET curves for Berlin database

Analyzing the DET curves and equal error rates for the Berlin database we can see that the best detected emotion is anger while for disgust the error rates are the highest. These findings confirm the recognition results of humans. Table 5.1 provides a comparison of the error rates of humans, SVM with one vs. one multi-class implementation and SVM with one vs. the rest approach.

Table 5.1: Equal error rates for Berlin database

Emotion	LIBSVM (1 vs. 1)	LIBSVM (1 vs. all)
Anger	11.6	11.0
Boredom	17.8	16.4
Disgust	29.2	42.1
Fear	18.2	14.8
Happy	18.9	23.2
Neutral	18.2	19.2
Sadness	14.8	15.1
Mean	18.38	20.25
Standard deviation	5.42	10.35

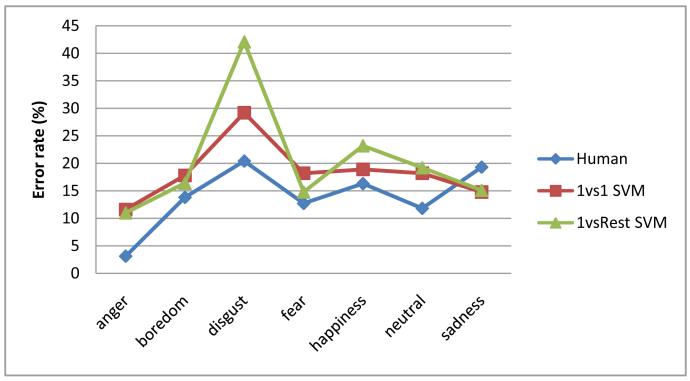


Figure 5.4: Error rates for humans and two types of multi-class SVM on Berlin database

The machine recognition comes close to the human recognition for boredom, fear and happiness. Sadness is the only emotion for which the automatic recognition outperforms human recognition.

Looking at the one versus one and one versus all approaches, we can observe that the average equal error rate is slightly lower for the former approach. Furthermore, analysing the standard deviation we can see more variability in the latter approach and more close-together results for the former.

The confusion matrices can give further insight into the results. The highest confusion seems to be between happiness and anger since almost half of the happiness samples were classified as anger. On the other hand anger is not that much confused with happiness. Sadness is often recognized as boredom, disgust as fear, boredom as neutral and neutral as boredom.

The strong confusion between happiness and anger does not come as a surprise since, as shown in Table (correlates of emotions), most pitch and intensity related features behave in the same manner for these two emotions. It is the same case with sadness and boredom, most features are strongly correlated. Therefore, a discrimination between these kinds of emotions with very similar acoustic features is difficult and the best would

Table 5.2: Confusion matrix for Berlin database

	Stimulus							
Response	fear	happy	anger	sadness	disgust	boredom	neutral	SUM
fear	**28**	2	3	4	13	4	3	57
happy	5	**26**	15	0	8	0	0	54
anger	4	29	**105**	0	3	5	1	147
sadness	4	0	1	**33**	2	10	7	57
disgust	8	7	3	0	**8**	2	6	34
bored	1	0	0	12	0	**40**	18	71
neutral	5	0	0	4	4	18	**43**	74
SUM	55	64	127	53	38	79	78	

be to employ another modality for extra information.

We can think of several reasons why disgust appears to be the most difficult to be detected. First of all, looking back at Figure 3.1, we can see that disgust is the emotion with the lowest number of samples. This means of course that the training set is smaller and the model is less accurate. Figure 5.5 contains two plots: the detection results for each emotion using 1 vs. 1 SVM, and the percentage of each emotion in the database. It is quite interesting to see that there appears to be a correlation between the two. Also, disgust was also the emotion with the lowest human recognition rate, so we might assume that there is indeed a difficulty in recognizing this emotion, based on the characteristics of the acoustic features of this emotion or the quality of the acted data.

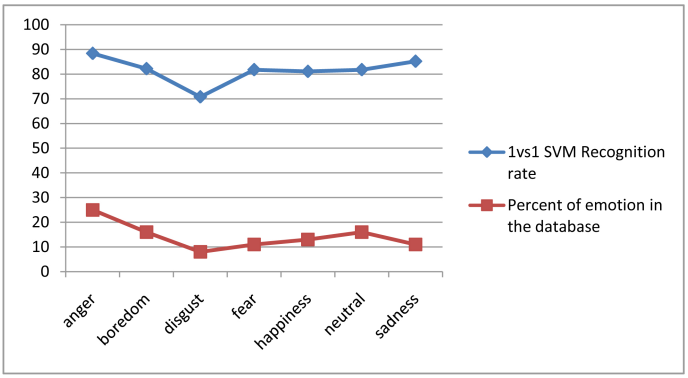
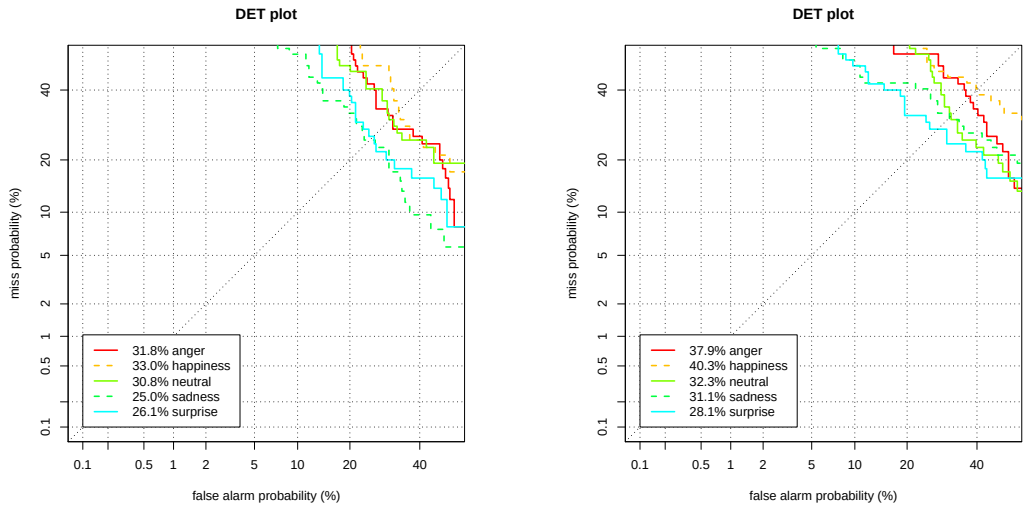


Figure 5.5: Recognition rates for the Berlin database using LIBSVM(1vs1) and percentage of samples for each emotion in the database

5.2.2.2 Results for the DES Database

As we can observe from Figure 5.6, the multi-class LIBSVM has the lowest error rate for sadness while the one vs. rest approach has the lowest error rate for surprise. However, the difference between the two is very small. Sadness is also the best recognized by



(a) One vs. one SVM

(b) One versus all SVM

Figure 5.6: DET curves for DES database

Table 5.3: Equal error rates for DES database

Emotion	**LIBSVM (1 vs. 1)**	**LIBSVM DB (1 vs. All)**
Anger	31.8	37.9
Happy	33.0	40.3
Neutral	30.8	32.3
Sadness	25.0	31.1
Surprise	26.1	28.1
Mean	29.34	33.94
Standard deviation	3.56	5.02

humans for this database. It seems to be a consistent conclusion that happiness is the most difficult to differentiate emotion from the outputs of our SVM classifiers as well as from the human recognition results.

The average results of the two types of multi-class classification are not very different. Again, as in the case of the Berlin database, we can observe more variability in the results obtained with the one versus all approach (see Table 5.3).

Figure 5.7 gives an insight into the performance of the two multi-class modalities and the performance of humans. There appears to be less correlation between the automatic classification results and the results for humans than in the case of Berlin database. For sadness and anger humans seem to achieve much better results than the machine classification. However, the latter prove to be more effective for emotions like happiness and surprise.

Further analysis can be made based on the confusion matrix obtained to LIBSVM multi-class classification presented in Table 5.4 and the one provided as documentation

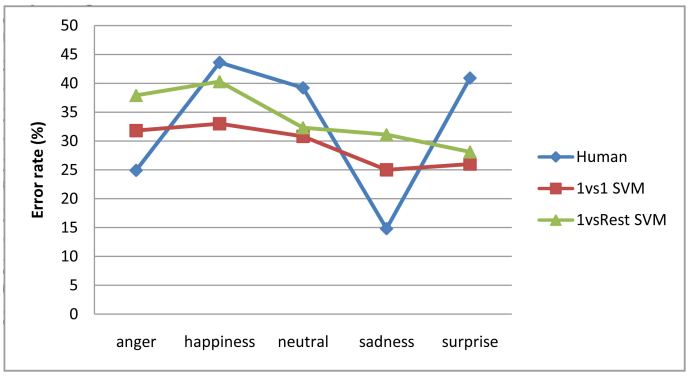


Figure 5.7: Error rates for humans and two types of multi-class SVM on DES

Table 5.4: Confusion matrix for DES database

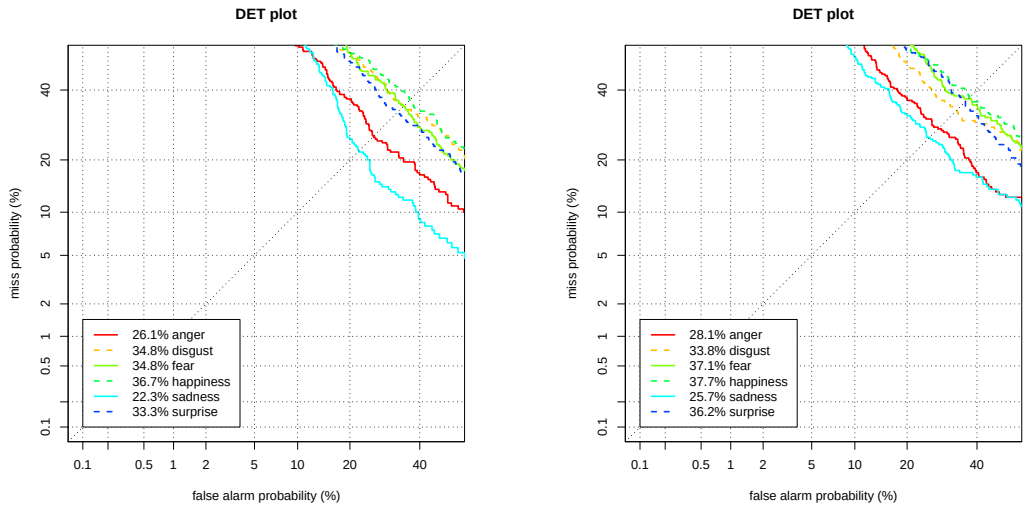
	Stimulus					
Response	anger	happy	neutral	sadness	surprise	SUM
anger	**21**	21	12	5	4	63
happy	17	**14**	1	1	14	47
neutral	6	1	**8**	13	4	32
sadness	1	1	23	**30**	3	58
surprise	5	15	8	3	**25**	56
SUM	50	52	52	52	50	

to the database for the human recognition performance (see Figure 3.5). As in the case of Berlin database, the greatest confusion was between happiness and anger for the automatic system. However, humans did not make many times this kind of mistake. Instead, they strongly confused happiness with surprise. In the case of humans anger was mostly mistaken for neutral while for SVM anger is many times recognized as happiness. For the other emotions, there are no differences between the most repetitive mistakes between human and machine classification: neutral is confused with sadness, surprise with happiness and sadness with neutral.

5.2.2.3 Results for the eNTERFACE'05 Database

Sadness wins the first place where SVM recognition is concerned, and it is followed by anger. The highest equal error rates are obtained for happiness. The results show that the best recognised emotion is sadness, while anger is the second best. The average error rates for the two modalities of multi-class classification are not very different, slightly better for the one versus one approach. For the first time in our experiments, the results show more variation for the one versus one approach.

Unfortunately results for human recognition on this database were not provided. From the confusion matrix shown in Table 5.6 we can analyse the most frequent confusions. Anger is mostly mistaken for happiness, disgust for anger and sadness, fear



(a) One vs. one SVM

(b) One vs. all SVM

Figure 5.8: DET curves for ENT database

Table 5.5: Equal error rates for ENT database

Emotion	**LIBSVM (1 vs. 1)**	**LIBSVM (1 vs. All)**
Anger	26.1	28.1
Disgust	34.8	33.8
Fear	34.8	37.1
Happy	36.7	37.7
Sadness	22.3	25.7
Surprise	33.3	36.2
Average	31.3	33.1
Standard deviation	5.75	5.04

for sadness, happiness for anger, sadness for fear and surprise for happiness. For this database it appears the strongest confusion was between fear and sadness. Anger and happiness for both mistaken for each other, but the model for anger seems more robust since the confusion is not that frequent.

From the analysed databases we can conclude that it is very difficult for a classifier on acoustic features to discriminate between happiness and anger, sadness and neutral, disgust and fear and happiness and surprise. The reasons for this confusion is that these couples of emotion share similar acoustic features and therefore a classifier trained on these features has sometimes problems in outputting the correct class. A solution for this problem can be finding extra features that are more relevant. However, this is not part of our research. Also, the combination between more modalities, for instance textual analysis or facial expression analysis and their fusion with such a system might improve the results.

