OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG
EIT
FAKULTÄT FÜR ELEKTROTECHNIK UND INFORMATIONSTECHNIK
INSTITUT FÜR ELEKTRONIK, SIGNALVERARBEITUNG UND KOMMUNIKATIONSTECHNIK (IESK)

Emotion Recognition within Spoken Dialog Systems

DISSERTATION

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

von

M.Sc. Bogdan Vlasenko

geb. am 06.04.1982 in Cherson, Ukraine

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

Gutachter: Prof. Dr. rer. nat. Andreas Wendemuth
Prof. Dr. rer. nat. habil. Dietmar Rösner
Prof. Dr. Dr.-Ing. Wolfgang Minker

Promotionskolloquium am 21.12.2011

Dedicated to
my parents

Acknowledgements

First of all, I would like to express my utmost appreciation to my scientific advisor, Prof. Dr. rer. nat. Andreas Wendemuth, for his guidance, suggestions and criticism throughout my studies at Otto-von-Guericke-Universität Magdeburg. His responsibility to his students is impressive, which has been inestimable to me. I learned a great deal from his preciseness in mathematics, motivation of concepts and transparent logic flow from his presentations and writing. The high requirements and close guidance on these aspects have given me the capability and confidence to carry out the research work of this thesis as well as works in the future. By initiating well-founded relevant research questions, offering recommendations based on experience and having creative and productive discussions, he is the most important person to have helped me make this work possible!

Special thanks go to Dr.-Ing. Björn Schuller for his helpful ideas during the inception of my research and later collaboration in the detailed research of speech-based emotion recognition. His insight, experience and wide-range of knowledge on emotional speech processing have also benefited me a lot.

I owe thanks for providing financial support for my studies and attendance of many international conferences and workshops through the framework of the project "Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems" (UC4) funded by the European Community through the Center for Behavioral Brain Science, Magdeburg. My research is also associated and supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion- Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

I must also thank my colleagues in the Cognitive Systems group from the Department of Electrical Engineering and Information Technology. Particular thanks Ronald Böck for his participation in collecting and processing data during usability test experiments.

Finally, the biggest thanks go to my parents to whom I owe everything I have! They have offered everything possible to support me all through my life.

Zusammenfassung

Systeme im Bereich der Mensch-Maschine-Interaktion (MMI) im Allgemeinen, und im Speziellen aktuelle Sprachdialogsysteme (SDS), die auf automatischer Spracherkennung (ASR) basieren, haben Defizite bei natürlicher und benutzerfreundlicher Kommunikation. Problematisch dabei ist, dass die meisten Systeme wichtige Informationsquellen über die Aktivität des Nutzers nicht in Betracht ziehen. Dies sind unter anderem die Motivation und Intention sowie der emotionale Zustand des Nutzers. Detaillierte Analysen dieser Eigenschaften können daher bedeutend zu Prinzipien der Entwicklung nutzerfreundlicherer Systeme beitragen. Die Notwendigkeit der Emotionsanalyse in der MMI liegt in den Beschränkungen der ASR: aktuelle, automatische Spracherkennungssysteme können nicht mit flexibler, spontaner, nicht im Vokabular eingeschränkter und emotional gefärbter, d.h. allgemein affektbetonter Sprache umgehen. Daher rückte in den letzten Jahren konsequenterweise die Analyse emotionaler Sprache in den Fokus der ASR und darüber hinaus auch in das Blickfeld der Sprachsynthese. Beide Techniken können einen Beitrag für eine intelligentere und nutzerbezogenere MMI leisten.

In dieser Arbeit werden neue Ansätze zur nutzerbezogenen Interaktion aus der Sicht der automatischen Emotions- und Intentionserkennung aus gesprochener Sprache untersucht. Dabei liegt das Hauptziel auf der Bereitstellung einer effektiven Emotionssprachverarbeitung (Emotionserkennung, Erkennung emotional gefärbter Sprache). Der Beitrag dieser Arbeit ist die Beschreibung affektbetonter Spracherkennungsmethoden auf Basis von Hidden-Markov-Modellen (HMMs) mit Gauß'schen Mischverteilungsmodellen (GMMs). Der dazu verwendete Framework enthält Konzepte der ASR, die auf Aspekten der HMMs/GMMs basieren: Auswahl von Wort-Untereinheiten und deren quantitativen und qualitativen Definitionen, dem Erkennunsgsalgorithmus für spontane Sprache und einem Sprachmodell, sowie Adaptationsverfahren zur robusten Emotionsspracherkennung. Im Speziellen werden Wort-Untereinheiten des Deutschen in der ASR beschrieben. Darüber hinaus werden phonologische Muster mit detaillierten Spezifikationen für Konsonanten, Vokale und Diphthonge des Deutschen vorgestellt. Für die Beschreibung der Vokale und Diphthonge wird das Vokal-Dreieck verwendet, anhand dessen die verschiedenen Charakteristiken von affektbetonter und neutraler Sprache verdeutlicht werden können. In dieser Arbeit wird gezeigt, dass auf Grund der Ähnlichkeiten in den Aussprachemustern von affektbetonter und neutraler Sprache, emotionsabhängige Eigenschaften von existierenden Emotions-Korpora auf andere Sprachkorpora übertragen werden können. Dabei werden die Modellparameter eines neutralen Modells durch geeignete Transformationen so verändert,

dass ein akustisches Modell für emotionale Sprache entsteht. Wir haben die Adaptionsmethoden anhand deutscher Sprachkorpora getestet und einen beachtenswerten Genauigkeitszuwachs für die Emotionsspracherkennung erreicht.