Table 5.6: Confusion matrix for ENT database

	Stimulus						
Response	anger	disgust	fear	happy	sadness	surprise	SUM
anger	**114**	36	28	56	13	20	267
disgust	14	**60**	24	31	20	23	172
fear	16	21	**57**	20	32	27	173
happy	27	31	11	**66**	4	38	177
sadness	17	35	66	15	**124**	34	291
surprise	22	27	24	19	17	**68**	177
SUM	210	210	210	207	210	210	

The lowest equal error rates were obtained for anger and sadness and the highest were obtained for happiness. The acoustic correlates of these two emotions are quite opposite. It is interesting to note that even though happiness is frequently recognized as anger, the opposite does not happen that often.

5.2.3 Conclusion

The chosen feature set leads to lower performances on the tested databases than results reported in the literature where more extensive feature sets and selections of those have been used. This is to be expected since the feature set is not adapted to the databases, and of course finding a more specific set leads to better recognition rates. The chosen feature set does reflect the human manner for recognizing emotions since the same kinds of confusions are made and the same emotions are best recognized in most of the cases.

We do not claim that the chosen feature set is optimal nor that it is bound to give good performance on any kind of data, and we are aware that the feature vector lacks information that would be beneficial for the classifier. We believe that extending the feature set taking special care to the added features would increase the overall performance.

Regarding the SVM classification, we were content with the performance of the RBF kernel and the default parameters. Also, considering the two approached for multi-class classification, we noticed that in general the 'one vs. one' approach gives slightly better results. This approach is provided within the LIBSVM toolbox, so does not require extra work in contrast to the 'one vs. all' method which we had to implement. Research presented in [Hsu & Lin, 2002] argues that 'one vs. one' is a faster method for SVM multi-class classification. Due to the reasons mentioned hereby, the rest of the experiments use the 'one vs. one' approach for multi-class classification.

5.3 Experiment 2 - Investigating the Generalization Capabilities of Extended Corpora

Part of the work of this thesis aims at investigating ways for making speech emotion recognition more general. Such was the case with the previous section for finding a feature set suitable for most applications. The current section continues the strive for generality with an attempt to use extended corpora for training in the attempt to gain a more robust system.

5.3.1 Related Work

We have already found such contributions in the work of [Shami & Verhelst, 2007]. This study investigates the problem of multi-corpus training and testing. Four emotional speech corpora are used: Kismet, BabyEars, DES and Berlin. They were grouped in two pairs: Kismet and BabyEars, both containing speech directed to children, and DES and Berlin with speech addressed to grown-ups.

A segment based approach and a utterance based approach, each with different features, were used. The purpose was to investigate whether a classifier trained to recognize certain emotions in one database is able to recognize the same emotions in a different database and also to test the performance on merged corpora. Three kinds of experiments are presented:

- within corpus experiments (train and test databases are the same),
- off corpus experiments (train and test databases are distinct),
- integrated corpus experiments (training and testing is done on a fusion of corpora).

The results on multi-corpora experiments show that training on one database and testing on another one in general does not work. However, when similar classes from the two corpora were fused and the classifier was trained on the merged corpus, the recognition rate was higher than 70% for both groups.

Another experiment proved that by integrating the classes from two corpora, and keeping them as separate classes for training and testing, the classification performance is somewhere in between the performance for the separated corpora. One difference between the two groups of corpora is that for the Berlin-DES group the classifier never mistakes instances from one database as belonging to the same emotion in the other database (e.g. anger samples from Berlin were nor classified as anger Des and the other way around). For the Kismet-BabyEars pair there was a tendency for generalization, since there was sometimes confusion between instances of the two initial corpora. These results can also be attributed to the differences in language (Kismet and BabyEars were both American English while Berlin and Danish were German and Danish respectively).

A study of Klaus Scherer [Scherer, 2000] investigates the accuracy of human recognition in a multi-language emotion encoding-decoding experiment. Five basic emotions were used, and the experiment was done in nine countries, over 3 continents. The speech utterances were context free, containing phonological units from different Indo-European countries. The overall recognition rate for all emotions and all countries was 66%, which

proves that humans from different cultures use similar inference rules for recognizing emotions. Apparently the Germanic language speakers had the best results, followed by the Romanic language speakers. The worst recognition rate was for Indonesian speakers, the only country that does not belong to the Indo-European language family. The reason might be that listeners from some countries perceive features in addition to the basic perceptive features; therefore they are able to better recognize emotions. In addition, it can also be the case that the cognitive space of these listeners fits better the labelling.

The results of the multi-corporal machine learning experiment are very different from Sherer's multi-language study. Classifiers depend only on the labeled data while humans perform the tasks without being trained. The perceptive features used by humans allow generalization, which is not possible in the case of machine classifiers. Similar emotions from different databases can become recognizable by machine classifiers only if an ideal set of features is found.

The portability of emotion recognition systems has also been studied in [Vidrascu & Devillers, 2008]. They made use of three databases, two with call center data and one acted. The approach was to see the performance of a system trained on real data and tested on acted data. Their findings show that only emotions with very strong characteristics like hot anger can be recognized correctly, and that in general one emotion recognition system trained on real-life call center data can not recognize the same emotions from acted data. Experiments involving the two call center databases give more optimistic results, showing there are similarities between them. When using one corpus for training and the other one for testing, the accuracy is better in one direction than the other.

5.3.2 Experiment Setup

For the purpose of further investigation on the generalization capabilities offered by multi-corpora approaches we used a set of three database: Berlin, DES and ENT. In fact, subsets of this databases were extracted because we wanted to have samples belonging to the same emotional classes from all the databases. Only three emotions were common for the mentioned databases, namely anger, happiness and sadness.

In a second step an extra database was used, namely the South-African database. Considering the assumption that the 'English' class from this database corresponds to neutral and the 'Emotional' class corresponds to anger, we used this database as part of generalization experiments on two emotions. Therefore, for the three previously mentioned databases only samples from the classes anger and neutral were selected. Unfortunately the eNTERFACE'05 database does not contain samples portraying neutral states. However, the decision was not for discarding this database, but using the anger samples for training. This decision can also be regarded as a trial for compensation of the unbalanced data in the South-African database where there are twice as much neutral samples as emotional ones. Unlike the approach in [Vidrascu & Devillers, 2008], we plan to use acted data as training and test it on real-call center data, expecting to receive more reliable result for the capabilities of acted data.

A number of different tests are designed for the purpose of this experiment.

1. **Within corpus tests.** This stage involves building models for each corpus. The

Table 5.7: Accuracies in % for the within corpus experiments and 3 emotions: anger, happiness and sadness

	Accuracy	
Corpus	**10 fold SD cross-validation**	**LOSO cross-validation**
Berlin	84.01	80.81
DES	62.98	58.39
ENT	68.26	64.79

idea is to look at the performance of these models on the same corpus. Speaker dependent 10 fold cross-validation as well as leave-one-speaker-out cross-validation results are provided. A more detailed explanation of the latter protocol was given in section 5.1.1.

2. **Off corpus tests.** Models from one database are used for testing on another database. The normalization is done feature-wise, with mean and standard deviation extracted from the training set, because in a real application we would not have available normalization parameters for the new data. Also, a larger corpus build as a sum of more databases is used for training and the left out from training dataset is used for testing, again with normalization on the training set.

3. **Integrated corpus tests.** For this section, samples corresponding to the same emotion but belonging to different databases will be considered as one class. Afterwards, speaker dependent 10 fold cross-validation as well as leave-one-speaker-out(LOSO) cross-validation are performed.

5.3.3 Results and Interpretation

The result section is divided into two parts. First, three databases are used for the experiment: Berlin, DES and ENT. The purpose is to use three emotion classes which are the only classes present in all of the three datasets. The second part makes use of these three datasets plus the TNO database. The addition of a new database leaded to a restriction at the class level. For this second part the classes anger and neutral are used.

5.3.3.1 Anger, happiness and sadness

Before starting with the multi-corpora experiments, it is important to examine the within corpus results in the conditions of the experiment. For comparison reasons, the results are shown in speaker dependent and speaker independent modes, both with cross-validation, see Table 5.7. As expected, the speaker dependent results are higher. From a fast look we can already observe that the Berlin database leads to the best accuracies. The DES and ENT databases show accuracies somewhat closer-together, slightly higher for ENT. This can be a starting point with regard to the expectation we can have on testing on these databases.

Table 5.8: Accuracies in % for the off corpus experiments and 3 emotions: anger, happiness and sadness

Train	Test	Accuracy
DES	Berlin	46.31
ENT	Berlin	42
DES & ENT	Berlin	50
Berlin	DES	40.25
ENT	DES	59.74
Berlin & ENT	DES	48.05
Berlin	ENT	43.38
DES	ENT	58.05
Berlin & DES	ENT	49.6

For the off corpus experiment the literature already suggest that we should not have high expectation. Indeed, training on one database and testing on the other, as Table 5.8 indicates, does not give high performances. However, the results are better than random.

What we were particularly interested to see was whether adding more databases for the training set will yield an improvement to the result generated by each of the databases used for training by itself, given the same unseen database as test. From the test we can see that training on DES and ENT and testing on Berlin gives better accuracy than training on either DES or ENT individually. This is the only situation when we can see an improvement by using a larger training set. For DES and ENT as testsets, the accuracy for training on two datasets is somewhere in between the results from the individual training sets, a bit lower that their average.

An interesting remark is that the Berlin database was the only one who was better classified using an enlarged corpus, but also, is the database that used as a train set gives the worse results. From our experiments on one corpus we already know that Berlin gives better results than the other two. Compared to DES, it also gives better results for human recognition. This can be a reason to believe that Berlin is an"easier" database. However, what makes it easier to get results on, at the same time makes it less efficient in providing a good model for new data. All these three databases are acted, but perhaps the emotions in Berlin are more exaggerated or more obvious.

An argument against the importance of the training size can be derived from the following. The three used databases have different sizes. ENT is the largest one, followed by Berlin and than by DES. However, when ENT is used as unseen data, the results from the model generated using DES are much higher than the ones generated using Berlin (a difference of approx. 15%). The results of testing on DES and training on ENT are almost 20% higher than those for training on Berlin. We can assume that indeed Berlin is very bad for generalization purposes, or that DES and ENT are somewhat similar, for which we have no objective proof.

Looking at the accuracies presented in Table 5.9 we can see the classification performances on combined datasets. The results are better when Berlin is included, probably because this dataset is easier to be classified. From the two situation where Berlin is in-

Table 5.9: Accuracies in % for the integrated corpus experiments and 3 emotions: anger, happiness and sadness

Train	**10 fold SD cross-validation accuracy**
Berlin & DES	77.88
Berlin & ENT	72.78
DES & ENT	65.55
Berlin & DES & ENT	70.34

Table 5.10: Accuracies in % for the integrated corpus experiments and **2 emotions**: anger and neutral
and the accuracies for within corpus for comparison

Train	**Test**	**LOSO cross-validation accuracy**
Berlin	Berlin	80.81
Berlin & DES & ENT	Berlin	71.17
DES	DES	58.39
Berlin & DES & ENT	DES	62.96
ENT	ENT	64.79
Berlin & DES & ENT	ENT	63.08

volved in a couple, the combination with DES leads to better results. One of the reasons might be that, as DES is the least numerous data, the amount of "difficult" to classify samples is less for Berlin-DES pair, and larger for Berlin-ENT pair.

Using the three databases together leads to an accuracy of approximately 70% in speaker dependent mode. This result is higher than the individual results for DES and ENT in speaker dependent mode and less than the results for Berlin. In order to get more insight in what is actually going on when using more databases, we want to examine the results on each database independently, in a leave-one-speaker-out cross-validation system. This means that the result outputted for one database is an average of the results obtained for each speaker in that database with a training set consisting of all samples from the other database and the data from all speakers withing the current database except the one used for testing. The results are presented in Table 5.10.

Using this approach with testing on each of the three databases leads to an improvement for DES, an almost similar accuracy for ENT and a decrease in accuracy for Berlin. The higher recognition for DES can also be accounted on the small size of DES which would benefit from more training samples, and most probably the ones from ENT are helping. The lower result for Berlin show that the samples from the other two databases do not fit that well to the style of the database. In the case of ENT, the multi-corpora approach does not have a positive nor negative influence, which is actually a good sign.

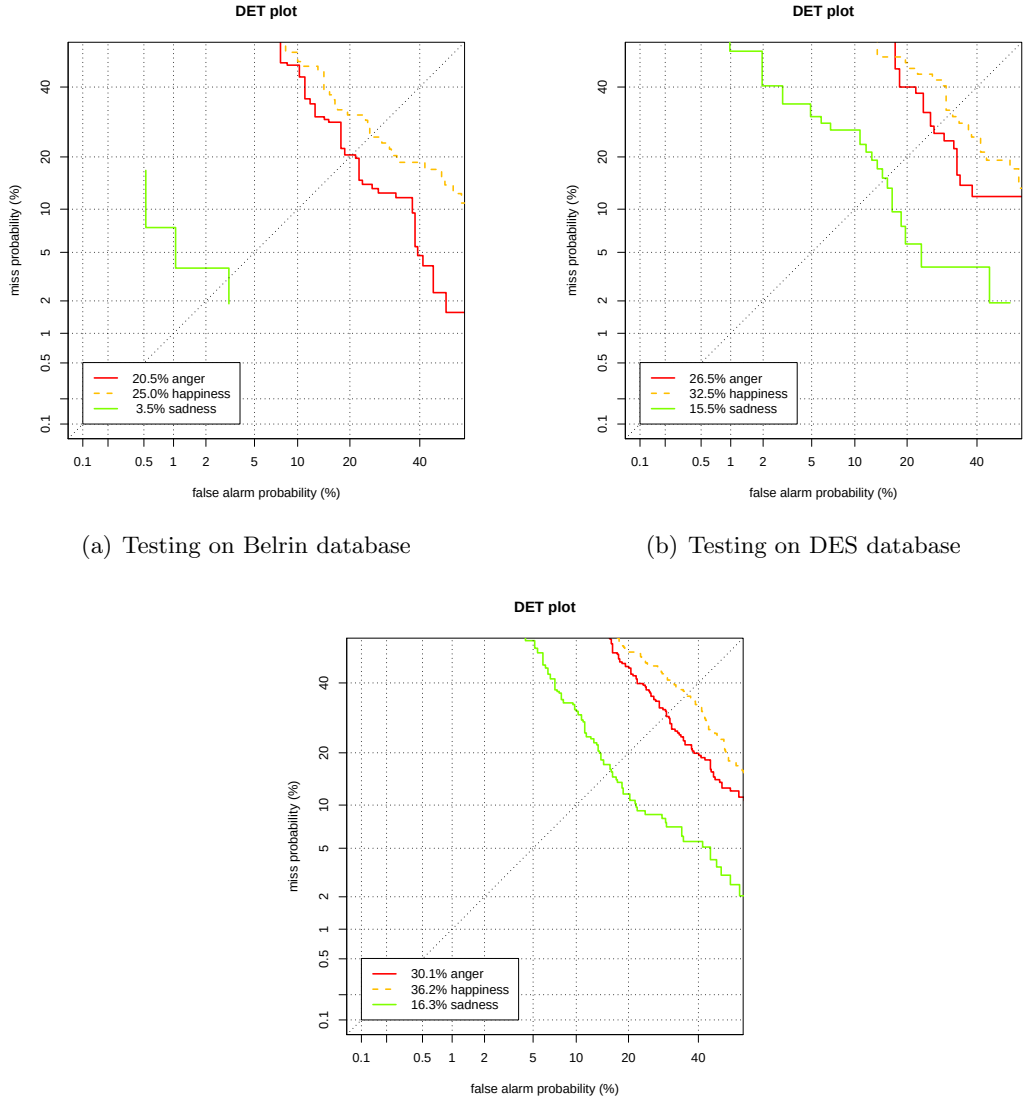
A more in depth analysis can be done based on the fluctuations of the equal error rates for different emotions in turn. The DET plot for Berlin database with training on all three candidates is plotted in Figure 5.9(a). The results should be compared with the ones from Figure 5.3(a). For a easier comparison, the equal error rates for each emotion are presented in Table tab:eerComparison.

Table 5.11: Equal error rates for 3 emotions on each database in comparison with multi-corpus results on each databse

Train	Test	Class	EER
Berlin	Berlin	anger	11.6
Berlin & DES & ENT			20.5
Berlin	Berlin	happiness	18.9
Berlin & DES & ENT			25.0
Berlin	Berlin	sadness	14.8
Berlin & DES & ENT			3.5
DES	DES	anger	31.8
Berlin & DES & ENT			26.5
DES	DES	happiness	33.0
Berlin & DES & ENT			32.5
DES	DES	sadness	25.0
Berlin & DES & ENT			15.5
ENT	ENT	anger	26.1
Berlin & DES & ENT			30.1
ENT	ENT	happiness	36.7
Berlin & DES & ENT			36.2
ENT	ENT	sadness	22.3
Berlin & DES & ENT			16.3

What we can easily notice from the comparison of the equal error rates is that sadness is the only emotion for which every time there is an improvement when a larger training corpus is involved. In the case of the Berlin database, sadness is the only emotion for which there is an improvement. For DES we can observe a small improvement in the case of anger, almost the same value for happiness and a almost 10% improvement for sadness. The case of ENT does not provide the same kind of direction in the evolution of the equal error rates. For sadness there is an improvement but not as strong as for the other databases and for happiness the improvement is almost insignificant. In the case of anger, the recognition is lower when more databases are used as a training set.

For these results it is quite difficult to draw a conclusion. Sadness is better recognized when more data is used for training. From what we have noticed so far from the previous experiments, we found out also that in most cases sadness is the easiest to be recognized. Anger was also an emotion for which small error rates can be achieved, but anger is many times confused with happiness. Happiness proved to be very difficult to be classified correctly in general, because it is many times confused with anger. This multi-corpora experiment however included only 3 emotions: anger, happiness and sadness. So the better results for sadness can be the caused by having more data for training but also to the fact that there is no real competition' for classes with the same features as sadness.



(a) Testing on Belrin database

(b) Testing on DES database

(c) Testing on ENT database

Figure 5.9: DET curves with training on Berlin, DES and ENT (LOSO)

5.3.3.2 Anger and neutral

The discrimination between anger and neutral is a very important problem with many applications in real life problems. Furthermore, this time the tests include a database with genuine recordings. The results obtained in a leave-one-speaker-out cross-validation procedure are presented in Table 5.12.

It is interesting to see that again for the Berlin database there is a decrease in performance. This time we are talking about just two classes which have quite different

Table 5.12: Accuracies in % for the integrated corpus experiments and 2 emotions: anger and neutral, and the accuracies for within corpus for comparison

Train	**Test**	**LOSO or 10 fold cross-validation accuracy**
Berlin	Berlin	98.4
Berlin & DES & ENT & TNO		90.4
DES	DES	70.5
Berlin & DES & ENT & TNO		77.62
ENT	ENT	-
Berlin & DES & ENT & TNO		92.56
TNO	TNO	83.72
Berlin & DES & ENT		63.77
Berlin & DES & ENT & TNO		83.59

features. Therefore we would expect good recognition results. In the case of Berlin, the results are worse even though much more data is provided for training. This is another clue that Berlin is a very specific database and does not perform well in combinations with other corpora. For DES, there is an increased performance when all corpora are used for training. DES has few samples and apparently it always benefits from addition of new corpora. In ENT only anger samples are present so we only provided the results for training on all databases.

The final step is testing on the South African database. As expected, training only on acted data and testing on real one does not perform very well but it is better than random. When including the SADB database in the training set, there is insignificant difference between using all corpora for training or just SADB. However, we cannot jump to the conclusion that the addition of new corpora does not have a negative effect. Especially in the case of this South African dataset we need to think about the sizes we are talking about. The three acted databases constitute an addition of approximately 400 samples for this experiment, which is far less than the amount of samples in SADB data - almost 3000. So in general using all that samples does not have a strong impact.

5.3.4 Conclusion

Designing an emotion recognition system that is general enough to provide reasonable recognition rate or new and different data would of course be a great achievement in the field of emotion recognition. Of course there is always a trade-off between building something geared towards a specific application, that would lead to very good performance in that specified domain, and building something general which can perform well on a variety of situation, but not as well as the specialized system.

In our tests we were interested to see how would the performance of a model change when more data is used. The results do not lead to a definite conclusion. What we are most interested in is how to make a system perform well on new data. Of course, intuitively a system trained on for instance three corpora should behave better than each system trained on only one of those corpora. This is unless that new data fits better

with the model provided by one corpus.

For this experiment we used the data which we had available, and that is not very much. Also, we used that same data in turn for testing. A more realistic experiment would contain more data for training, but also more data for testing, and one important condition is that the test data should be different. Otherwise, the specialised system will always work better.

Another issue that was not mentioned before is the use of recordings in different languages: German, Danish, English, South African languages. The prosody features contain a lot of language specific information, and since this information is encapsulated in the model, it can be seen as an impediment for generalization. Further investigations using more corpora of the same language might have better results. However, many signs of emotions are inter-cultural and it is also an important feature of a system to be able to deal with a change in language.

5.4 Experiment 3 - Reaching Higher Detection Performance on the South-African Database

This section describes a series of experiments which we made on the South-African database. The experiments include using different classifiers with different features and also some methods for fusing more classifiers at the decision level. The purpose is to see how well the classifiers combine, how well the features complement each other, and of course which method yields the best results.

5.4.1 Background

Gaussian Mixture Models represent weighted averages of a number of Gaussian probability distribution functions. They have been successfully used for different kinds of applications involving speech, like emotion recognition, speaker recognition or language recognition. GMMs are very good when it comes to modeling data and they do not have any constraints regarding the lengths of the feature vector. Therefore, they are a very good choice when using spectral features.

Previous work [Truong & van Leeuwen, 2007] shows that the fusion between classifiers trained on spectral features and classifiers trained on utterance level acoustic features is beneficial. These two feature types complement each other. The paper also uses GMM and SVM approaches, and their fusion leads to better performance.

5.4.2 Experiment Setup

The experiments described within this section made use of three kinds of features: spectral features descrived in section 3.2.2.1 (RPLP) extracted at the frame level, the feature set described in 3.2.1.1, and zero and first order statistics. For all these feature types different classification methods are used. As a final step late fusion between different classifiers is performed. All experiments are implemented using 10 fold speaker independent cross-validation.

5.4.2.1 GMMs with RPLP Features

As stated in section 3.1.5 there are two classes in the South African database: emotional and plain English. A first classification approach was to model each class by a mixture of Gaussians. The classification takes place by means of four iterations of the Expectation Maximization algorithm. The experiments include using different number of Gaussian mixtures in order to inspect which one is the most beneficial and what is the influence of varying the number of mixtures on the discrimination result.

In the case of RPLP features the number of features is high: 26 features for one frame, therefore a large number of features for an entire file. Because of this, in order to obtain a good representation, the data needs to be modelled with an adequate number of mixture. Experiments with 64, 128 and 256 GMM were made.

A precomputed Universal Background Model (UBM) was also used as a basis for GMM adaptation. The UBM was trained on English telephone conversations, so the recording conditions of these samples and the ones in the database we use are similar.

The UBM contains 512 mixtures. In general, adapting a GMM from a UBM trained on similar data results in a very fast training and the results are comparable.

5.4.2.2 GMMs with Praat Features

The feature set described in section 3.2.1.1 is also used for modeling GMMs. The feature preparation procedure is similar to the one previously described. After the features are extracted for each sample in the database, they are normalized by subtracting the mean of the feature elements and dividing the result by the standard deviation of the feature column, on the training set. Because some files were very short, Praat was unable to extract features from them, and those files were discarded. Of course, special care was taken to fulfil the speaker-independence constrain, and the ten folds used for cross-validation were the same as with the RPLP features, except for the files which were not processed.

Since for one utterance only 20 features are extracted, less Gaussians are necessary for modeling the data. Therefore, models with 2, 4 and 8 mixtures were built and tested by means of expectation-maximization algorithm.

5.4.2.3 The UBM-GMM-SVM Approach

The UBM-GMM-SVM approach is based on the idea of modeling the data by means of GMMs that are adapted from one UBM. GMMs are characterized by the means, covariances , as well as the weights of each Gaussian. From each sample in the training and in the test data, a GMM is adapted from the UBM and the means of the GMM are kept as feature files. In this approach, the idea is to use these means as training data for SVM and also for testing. The approach is very similar to the previously described SVM approach, given that a SVM does the actual classification. The difference is that instead of using acoustic features, the features are now the means of the Gaussian mixtures. The training files for this approach are very large, so one disadvantage of the approach is that it is very expensive.

5.4.2.4 The Dot-Scoring Approach

The dot-scoring technique is described in section 3.3.3. For using it, the first step is to prepare the data in such a way that the zero and first order statistics can be extracted. The tool introduced in section 3.3.3 was used for implementation and obtaining the results.

5.4.2.5 SVM with Praat Features

The features mentioned in section 3.2.1.1 are used for this experiment, along with SVM with RBF kernel. The experiment is designed in the speaker independent cross-validation framework. The features are z-normalized according to section 5.1.2.

5.4.2.6 Fusion

All fusion types applied in this work are at the score level. This means that there is no fusion taking place before each classifier gives an output, and the features from different approaches are not combined. The methods previously described lead to scores in different ranges. The aim is to use a linear combination of the scores in order to find a final result. In our case the scores from different classifiers are considered equally important, therefore we will consider their sum as the linear combination.

The scores come out from different classifiers, therefore they span different ranges (some are log likelihoods, some probabilities). As we want to perform a linear combination of the scores, a first step is to bring all the scores into the same range. This can be done using an adapted t-normalization [Auckenthaler *et al.*, 2000].

The following t-norming procedure was used for normalization (see section 5.1.2). The data was divided in training, development and test sets, for each possibility in the cross-validation framework. For an easy understanding, we name the Emotional samples as target, and the English samples as non-target. The mean and standard deviation of the scores of the non-target development set are used in order to normalize both the target and the non-target scores of the evaluation set.

For a successful fusion, it is important that the approaches being fused handle the problems differently, so that they would benefit from information and complement each other. In our case, the approached previously described can be roughly divided into GMM-based approached and SVM-based approaches, from the classification point of view, and in frame level features and utterance level features from the features point of view.

Given these different approaches, we can expect to benefit from the fusion between GMM-based and SVM-based approaches, since the classification procedure is very different, as well as from the fusion between different types of features.

Three examples of late fusion are implemented within this work:

- fusion between GMM with RPLP features and SVM with Praat features,
- fusion between the UBM-GMM-SVM approach and SVM with Praat features,
- fusion between the dot-scoring approach and SVM with Praat features.

Besides the fusion done using a linear combination of the t-normalized scores, linear logistic regression fusion was performed. This approached was implemented using the FoCal tool described in section 3.3.4. The tool provides simultaneous fusion and calibration. The fusion is linear, and a specific weight is assigned to each component in such a way to make the fusion optimal. The results of the logistic regression tend to be well-calibrated detection log-likelihood-ratios.

5.4.3 Results and Interpretation

Figure 5.10 depicts the equal error rates for GMM classification with different numbers of mixtures and RPLP features. It is important to note that for 512 Gaussians, the GMM was adapted from a UBM of different data. Apart from this result, be can see that the

trend is to have smaller equal error rates as the number of Gaussians in the mixture increases. Of course there is always a compromise that needs to be made, because when the number of Gaussians is high, the process become more expensive. In our case, 256 Gaussian lead to the smallest equal error rate (19.5).

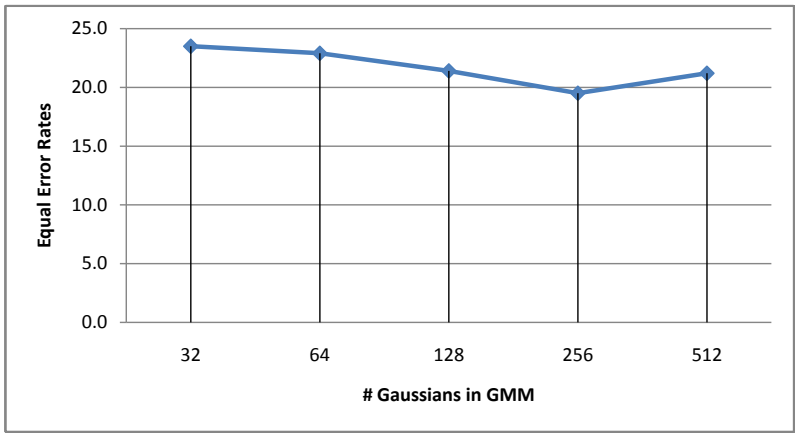


Figure 5.10: Equal error rates for different number of mixtures and RPLP features

Looking at the outputs for GMM with the Praat features depicted in Figure 5.11, we can see that the errors seem to be minimum for a mixture with 4 Gaussians, and that changing the number of Gaussians in any direction leads to worse performance.

Keeping the classifier fixed to GMM and using different types of features shows that RPLP features are more appropriate for GMM and lead to better results.

Figure 5.12 shows the DET plots for five methods: a 512 Gaussians adapted GMM with RPLP features, SVM with Praat features, and their fusion made in different modalities. The DET curve called GMM_512+PraatSVM(TN) is the result of the addition of the t-normalized scores from GMM and SVM. Apparently, but normalizing the scores and using a simple linear combination of them in which they are assigned equal weights leads to a high improvement in the performance.

Looking also at the previous figure, we can see that the Praat features lead to better results with SVM than with GMM. Also, Praat and SVM yield smaller equal error rates than RPLP and GMM. Their fusion leads to better performance than each of the individual approaches, which shows that the features and the modelling techniques are able to complement each other.

The dashed green curve represents the DET plot of the fusion between SVM using Praat features and GMM using RPLP by means of logistic regression (see section 3.3.4 using FoCal. In the case of both types of scores initial scores no normalization is done. The results are very simmilar with the performance of the SVM classifier by itself.

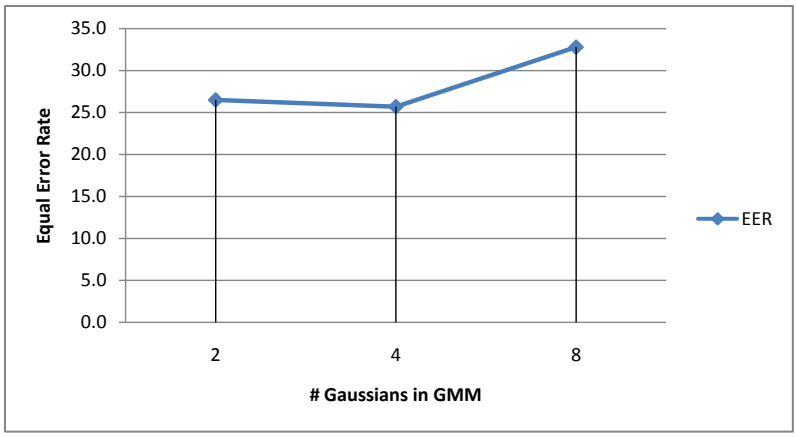


Figure 5.11: Equal error rates for GMM with Praat features and different number of mixtures

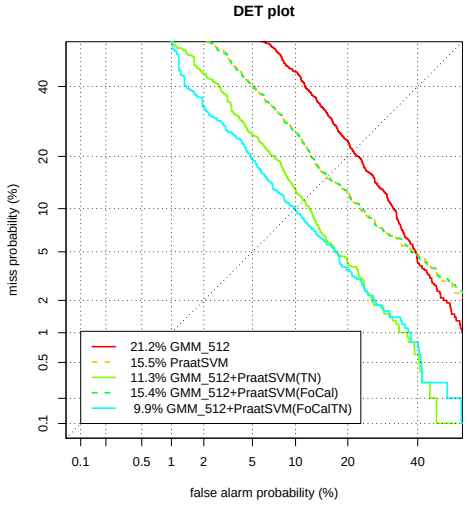


Figure 5.12: DET curves for GMM with RPLP, SVM with Praat features and their fusion

When the scores from GMM are t-normalized as explained in section 5.1.2 and they are fused with the scores from SVM using FoCal, the equal error rates drop to 9.9%, which is an improvement to all previously investigated methods. It is interesting to examine the coefficients which FoCal assigns to each of the classifiers, presented in Table 5.13. We can observe that the influence of SVM is stronger and that it is considered

Table 5.13: Fusion coefficients for GMM and Praat SVM fusion

Classifier	Weight
GMM	3.03
SVM	5.77

more important.

The UBM-GMM-SVM approach gives by itself better results than the regular GMM approach as we can see from Figure 5.13. This proves that the means of the Gaussians, which are used as input for an SVM classifier are a very meaningful representation of the data. For some areas in the plot, this approach performs better than SVM using Praat features, but mostly it is still inferior.

Again, these two approached are fused. The first type of fusion is the addition of their t-normalized scores. This approach works very well, since we can notice a decrease of equal error rates to almost 10%, and we are starting with two classifiers one with almost 20% and one with almost 15% equal error rate.

Using the fusion and calibration provided by FoCal which assigns different weights to the outputs of the classifiers, the results are further improved. The coefficients are presented in Table 5.14. We can see that the two classifiers are given almost the same importance. Interestingly, the weight of UBM-GMM-SVM is slightly higher, while its performance is slightly worse than the one of SVM.

There is not a big difference between the performances of two types of fusion in this case, logistic regression leading to only slightly better improvement (from 10.5 to 10.1).

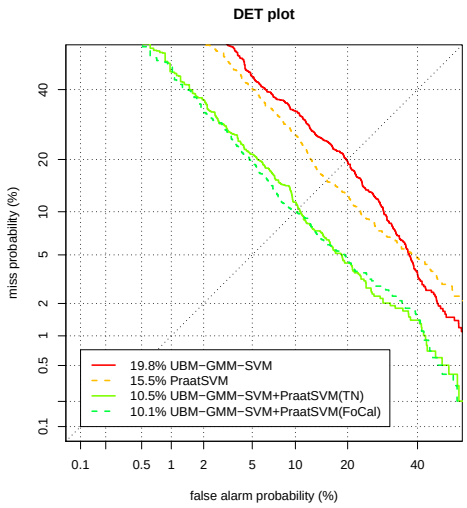


Figure 5.13: DET curves for the UBM-GMM-SVM approach, SVM with Praat features and their fusion

The DET curve for the dot-scoring approach is plotted in Figure 5.14 along with the curves for Praat SVM for comparison and their fusions. DotScoring leads to results very similar to the UBM-GMM-SVM approach. Interestingly, the fusion between dot-scoring

Table 5.14: Fusion coefficients for UBM-GMM-SVM and Praat SVM fusion

Classifier	Weight
UBM-GMM-SVM	5.62
SVM	5.44

Table 5.15: Fusion coefficients for UBM-GMM-SVM and Praat SVM fusion

Classifier	Weight
Dot scoring	1.81
SVM	5.27

and Praat-SVM based on t-norming leads to the same equal error rate as RPLP-GMM fused with Praat-SVM. Fusion by FoCal has also been performed, both on the initial and t-normed scores of Dot scoring. The results of all these types of fusion are very similar and in many points their DET curves are almost overlapping. The lowest equal error rate is achieved using FoCal and using the t-normed scores did not make a difference. The coefficients assigned by FoCal to SVM and Dot scoring are presented in Table 5.15. As in the case of using GMM, SVM appears to have a very strong impact on the final score.

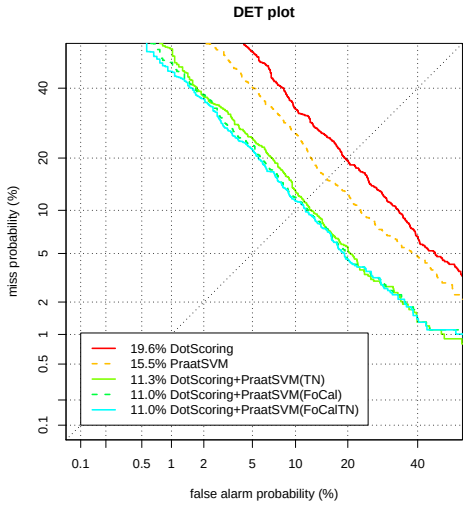


Figure 5.14: DET curves for the Dot-scoring approach, SVM with Praat features and their fusion

As a final step, we have fused all the individual classifiers using logistic regression. This was done on both the original scores and on the t-normed scores for GMM and Dot scoring. The DET curves of these fusions are presented in Figure 5.15. The fusion of all four individual classifiers: GMM, SVM, UBM-GMM-SVM and Dot scoring lead to an equal error rate of 7.1% if FoCal is used on the original scores. When the scores are t-normalized and then fused the equal error rate drops to 4.2% which is a major improvement. The weights used for fusion are displayed in Table 5.16. Here UBM-

Table 5.16: Fusion coefficients for the fusion of all individual classifiers

Classifier	**Weight**
GMM	3.35
UBM-GMM-SVM	5.92
Dot scoring	1.72
SVM	5.13

Table 5.17: Cost values for the fused systems

Method	C_{llr}	min C_{llr}	EER	C_{det}	min C_{det}
FoCal	0.2613	0.2505	7.05	0.0695	0.0684
FoCal TN	0.1624	0.1533	4.23	0.0416	0.0411

GMM-SVM seems to be the most important for the final result, followed closely by SVM. Dot scoring has the least importance.

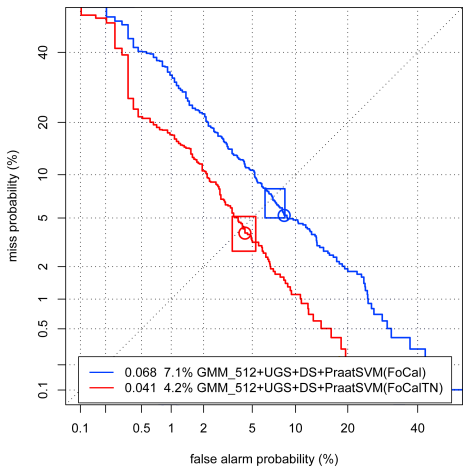


Figure 5.15: DET curves for the fusion of Dot-scoring approach, SVM with Praat features, adapted GMM and UBM-GMM-SVM

For the first time in Figure 5.15 we see two circles; they represent the minimum detection costs. Table 5.17 gives more insight into the cost associated with making a decision for the system based on the fusion of all individual classifiers, with and without t-norming. We can see that for these well calibrated systems the actual detection costs (C_{det}) are very close to the minimum costs ($minC_{det}$). This is the same in the case of C_{llr}, the actual cost is very close to the minimum cost. This means that by the chosen threshold we do not loose much performance.

We refer to the area on the DET plot corresponding to a threshold as an *operating point*. In Figure 5.15 boxes are drown around the operating point based on the 95% confidence intervals of P_{fa} and P_{miss} (recall section 5.1.3). We can see that in both cases the opperating point is very close to the point of minimum cost, which proves that the calibration is right. The situation in Figure 5.15 for the red curve is optimal, since

Table 5.18: Equal error rates of individual classifiers and their fusion

Classifier	**EER**
GMM	21.2
Dot scoring	19.6
UBM-GMM-SVM	19.8
SVM	15.5
FoCal fusion TN	**4.2**

the minimum cost is in the operating point confidence area.

5.4.4 Conclusion

In this section more classification techniques were experimented, and more feature types. We can regard them however as being part of two major classes: SVM and GMM. This is because the UBM-GMM-SVM approach as well as Dot scoring are also based on Gaussian Mixture Modelling. To recall the results of the individual classifiers, please see Table 5.18. All GMM-based approaches have equal error rates close to 20%. SVM performs better showing an equal error rate of 15.5%. Therefore, we start with a set of relatively weak classifiers, but by fusing them we can build a much stronger one.

Also, by using special tools for fusion and calibration we can obtain scores that can be interpreted as well calibrated log-likelihoods. This means that using proper information about the aplication parameter (C_{fa}, C_miss and P_{tar}), we can set a threshold at θ where θ is defined as:

$$\theta = -log\frac{C_{miss}P_{tar}}{C_{fa}(1-P_{tar})},$$

and expect this to be an optimum threshold in terms of cost. The point we operate in is located in the DET curve at the (P_{miss},P_{fa}) point given the threshold.

5.5 Experiment 4 - Analysis on the HUMAINE Database Based on the Continuous Model of Emotion

In many real-life application it might be less important to distinguish the specific emotion of a user based on making a choice from a closed set, which is the case of the previously discribed experiments. This closed set, based on the discrete model of emotion, might not be covering the essential information for a certain application. Sometimes it might just be important to see wether the experienced emotion is positive or negative, and the level of arrousal.

The only database we found which had a valence and arousal annotation available is the HUMAINE database, introduced in section 3.1.4. As part of this experiment, given a sound input, we try to predict values for valence and activation for each frame. This procedure is totally different than the previous approaches. We no longer have classes, we just have real values, and based on these real values (the labels) of a training set, we would like to build a model of the data that is able to predict appropriate real values for new speech samples. This is obviously a more difficult problem than classification.

Before presenting the experiement, it is important to provide a clear understanding on how we would like things to work and what are the impediments. As a general idea, having a database with time aligned annotation on a continuous scale, we can extract some features and use these features and the labels as training data for Support Vector Regression (SVR). However, an important requirement is that the data should be labeled on a absolute scale. The HUMAINE database provides a time-aligned labeling on the valence and activation dimensions of emotion. The method used for labeling is of main importance, and we provide a description in section 3.1.4. The annotation was done using Trace programs, and the users were supposed to move the mouse according to the value of the emotion they perceived. They could use the mouse within a given interval which included some indicators for extreme and intermediate states. In Figure 5.16 and 5.17 we can see a picture taken from the program used for annotation for valence and activation.

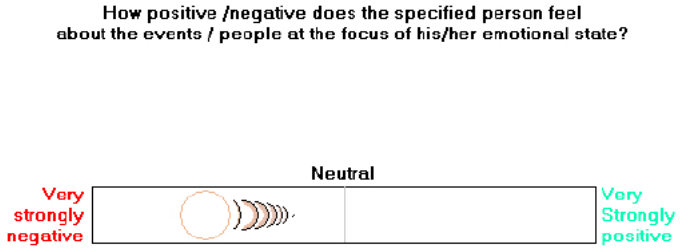


Figure 5.16: Trace program for labeling valence

In this case we believe Model II from section sec:model2 was used for annotation. This means that it is easy to understand the dynamic of emotion on the valence and arousal scale with regard to direction, but it is not so clear how big are the steps. Newertheless, we try to see if it is possible to get any positive result by using the labels as they are, and to predict values on the valence and activation scale.

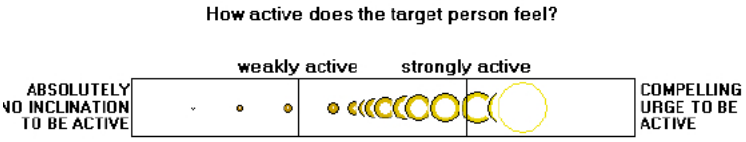


Figure 5.17: Trace program for labeling activation

5.5.1 Experiment Setup

In the case of the HUMAINE database, six expert labelers were employed to annotate the data. Their work resulted in trace-like files for each clip. Since no final annotation was provided, we have interpolated and averaged these annotations, to result in a general one. For more information on the procedure and the consistency between labelers, please refer to section 3.1.4.

As features for this experiment we have used MFCCs (see section 3.2.2.2). Every 10 ms, 39 coefficients are extracted over a 25 ms window, and are normalized per file. They are used as features for support vector regression using LIBSVM (see section 3.3.1.2). Separately, values for pitch at intensity were added to the MFCC feature vectors, to see if there is noticeable improvement. The pitch and intensity values were z-normalized on each utterance. If for a specific frame there was an undefined pitch or intensity value, we replaced it with 0. For normalization only the defined values were used.

Since our analysis is done on very small units of analysis, and in general a label is assigned to set of MFCC-based features, it would not make any sense to give labels to silences because that would really confuse the system. Besides the labeling for different dimensions of emotion, for the HUMAINE database transcriptions of words on a time line are also available. We have used this transcription in order to determine the pauses in speech. In the word annotation files, the precise time of beginning and ending of each word is given (resulting in a sequence of borders between word for continuous speech), and also the beginning and ending of pauses can be extracted. Therefore, we have exluded from our train and test data all the the silences. From the total of 48 clips in the HUMAINE database we have selected only the files for which word transcriptions were provided, resulting in a subset of 37 clips.

For training SVR we used the MATLAB library of LIBSVM. All reported results were performed in a 10-fold cross-validation manner. We have performed experiments using different units of analysis. Besides regression, we have devided the data in extreme classes for each dimension: active versus passive and activated versus deactivated.

5.5.2 Results and Interpretation

The results for regression are expressed by means of squared correlation coefficients. We have performed SVR for MFCCs averaged over different time intervals, ranging from 0.1 to 1.0 seconds. The labelss were also averaged over these interval. The best result are obtained for the shortest window length (0.1 miliseconds), in the case of both dimensions. The tests were run first only using the MFCC features, and for a second trial the pitch and intensity features were added to the feature vector.

Table 5.19: Accuracy for SVM classification for valence and arousal

Classes	Accuracy (%)
Positive vs. negative	66
Active vs. passive	75

Table 5.20: Inter-rater correlation for HUMAINE database

Dimension	Correlation Coefficient
Activation	0.33
Valence	0.40

The highest squared correlation coefficient(SCC) in the case of activation was obtained for a 100 miliseconds window using both MFCC and pitch and intensity. However, there is very little improvement compared to using only MFCCs: the SCC value for MFCC is 0.1731 and when pith and intensity features are added it increases to 0.1770.

In the case of valence, the scores were slightly lower: the highest SCC was obtained again for the smallest analysis window: 0.14, and again the addition of pitch and intensity features did not make much difference. These results show that there is it very difficult to predict a valid value for valence or for activation based on the SVR model, and that the situation is a bit more optimist in the case of activation.

Besides SVR we have also performed classification. The classification accuracy is depicted in Table 5.19. Again, in the case of activation the performance is better.

5.5.3 Discussion

By listening to the clips in the HUMAINE database, we can see that there are no extreme emotions involved, but many blended emotions. Another proof is that by examining the annotations for valence and arousal, they never go close to 100 or -100 which were the labels for extreme emotions. This can be observed from the plot in Figure 5.18 where the labels of all the files which we used from this database are plotted. We can see that they are clustered around the center and rarely get over into the extreme parts of the graph. However, this is probably what we can expect from real data. Probably the lack of intensity of these emotions is one of the reasons for the low inter-rater correlation shown in Table 5.20.

We expect that by visual inspection of the time aligned annotations for valence and activation and the values for pitch and intensity, we can see a correlation between the annotation and the trends in these features. However, there is no clear indication of such a correlation, as can be observed from Figure 5.19, where we decided to plot the data for the same speech sample for pitch, intensity, valence and arousal. It is very difficult to come to a conclusion by visual inspection. What we would expect is for instance to have strong positive correlation between pitch and activation for example. It is also very difficult to draw any conclusion by looking at the intensity plot and comparing it to the annotation plots. Only valence and pitch seem to be correlated, but we can see that the problem we are dealing with is complex.

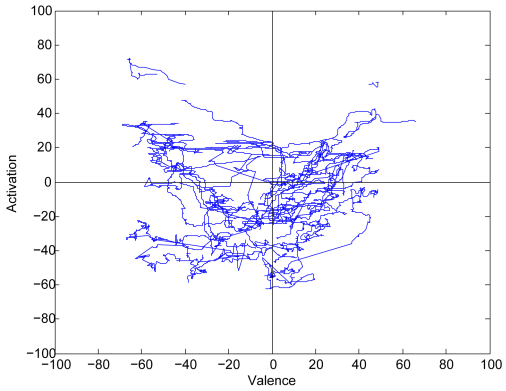


Figure 5.18: 2-dimensional trace annotation for all files used from HUMAINE database

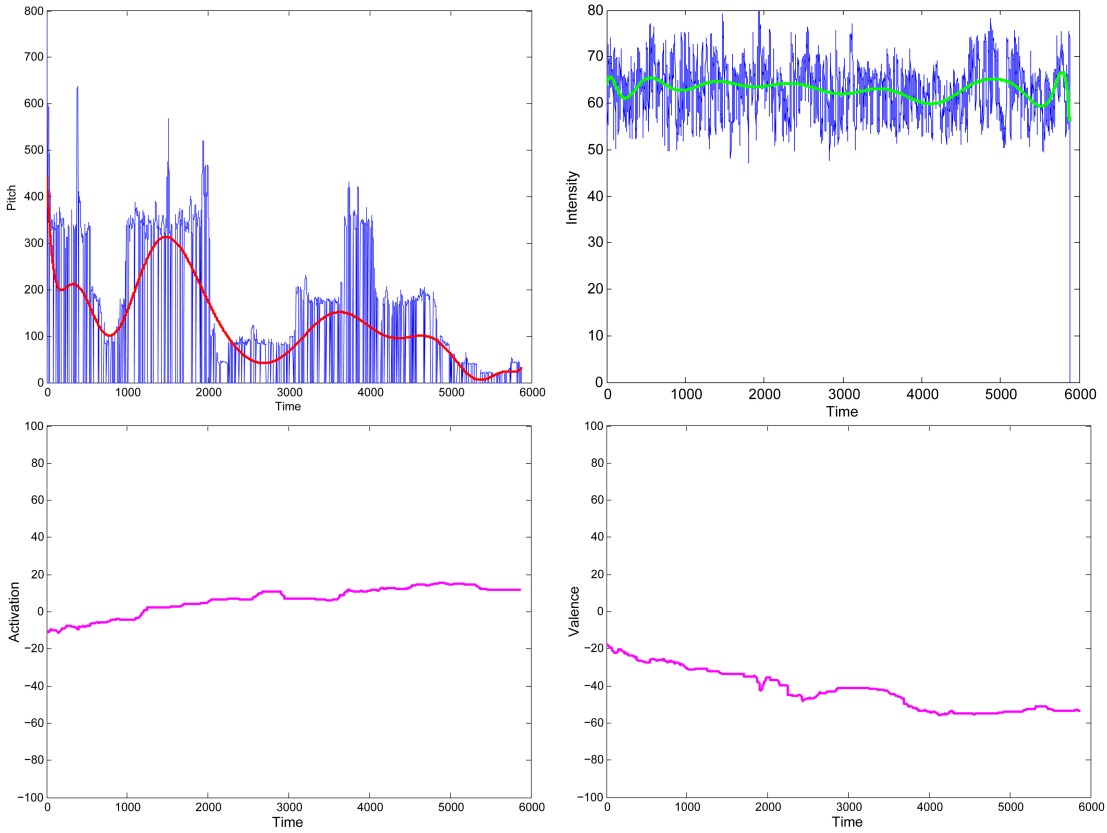


Figure 5.19: Plots of pitch, intensity, valence and activation for one clip from the HUMAINE database

5.5.4 Conclusion

Obtaining reasonable results for regression on the HUMAINE database seems to be very difficult. There are more possible causes for the small correlation coefficients we found.

First of all we have to come back to our first concern which is the labelling of the data. It might be indeed that the annotation is purely directional, and we can not put very much base on the absolute values. We can not say for sure that a value of 50 for valence (on a scale from -100 to 100) is necessarily higher that a value of 40 on the same scale in a different clip or for a different annotator. Therefore, we can expect that machine learning is not very feasible under the circumstances.

Another reason might be that the database contains only mild emotions, there are not really pure intense emotions recorded. This means that on one hand it is also more difficult for the human labelers to give a consistent annotation, and also that the learning process on the data is expected to be difficult. This is also demonstrated by the small inter-labeler agreement (see 3.1.4). Furthermore, the accuracy when just the extreme classes are selected is also not too high. Probably by having more knowledge about the conditions of the recording in the HUMAINE database, and how they were annotated, we would be able to have a clearer view on the problem and find more suitable solution to it.

EmoReSp - A Real-Time System for Emotion Recognition from Speech

6

The final part of the project involves the development of a system that is able to recognize emotions given a speech input, in a real-time setting. The purpose of the real-time system is to demonstrate and give a more intuitive view of the work done within this project. It is a very suitable method for experimenting and finding out the shortcomings and strengths of the emotion recognizer. The system is not designed for a specific application. Instead, it can be regarded as a very general application which is supposed to be suitable for most environments and users.

The emotion recognizer should fulfil the following requirements:

- it should acquire sound input from a microphone,
- it should analyse the input and provide hints (or probabilities) for different emotions,
- it should give the results almost in real time.

6.1 Design

In chapter 4 we show several models which can be used for emotion recognition from speech. Using the first model would mean that we are basing our analysis on the spoken words and the DAL scores for valence and arousal. This can of course be an option, but it means that we should use a very powerful real-time speech recognizer. Even though speech recognition has improved very much in the past years, 100% recognition accuracy is not realized. This means that the error made by the speech recognizer will propagate to the emotion recognizer, and this would make the assessment of the emotion recognizer more difficult.

The second model, described in section 4.3, is a strong candidate for our real-time system. It implies that the system should output a result in the 2-dimensional space of valence and arousal, and that we should be able to track how the emotions are changing. The prerequisites of such a system include an emotional database annotated on valence and arousal. It is important that this database should include major changes in the emotional states that should be tracked, and also more intense emotions. Since in the HUMAINE database the emotions are not too intense it is very difficult to build a system that will learn from it and will be able to predict the dynamics of valence and arousal for new speech samples. Therefore, we refrain from using this model but we believe that it is a very interesting topic for future research, including its combination with the first model.

We consider the third model (section 4.4) the most suitable choice in our case. It means that the situation is similar with the ones presented in experiments 1-3: we are concerned with the acoustics of speech, we can use databases which are split according to emotion classes, and we can use classifiers to build models and predict emotions on new data.

The decisions made for building the real-time system are based on the knowledge we have gathered so far. Looking at the third experiment, where we can compare the performance of different classifiers and different features, we see that the best performing individual classifier is SVM using the Praat features (see section 3.2.1.1). The highest performance however is obtained when all the classifiers are fused, but this would be very difficult to achieve in a real-time application. One reason would be for instance the t-normalization based on a development set which appears to be necessary. The best results would probably be achieved when we can build a development set out of the history of the real-time recordings. This presumes of course that labelling of this data can be provided, which is usually not the case. The availability of such development data could also be used for GMM adaptation, and the resulting system can be expected to perform better.

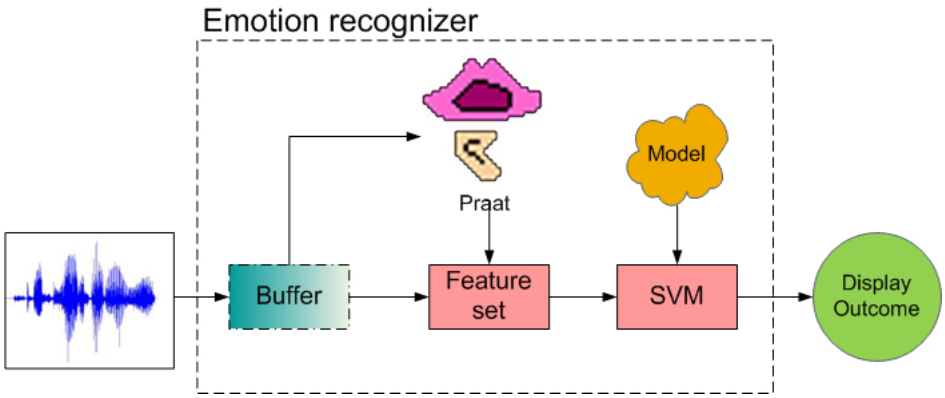


Figure 6.1: Design of the real-time emotion recognizer

Another important aspect is the feature normalization at different granularities. So far we have used z-norming over feature type for utterance level features, and over utterances in the case of frame level features. Deciding how this normalization should be done in real-time is a also a challenge. In the case of utterance level features, we can do the normalization using the mean and standard deviation of the development set. In the case of frame level features, we have to come up with some unit of history for which we can normalize, which can also be the entire history.

For our application we decide not to delve into the problem of fusing more classifiers in real time, but to stick to the best performing individual classifier: SVM using Praat features. The design of our system can be seen in Figure 6.1.

The decision for a database to use was based on the findings of experiment 2 (see section 5.3), where we explored the generalization capabilities enabled by using a larger emotional speech corpus. Even though the results of the experiment do not give a very clear direction for using a larger dataset, we still believe that a system trained on more

databases is more likely to perform well on unseen data, especially if the unseen data is not very similar to the recordings of one of the databases. This is why we decide to use a model trained on the three acted databases Berlin, DES and ENT, that contains the three emotions: anger, happiness and sadness.

The choice of the model is not limited (here by model we mean the model built by the classifier). The system can be easily adapted to a new model given that the train data is provided. The results of the emotion recognition process are displayed as probabilities for each emotion in the model.

6.2 Implementation Details

For implementing the model we had to find an environment that would fulfil the following requirements:

- acquire sound from the microphone and to buffer it,
- communicate with Praat,
- there should exist a LIBSVM library for that environment,
- provide facilities for developing a graphic user interface.

The main candidate environments were MATLAB and C#. For both of them sound acquisition was possible. However, C# does not provide integrated support for audio recording, but implementation attempts are already available from the community. Recording sound using MATLAB is very easy, but the drawback is that we could not buffer the sound while is was being acquired. The Data Acquisition Toolbox of MATLAB provides an object called Analog Input which provides buffering functionalities. For the rest of the requirements, both environments were able to communicate with Praat, LIBSVM libraries are available for both, and they also provide GUI capabilities. Our final choice was MATLAB mainly because of the Data Acquisition Toolbox which we found very useful.

The Analog Input (AI) object converts real-world analog signals (in our case the signal from the microphone) into bits that can be read by computer. Its use can be adapted very easy to many applications, since it has a large set of properties and functions. The main functionality of this object is to collect data, and when a command is called it will dump the data into a variable. The first step for using AI is creating a channel, which in our case is the standard microphone. A set of properties and their value in this application are listed in Table 6.1.

The sample rate is the rate at which an analog input subsystem converts analog data to digital data. A trigger of the AI object is an event that initiates data logging to memory or to a disk file. There are more trigger types available:

- *immediate* - starts as soon as the AI object starts,
- *manual* - starts when a trigger function is manually issued, and
- *software* - occurs when a specific trigger condition is satisfied.

Table 6.1: Parameters of the Analog Input object

Attribute	Value
Sample rate	8000 (samples/sec.)
Samples per trigger	8000 (samples)
Trigger type	Software
Trigger condition	Rising
Trigger condition value	0.2 (voltage)
Trigger repeat	infinity
Trigger delay	−500 (samples)
Timer period	0.5 (sec.)

Our choice was the *software* trigger type, since we intend to acquire data only if somebody is indeed speaking, and otherwise the system should be in a stand-by mode. There are more condition types for the software trigger, from which our option was to use the *rising* condition, which appears to work well for voice activity detection. This means that data acquisition begins when the signal has a positive slope when passing through the specified threshold value. In our case, when the trigger is issued, the AI object takes 8000 samples from the signal (equivalent for 1 second). By empirical testing we found out that many times the very first bits of a word can be missed in this way. Therefore, we set the trigger delay property to −500 samples, which implies that when the trigger is executed, the AI object collects 8000 samples, but starting with the previous 500 samples. The trigger repeat property is set to infinity, so any time there is a new sound input, the system should leave the stand-by mode and start collecting sound. The buffer can be regarded as a queue (FIFO) implemented using a sliding window with different amount of overlap for different units of analysis.

A very important problem for the real-time system is calibrating the parameters. By this we refer to threshold for the trigger to be issued which is determining voice activity detection, which should have different values in silent or in noisy environments for instance. An even more important parameter is the unit of analysis: how much speech do we need to acquire so that we can gather information about emotions. There is still no clear answer to this question. Looking at the recordings in the database we can expect that a value of 1,2 or 3 seconds could give a reasonable result. These are values with which we have experimented.

Besides triggers, there is another function which played a very important role in our system: the timer function. The analog input object has a timer period attribute, which specifies the interval at which the timer function is executed. In our case, the most important actions are executed in the timer function. This mainly means that we are checking if the desired amount of data has been acquired and if so, we are analysing these data and outputting a result. A detailed picture of the data acquisition along with the other operations is available in Figure 6.2.

For analysing the data we need to extract features, and this is done using Praat. The communication between MATLAB and Praat is done using the *sendpraat* program introduced in see section 3.2.1.2. It is important to save the data from the buffer to a temporary file, where Praat can find it. With *sendpraat* we send a command to Praat

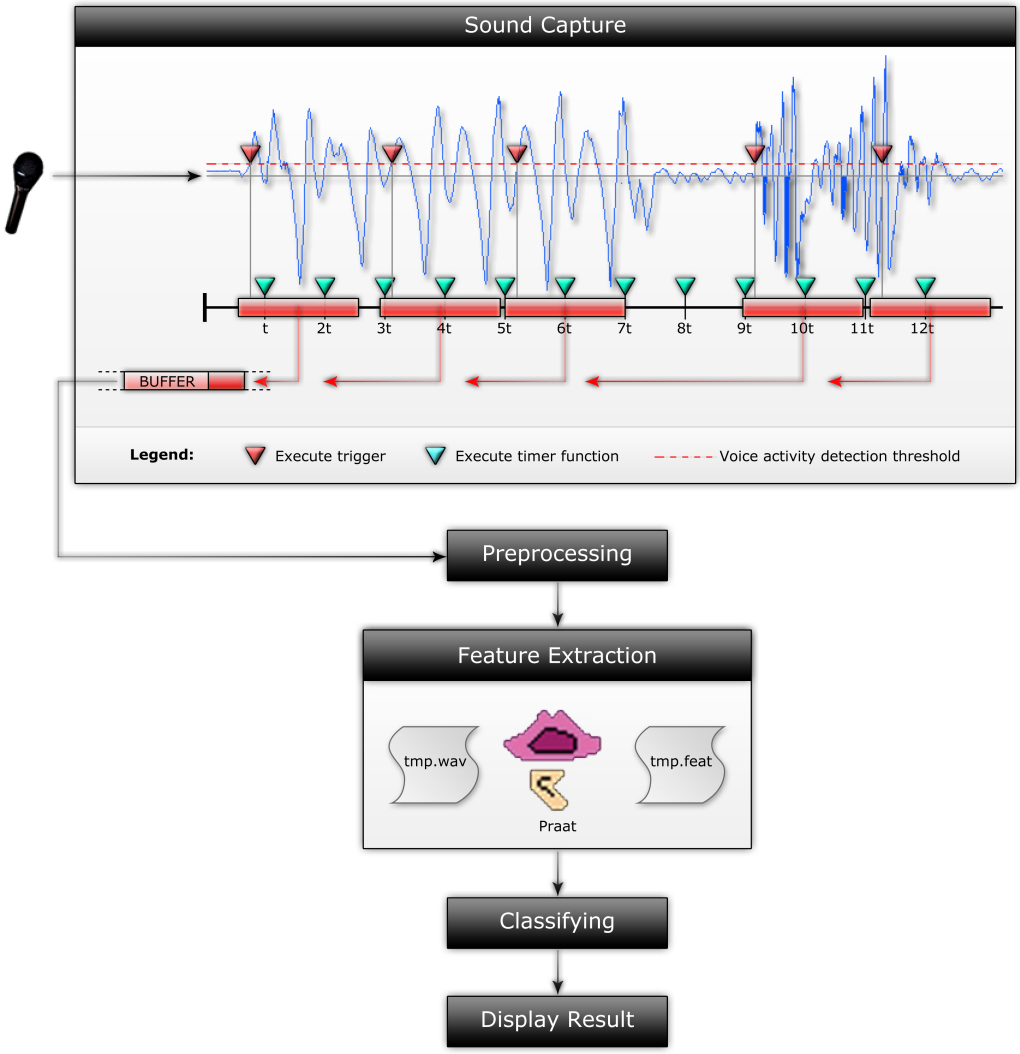


Figure 6.2: Action flow in the emotion recognizer

(which is listening) to execute a feature extraction script on the temporary file. The result is then outputted to a feature file and read into MATLAB.

After the features are available in MATLAB we need to normalize them. We use z-normalization on the training set (which is the collection of more databases). This having been done, the classification process can start. Using the LIBSVM library for MATLAB everything goes very smooth. The output of the classification is a set of probabilities of the analysed sample of speech belonging to each of the classes in the model. These are then outputted a bar graphs.

When the program is opened the important parameters are initialized, including the SVM model. In the first stage the system is inactive and after pressing the Start button we can begin experimenting. Once started, the emotion recognizer is listening and waiting for triggers to be issued. The triggers determine a certain amount of data

to be sent to the buffer. Every 0.5 seconds the timer function is called, and it looks at the amount of data in the threshold. If the data is considered sufficient for analysis, it is saved to a temporary file. From there on Praat and the LIBSVM library are employed for getting the final outcome.

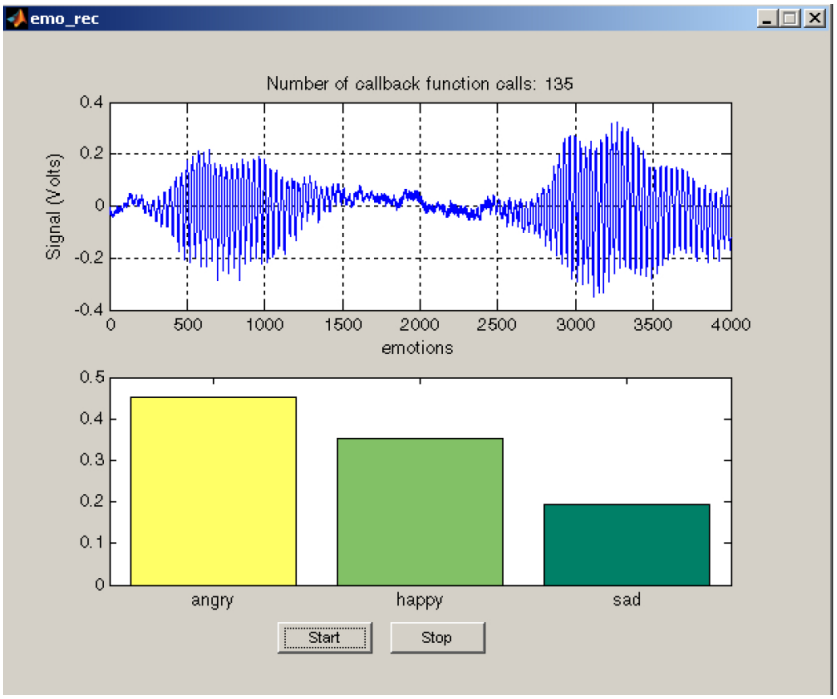


Figure 6.3: GUI of the emotion recognizer

6.3 Testing EmoReSp

Testing the real-time emotion recognition system in a proper way is a fairly difficult problem because there is not really a metric available for this. Also, we do not know for sure the what would have been the correct outcome. If we test the emotion recognizer on recordings from the databases we do nothing else but evaluating the performance of the model, which we have already done in the previous experiments. What we really want to see is if the system based on the extended corpus of acted data is able to give an output for new data that can actually make sense. Using actors and asking them to portray different emotions can be an option for testing, but again, there is no accurate way of measuring how well the system performs because the result is depending on how well the participants are acting, how good the model fits their concept for those emotions, and in the end there is no standard for such a thing.

Our decision is to ask a small number of 5 users to test our system in an unconstrained manner, to speak with different emotions, and see if the output of the system is correlated to their intended emotion. The setting of the experiment is depicted in Figure 6.4. They are asked to rate the correlation they found answering the following question on a 4-point

Figure 6.4: Testing EmoReSp

scale:

Did the output of the emotion recognizer match the emotions you were trying to express?

- not at all,
- very little,
- to some extent,
- very well.

After the test session we had a small discussion with the participants in order to get more insight in their experience. The outcome is the following: all participants chose option (c) for answering the previous question. In the discussion, we noticed that some emotions were easier to recognize than others. There were participant for which anger was always confused with happiness, and some for which sadness was difficult to recognize. Happiness seems to be easily detected, but not for everybody. It is interesting that pure laughter was also interpreted as happiness, even though no training data was available for laughter.

6.4 Conclusion

As we have tested the system with different units of analysis, we can also try to have some observations on that. However, we have to remind the reader that no metric was used for these finding, and everything is based on common sense. A first remark was that the users try to say something and expect immediately a result. The sound captured might be shorter than the unit of analysis, and in this case there is no output. The sound is merged with later data, which we regard of a drawback of the choice of fixed length unit of analysis. We believe that using segments between pauses can be a better way for the decision of what segments to use. However, this is not always feasible because sometimes we are dealing with speech segments that are very short for extracting any emotion. Furthermore, Praat gives errors when the sound to be analysed is too short for the computation of certain statistics, and it can even be that some features can not be calculated.

It is difficult to draw a conclusion about the most suitable unit of analysis, but a hint would be that one second is too short, and perhaps three seconds is too long. When analysing three seconds of speech, with an overlap of two seconds, many times the output changes very fast and in a strange way. This can be caused by the location of the emotion in the speech segment. For example, when we analyse a speech segment containing anger, after which the state is becoming neutral, and we use a sliding window over this segment, the trend of the features is very different in the windows, and this can lead to wrong interpretations.

We can conclude that building a real-time system is an effective way of exploring the problems of emotion recognition from speech. Also, we can see that the emotion recognizer gives the expected outcome in some percent of the cases, but coming up with more reliable metrics is not really possible. However, we can see that we are on the right track and, of course, that more research is needed to improve emotion recognition from speech.

7 Conclusions and Future Work

7.1 Conclusions

The research conducted in this thesis aimed at finding and applying an appropriate model in order to extract emotional content from speech. This meant in our case to learn from the state of the art of emotion recognition, to examine it with a critical eye, try to keep the best aspects and combine them in a proper way. For searching for a proper way more experiments have been designed, each aiming at clarifying one or more questions. After completing all the research steps we had planned, it is time to come back to our research goals and draw some conclusions.

7.1.1 Models of Emotion Recognition from Speech

The initial step was designing a model for emotion recognition from speech. This was done based on the knowledge gathered by examining more models which were available from psychology, analysing the way in which humans perceive emotions, and trying to adapt this information to the requirements of an automated system. The chosen model is based on the idea that emotions are regarded in a discrete way. From the practical perspective the model is appropriate for using machine learning techniques, and results of several tests show that it is a successful choice.

7.1.2 Combining Emotional Speech Databases

A prerequisite for using machine learning for emotion recognition is the availability of databases of emotional speech which should be regarded as the examples generating the model. There is a lot of debating in the research community about shifting from databases of acted speech to ones of spontaneous speech. Privacy issues and sparse data problem are just an example of the problems that appear when real data is involved. However, in this thesis we had the opportunity to experience both acted and spontaneous databases.

Our research is aiming at building a robust and portable emotion recognizer. Therefore, the performance of emotion recognition using the same model has been tested against several databases, including experiments were the training and testing sets were extracted from different databases, and training and testing on a merged corpus from more databases. Our expectation was that a model built based on one database would have low performance when used as predictor for another database, due to the fact that the databases lack generality and determine database-tailored models. This expectation was confirmed, including the remark that some databases have more generalization power than others. Another expectation was that when we want to have results on one database, there should be an improvement when we use a merged training corpus rather

than individual databases. However, the results do not lead to a very clear conclusion in this direction. We get the impression that it is very important how similar the databases from training and testing are. For a more reliable result in this direction we believe that more general databases should be used, and perhaps one additional requirement should be that the recordings from all databases involved should be in the same language.

7.1.3 Key Parameters

Building an emotion recognizer based on speech and according to the chosen model is a process that needs to take into account several important parameters. After the database is chosen, a relevant set of features for emotional speech needs to be decided upon. The decision of a feature set should be based on knowledge about the features and their changes in behavior when emotional stimuli are present. Choosing such a set can be regarded as a compromise, or at least at this moment there is no clear recipe for an optimal choice. However, a good choice should avoid redundancies and still capture the most relevant information. As opposed to recent trends in choosing the features set by a selection based on the performance of a larger feature set on the data, we propose a minimal feature set that proves to be relevant and database independent.

Our experience with acted and spontaneous databases confirms the belief that real data is more problematic. We have used two databases of spontaneous speech which are very different. SADB was recorded in call-centers while the HUMAINE database mostly contains recording from TV-shows. The call center data fitted very well the model used also for the acted databases, and encouraging results were obtained using classification techniques. On the other hand, the HUMAINE database provided a completely different experiment setting which actually involved changing the model. As opposed to classification and detection where we can say that the results are satisfying, we were not able to predict values for the continuous model of emotion (valence and arousal). Therefore, we can say that besides the features chosen, the unit of analysis and other parameters that can be tuned, the database which is one of the most important parameters.

7.1.4 Fusing Classifiers

As expected from literature, we found out that using different types of features, different classifiers and in the end fusing everything in an optimal manner can lead to strong improvements of the results. Furthermore, by fusing the results of different classifiers using logistic regression we receive well-calibrated log-likelihood scores which can be used in a decision process.

7.1.5 Building a Real-Time System

Besides the off-line analysis mentioned so far, we have developed a real-time system that aims at detecting emotions from speech, called EmoReSp. This proved to be a very good modality of exploring the problems that can appear when emotion recognition from speech is required on line, which of course in the end is the aim of all this research. We have found that in real time the situation is very different than with using databases, because we no longer have boundaries for the speech that needs to be analyzed at once.

Therefore, we have experimented with different units of analysis, but because of the real-time conditions we are not able to draw a very clear conclusion. Nevertheless, analysing segments of 2 seconds seemed a good compromise.

To conclude, we have focused our work on building a system that is general and robust. We have inspected therefore several opportunities that could have provided this, as we have tried to base our decisions from what seemed best and most reasonable from the literature. Hopefully, this work is a step forwards in affective computing research.

7.2 Future work

Our analysis, trials, and of course the confirmations of our expectations or sometimes their lack, lead of course to many ideas for future work and improvements.

Even though it is something that researchers said many times before, we need to mention one more time that there is a strong need for new databases and of course especially databases of real emotional speech. It is obvious that all the research in this area is depending on the databases, and databases of real speech bring new challenges that need to be overcome. This way the research will get closer to the real-life application purpose. Besides this, we felt the need of standards in labelling of emotional states.

The general feature set that we have proposed is apparently a good minimal set. However, more research could focus on feature selection based on different databases, including ones with real speech, and a more advanced set could be found.

We believe that there is a strong opportunity to build more robust and portable systems by using extended corpora. However, the data we used did not help proving completely our theory. It is very likely that by using more appropriate databases from training as well as for testing, and also testing with different types of unseen data, to find out that the model is more general than the ones from individual databases.

A last direction is based on our experience with the real-time emotion recognizer. We suggest that by using frame level features and a better limitation of segments between pauses, the performance of the system can be improved. A later step would be to fuse classifiers in real-time and to include adaptive learning.

Bibliography

[Ang *et al.*, 2002] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. 2002. Prosody-Based Automatic Detection Of Annoyance And Frustration In Human-Computer Dialog. *Pages 2037–2040 of: in Proc. ICSLP 2002*.

[Auckenthaler *et al.*, 2000] Auckenthaler, R., Carey, M., & Lloyd-Thomas, H. 2000. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 42 – 54.

[Audibert *et al.*, 2008] Audibert, N., Auberg, V., & Rilliard, A. 2008. Acted vs. spontaneous expressive speech: perception with inter-individual variability. *In: in Proc. LREC 2008*.

[Austermann *et al.*, 2005] Austermann, A., Esau, N., Kleinjohann, L., & Kleinjohann, B. 2005 (Aug.). Prosody based emotion recognition for MEXI.

[Batliner *et al.*, 2003] Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E. 2003. How to find trouble in communication. *Speech Communication*, 117 – 143.

[Batliner *et al.*, 2006] Batliner, A., Steidl, S., Schüller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., & Aharonson, V. 2006. Combining Efforts for Improving Automatic Classification of Emotional User States. *In: Proc. IS-LTC 2006*.

[Batliner *et al.*, 2008] Batliner, A., Steidl, S., & Nth, E. 2008. Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus.

[Bechara & Damasio, 2005] Bechara, A., & Damasio, A. R. 2005. The somatic marker hypothesis: a neural theory of economic decision. *Games and Economic Behavior*.

[Boersma & Weenink, 2009] Boersma, P., & Weenink, D. 2009. *Praat: doing phonetics by computer (Version 5.1.10) [Computer program]*. download at from http://www.praat.org/.

[Boser *et al.*, 1992] Boser, B. E., Guyon, I. M., & N., Vapnik V. 1992. A training algorithm for optimal margin classifiers. *5th Annual ACM Workshop on COLT*, 144–152.

[Brümmer, 2009] Brümmer, N. 2009. Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics. *In: ISCA Proc. Interspeech*. ISCA. sumbitted.

[Brümmer & du Preez, 2006] Brümmer, N., & du Preez, J. 2006. Application-independent evaluation of speaker detection. *Computer Speech Language*, 230 – 275. Odyssey 2004: The speaker and Language Recognition Workshop - Odyssey-04.

[Burkhardt *et al.*, 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. 2005. A database of German emotional speech. *Interspeech*, 1517–1520.

[Busso & Narayanan, 2008] Busso, C., & Narayanan, S.S. 2008 (May). Recording audio-visual emotional databases from actors: a closer look. *Pages 17–22 of: Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008).*

[Chang & Lin, 2001] Chang, C.-C., & Lin, C.-J. 2001. *LIBSVM : a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[Chitu *et al.*, 2008] Chitu, A., van Vulpen, M., Takapoui, P., & Rothkrantz, L. J. M. 2008. Building a Dutch Multimodal Corpus for Emotion Recognition. *Pages 53–56 of: LREC 2008, Workshop on Corpora for Research on Emotion and Affect*ELRA, for ELRA.

[Chuang & Wu, 2004] Chuang, Z., & Wu, C. 2004. Multi-Modal Emotion Recognition from Speech and Text. *Pages 45–62 of: In ISCSLP 2002.*

[Cichosz & Slot, 2007] Cichosz, J., & Slot, K. 2007. Emotion recognition in speech signal using emotion extracting binary decision trees. *In: In ACII.*

[Collobert & Bengio, 2001] Collobert, R., & Bengio, S. 2001. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. *Journal of Machine Learning Research*, 143–160.

[Cortes & Vapnik, 1995] Cortes, C., & Vapnik, V. N. 1995. Support–Vector Networks. *Machine Learning Journal*, 273–297.

[Cowie & Cornelius, 2003] Cowie, R., & Cornelius, R. R. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, April, 5–32.

[Cowie *et al.*, 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. 2001. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 32–80.

[Datcu & Rothkrantz, 2006] Datcu, D., & Rothkrantz, L. J. M. 2006. The recognition of emotions from speech using GentleBoost classifier. A comparison approach. *In: International Conference on Computer Systems and Technologies - CompSysTech06.*

[Dempster *et al.*, 1977] Dempster, A. P., Laird, N. M., & B., Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal Royal Statistics Society*, 1–21.

[Desmet, 2002] Desmet, P. 2002. *Designing Emotion.* Ph.D. thesis, Delft University of Technology.

[Devillers & Vidrascu, 2006] Devillers, L., & Vidrascu, L. 2006 (September). Real-life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs. *Pages 801–803 of: In Interspeech 2006 - ICSLP.*

[Douglas-Cowie *et al.*, 2003] Douglas-Cowie, E., Campbell, N, Cowie, R., & Roach, P. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication*, 33 – 60.

[Douglas-Cowie *et al.*, 2007] Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J.-C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., & Karpouzis, K. 2007. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. *Pages 488–500 of: ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*. Berlin, Heidelberg: Springer-Verlag.

[Drucker *et al.*, 1996] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. 1996. Support Vector Regression Machines. *Advances in Neural Information Processing Systems*, 155–161.

[Egan, 1975] Egan, J. P. 1975. *Signal Detection Theory and ROC Analysis*. Academic Press.

[Ekman, 1994] Ekman, P. 1994. *The nature of emotion: Fundamental questions*. Oxford University Press. Chap. All emotions are basic, pages 7–19.

[El Ayadi *et al.*, 2007] El Ayadi, M.M.H., Kamel, M.S., & Karray, F. 2007 (April). Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models. *Pages IV–957–IV–960 of: Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*.

[Engberg & Hansen, 1996] Engberg, I. S., & Hansen, A. V. 1996. *Documentation of the Danish Emotional Speech Database (DES)*. Internal AAU report, Center for Person Kommunikation.

[Fitrianie & Rothkrantz, 2006] Fitrianie, S., & Rothkrantz, L. J. M. 2006. Two-Dimensional Visual Language Grammar. *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 4188*, 573–580.

[Forbes-Riley & Litman, 2004] Forbes-Riley, K., & Litman, D. J. 2004. Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. *Pages 201–208 of: HLT-NAACL*.

[Fragopanagos & Taylor, 2005] Fragopanagos, N., & Taylor, J. G. 2005. Emotion recognition in human-computer interaction. *Neural Networks*, 389 – 405. Emotion and Brain.

[Frijda, 1986] Frijda, N. H. 1986. *The emotions*. Cambridge University Press.

[Graciarena *et al.*, 2006] Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., & Kajarekar, S. 2006 (May). Combining Prosodic Lexical and Cepstral Systems for Deceptive Speech Detection. *Pages I–I of: Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*.

[Grimm *et al.*, 2008] Grimm, M., Kroschel, K., & Narayanan, S. 2008 (23 2008-April 26). The Vera am Mittag German audio-visual emotional speech database. *Pages 865–868 of: Multimedia and Expo, 2008 IEEE International Conference on.*

[Hermansky, 1990] Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis for speech. *J. Acoust. Soc. Am.*, 1738–1752.

[Hermansky *et al.*, 1992] Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. 1992 (Mar). RASTA-PLP speech analysis technique. *Pages 121–124 vol.1 of: Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on.*

[Hess, 1992] Hess, W. J. 1992. *Furui, S., Sondhi, M.M. (Eds.), Advances in Speech Signal Processing.* Marcel Dekker. Chap. Pitch and voicing determination.

[Hsu & Lin, 2002] Hsu, C.-W., & Lin, C.-J. 2002. A comparison of methods for multiclass support vector machines. *Pages 415–425 of: IEEE Transactions on Neural Networks.*

[Hu *et al.*, 2007] Hu, H., Xu, M.-X., & Wu, W. 2007 (August). Fusion of Global Statistical and Segmental Spectral Features for Speech Emotion Recognition. *Pages 2269–2272 of: In Interspeech 2007 - Eurospeech, 10th European Conference on Speech Communication and Technology.*

[Izard, 1977] Izard, C. 1977. *Human emotions.* Plenum Press.

[James, 1884] James, W. 1884. *What is an emotion?* In Mind.

[Johnstone & Scherer, 1999] Johnstone, T., & Scherer, K. R. 1999. The effects of emotions on voice quality.

[Juslin & Scherer, 2005] Juslin, P.N., & Scherer, K.R. 2005. *In J. Harrigan, R. Rosenthal, K. Scherer, (Eds.) - The New Handbook of Methods in Nonverbal Behavior Research.* Oxford University Press. Chap. Vocal expression of affect, pages 65–135.

[Kim *et al.*, 2007] Kim, S., Georgiou, P.G., Lee, Sungbok, & Narayanan, S. 2007 (Oct.). Real-time Emotion Detection System using Speech: Multi-modal Fusion of Different Timescale Features. *Pages 48–51 of: Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on.*

[Krajewski & Kroger, 2008] Krajewski, J., & Kroger, B. 2008 (August). Using Prosodic and Spectral Characteristics for Sleepiness Detection. *Pages 1841–1844 of: In: Interspeech 2007 - Eurospeech, 10th European Conference on Speech Communication and Technology.*

[Laukka *et al.*, 2008] Laukka, P., Elenius, K., Fredrikson, M., Furmark, T., & Neiberg, D. 2008. Vocal expression in spontaneous and experimentally induced affective speech: Acoustic correlates of anxiety, irritation and resignation. *In: LREC 2008.*

[Laver, 1980] Laver, J. 1980. *The Phonetic Description of Voice Quality.* Cambridge University Press.

[Lazarus, 1999] Lazarus, R. S. 1999. *Stress and Emotion: A New Synthesis.* New York: Springer.

[Lee & Narayanan, 2005] Lee, C. M., & Narayanan, S.S. 2005. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing, IEEE Transactions on*, March, 293–303.

[Lee *et al.*, 2004] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., & Narayanan, S. 2004. Emotion Recognition based on Phoneme Classes. *In: Eighth International Conference on Spoken Language.*

[Lin & Wei, 2005] Lin, Y. L., & Wei, G. 2005 (Aug.). Speech emotion recognition based on HMM and SVM. *Pages 4898–4901 Vol. 8 of: Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on.*

[Martin *et al.*, 1997] Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. 1997. The Det Curve In Assessment Of Detection Task Performance.

[Martin *et al.*, 2006] Martin, O., Kotsia, I., Macq, B., & Pitas, I. 2006. The eNTERFACE'05 Audio-Visual Emotion Database. *Data Engineering Workshops, 22nd International Conference on.*

[Matos *et al.*, 2006] Matos, S., Birring, S. S., Pavord, I. D., & Evans, H. 2006. Detection of cough signals in continuous audio recordings using hidden Markov models. *Biomedical Engineering, IEEE Transactions on*, 1078–1083.

[McNahon *et al.*, 2008] McNahon, E., Cowie, R., Wagner, J., & Andre, E. 2008. Multimodal records of driving influenced by induced emotion. *In: In LREC 2008.*

[Morrison *et al.*, 2007] Morrison, D., Wang, R., & De Silva, L. C. 2007. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 98 – 112.

[Neiberg *et al.*, 2006a] Neiberg, D., Elenius, K., Karlsson, I., & Laskowski, K. 2006a. Emotion Recognition in Spontaneous Speech. *Pages 101–104 of: Working Papers 52: Proceedings of Fonetik 2006.*

[Neiberg *et al.*, 2006b] Neiberg, D., Elenius, K., & Laskowski, K. 2006b. Emotion Recognition in Spontaneous Speech Using GMM. *Pages 809–812 of: Proc. Int. Conf. Spoken Language Processing (ICSLP '06).*

[Nicholson *et al.*, 1999] Nicholson, J., Takahashi, K., & Nakatsu, R. 1999. Emotion recognition in speech using neural networks. *Pages 495–501 vol.2 of: Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on.*

[Nogueiras *et al.*, 2001] Nogueiras, A., Moreno, A., Bonafonte, A., & Marino, J. B. 2001 (September). Speech Emotion Recognition Using Hidden Markov Models. *In: European Conference on Speech Communication and Technology EUROSPEECH 2001.*

[Nwe *et al.*, 2003] Nwe, T. L., Foo, S. W., & De Silva, L. C. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication*, November, 603–623.

[Ortony *et al.*, 1988] Ortony, A., Clore, G. L., & Collins, A. 1988. *The Cognitive Structure of Emotions.* Cambridge University Press.

[Panksepp, 1982] Panksepp, J. 1982. Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, 407–467.

[Panksepp, 1998] Panksepp, J. 1998. *Affective neuroscience: the foundations of human and animal emotions.* Oxford University Press.

[Pao & Chen, 2003] Pao, T. L., & Chen, Y. T. 2003 (Nov.-3 Dec.). Mandarin emotion recognition in speech. *Pages 227–230 of: Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on.*

[Petrushin, 1999] Petrushin, V. A. 1999. Emotion in Speech: Recognition and Application to Call Centers. *Pages 7–10 of: In Engr.*

[Plutchik, 1980] Plutchik, R. 1980. *In R. Plutchik, H. Kellerman, Emotion: Theory, research, and experience.* Academic Press. Chap. A general psychocvolutionary theory of emotion, pages 3–31.

[Posner *et al.*, 2005] Posner, J., Russell, J. A., & Peterson, B. S. 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 715–734.

[Razak *et al.*, 2005] Razak, A.A., Komiya, R., Izani, M., & Abidin, Z. 2005 (July). Comparison between fuzzy and NN method for speech emotion recognition. *Pages 297–302 vol.1 of: Information Technology and Applications, 2005. ICITA 2005. Third International Conference on.*

[Russell, 1980] Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1161–1178.

[Russell & Barrett, 1999] Russell, J. A., & Barrett, L. F. 1999. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, May, 805–819.

[Schacter & Singer, 1962] Schacter, S., & Singer, J. 1962. *Cognitive, Social and Physiological Determinants of Emotional States. Psychological Review.*

[Scherer, 2000] Scherer, K. R. 2000. A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. *In: Proc. ICSLP.*

[Scherer, 2003] Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 227–256.

[Scherer *et al.*, 2003] Scherer, K. R., Johnstone, T., & Klasmeyer, G. 2003. *Vocal Expression of Emotion.* R. j. davidson, h. goldsmith, k. r. scherer (eds.) edn. Handbook of the Affective Sciences. New York and Oxford: Oxford University Press.

[Schüller *et al.*, 2003] Schüller, B., Rigoll, G., & Lang, M. 2003 (April). Hidden Markov model-based speech emotion recognition. *Pages II–1–4 vol.2 of: Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on.*

[Schüller *et al.*, 2004] Schüller, B., Rigoll, G., & Lang, M. 2004. Speech Emotion Recognition Combining Acoustic Features And Linguistic Information in a Hybrid SVM-BN Architecture. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2004)*, 577–580.

[Schüller *et al.*, 2005a] Schüller, B., Villar, R. J., Rigoll, G., & Lang, M. 2005a (18-23,). Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition. *Pages 325–328 of: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on.*

[Schüller *et al.*, 2005b] Schüller, B., Muller, R., Lang, M., & Rigoll, G. 2005b. Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. *In: Interspeech 2005.*

[Schüller *et al.*, 2005c] Schüller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. 2005c (July). Speaker Independent Speech Emotion Recognition by Ensemble Classification. *Pages 864–867 of: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.*

[Schüller *et al.*, 2007] Schüller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., & Aharonson, V. 2007. The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. *Pages 2253–2256 of:* ISCA (ed), *Proceedings Interspeech.*

[Shami & Verhelst, 2007] Shami, M., & Verhelst, W. 2007. An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *In: Elsevier Editorial System(tm) for Speech Communication.*

[Shami & Kamel, 2005] Shami, M.T., & Kamel, M.S. 2005 (July). Segment-based approach to the recognition of emotions in speech. *In: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.*

[Sousa, 2008] Sousa, R. 2008. *Emotion.* In The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.).

[Steidl *et al.*, 2008] Steidl, S., Batliner, A., Nöth, E., & Hornegger, J. 2008. Quantification of Segmentation and F0 Errors and Their Effect on Emotion Recognition. *Pages 525–534 of: TSD '08: Proceedings of the 11th international conference on Text, Speech and Dialogue.* Berlin, Heidelberg: Springer-Verlag.

[Tomkins, 1984] Tomkins, S. S. 1984. *Affect theory.* In: K. R. Scherer and P. Ekman, Eds., Approaches to Emotion, pp. 163-196, Erlbaum, Hillsdale, NJ.

[Truong & Raaijmakers, 2008] Truong, K. P., & Raaijmakers, S. 2008. Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features. *Pages 161–172 of: MLMI '08: Proceedings of the 5th international workshop on Machine Learning for Multimodal Interaction.* Berlin, Heidelberg: Springer-Verlag.

[Truong & van Leeuwen, 2007] Truong, K. P., & van Leeuwen, D. A. 2007. Automatic discrimination between laughter and speech. *Speech Commun.*, 144–158.

[van Leeuwen & Brümmer, 2007] van Leeuwen, D. A., & Brümmer, N. 2007. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. 330–353.

[van Vulpen, 2008] van Vulpen, M. 2008. *Analysis and Recording of Multimodal Data.* M.Phil. thesis, Delft University of Technology.

[van Willigen, 2009] van Willigen, I. 2009. *Reasoning about Emotions. An affective natural language processing environment, using lexical relations to measure activation and evaluation, and extracting semantics from natural language.* M.Phil. thesis, Delft University of Technology, The Netherlands.

[Ververidis & Kotropoulos, 2006] Ververidis, D., & Kotropoulos, C. 2006. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, September, 1162–1181.

[Ververidis *et al.*, 2004] Ververidis, D., Kotropoulos, C., & Pitas, I. 2004 (May). Automatic emotional speech classification. *Pages I–593–6 vol.1 of: Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on.*

[Ververidis *et al.*, 2008] Ververidis, D., Kotsia, I., Kotropoulos, C., & Pitas, I. 2008. Multi-modal emotion-related data collection within a virtual earthquake emulator,. *In: LREC 2008.*

[Vidrascu & Devillers, 2008] Vidrascu, L., & Devillers, L. 2008. Anger detection performances based on prosodic and acoustic cues in several corpora. *In: LREC 2008.*

[Vlasenko *et al.*, 2007] Vlasenko, B., Schüller, B., Wendemuth, A., & Rigoll, G. 2007. Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. *In: Interspeech 2007.*

[Vogt & Andre, 2005] Vogt, T., & Andre, E. 2005 (July). Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition. *Pages 474–477 of: Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on.*

[Walker *et al.*, 2001] Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Seneff, S., & Stallard, D. 2001. DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection. *Pages 1371–1374 of: In Eurospeech 2001.*

[Watzlawick *et al.*, 1967] Watzlawick, P., Helmick-Beavin, J., & Jackson, D. D. 1967. *Pragmatics of Human Communication.* W. W. Norton Company.

[Whissell, 1989] Whissell, C.M. 1989. *Emotion: Theory, Research and Experience: Vol. 4, The Measurement of Emotions, R. Plutchik and H. Kellerman.* New York: Academic. Chap. The dictionary of affect in language.

[Zeng *et al.*, 2007] Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. 2007. A survey of affect recognition methods: audio, visual and spontaneous expressions. *Pages 126–133 of: ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces.* New York, NY, USA: ACM.

[Zhang *et al.*, 2004] Zhang, T., Hasegawa-Johnson, M., & Levinson, S. E. 2004. Children's Emotion Recognition in an Intelligent Tutoring Scenario. *Proc. of ICSLP.*