Der zweite Teil der Arbeit beschreibt unsere verschiedenen Methoden zur Klassfikation von Emotionen in detaillierter Weise. In Kapitel 4 geben wir einen Überblick über existierende Techniken der Emotionserkennung aus Sprache und besprechen akustische Features, die für die Unterscheidung von emotionalen Ereignissen am geeignetsten erscheinen. Zwei Klassifikationstechniken werden dabei näher vorgestellt: die statische (turn-level) und die dynamische (frame-level) Methode. Zur Entwicklung der dynamischen Emotionserkennung verwenden wir Hauptkonzepte der aktuellsten Methoden der Spracherkennung, die auf HMM/GMM Modellen basiert. Im Speziellen präsentieren wir verschiedene Methoden der Emotionsklassifikation basierend auf der Analyse unterschiedlicher Einheiten der Spracherkennung: Äußerungen, Satzteilen (Chunks) und Phonemen. Zwei Arten der Analyse auf Phonem-Ebene werden detailliert vorgestellt: emotionale Phonemklassen und Formant-Verfolgung von Vokalen. Darüber hinaus diskutieren wir zwei Arten der Fusion von Klassifikationsergebnissen. Diese sind: zweistufige Fusion und Fusion auf mittlerem Abstraktionsniveau. Abschließend werden die Erkennungsleistungen für einheitenspezifische (kontextabhängige) und allgemeine (kontextunabhängige) Modelle verglichen. Dabei können wir zeigen, dass die Emotionserkennung auf Basis von einheitenspezifischen Modellen solche mit kontextunabhängigen in der Erkennungsleistung übertreffen, vorausgesetzt es steht pro Einheit genügend Trainingsmaterial zur Verfügung.

Beide vorgestellten Ansätze werden auf verschiedenen Sprachkorpora evaluiert. Für die Experimente mit affektbetonter Sprache werden unterschiedliche Strategien zur Verifikation verwendet und diverse Erkennungsmaße benutzt. Durch Verwendung von Formantverfolgung auf Vokalebene können wir zeigen, dass unimodale, akustische Merkmale (gemittelte F1 Werte) stark mit dem Grad der Erregung (arousal) eines Sprechers korreliert sind. Mit diesen Merkmalen, dem Neyman-Pearson Kriterium und einer kleinen Menge an Trainingsmaterial (1-2 Äußerungen pro Sprecher) zur Adaption erhalten wir Ergebisse in der Emotionserkennung, die mit den auf affektbetonten Korpora trainierten Klassifikatoren vergleichbar sind. Mit unserer Methode der Erkennung, basierend auf dynamischer Analyse, und der Verwendung von spektralen Merkmalen (Mel-Frequency Cepstral Coefficients) konnten wir eines der besten Klassifikationsergebnisse auf spontaner, emotionaler Sprache während der INTERSPEECH 2009 Emotion Challenge erreichen.

Einige der Resultate dieser Arbeit wurden in einem prototypischen Dialogsystem, welches vom Autor und einigen Kollegen unter fortdauernder Ko-

operation seit 2005 entwickelt wurde, umgesetzt. Hierbei wurde das System so erweitert, dass es sich an den emotionalen Zustand des Nutzers anpassen kann. In Nutzertests fanden wir heraus, dass besonders in frustrierenden Situationen, ein solches System, mit Adaption an den emotionalen Zustand, erfolgreich Hilfestellungen und Lösungsvorschläge im Zusammenhang mit den aktuellen Aufgaben geben konnte. Sprecheradaptive Sprachdialogsysteme basierend auf akustischer Emotionserkennung in Kombination mit einer affektbetonten Adaption des ASR Modells senken die Zeit, die zur Interaktion und zur Anpassung an das Vokabular benötigt wird, signifikant, wodurch die MMI benutzerfreundlicher und nutzerbezogener wird.

Abstract

General human-machine interaction (HMI) systems, and in particular current state-of-the-art spoken dialog systems (SDS) based on automatic speech-recognition (ASR) technology, have a number of deficiencies in communicating with a user in a natural and friendly way. One problem is that most of these systems do not take into account important sources of the user's activities such as his/her motivation, intention and emotional state. Detailed analysis of these activities could, therefore, be an essential feature of a user-friendly interaction interface. The importance of user's emotional state analysis during HMI lies in existing limitations of ASR: current ASR methods still cannot deal with flexible, unrestricted user's language, spontaneous and emotionally colored speech. Consequentially, emotional speech processing is a topic that has received a great deal of attention during the last decade within speech synthesis as well as in ASR. Emotional speech synthesis and recognition of emotions within HMI can contribute to more intelligent and user-centered interaction.

In this thesis, new approaches for user-centered interaction are investigated from the point of view of emotions and intentions automatically estimated from speech. The main research goal of this thesis is to provide an effective emotional speech processing (emotion recognition, emotional speech recognition). The first contribution of this thesis is to describe automatic affective-speech-recognition methods based on hidden Markov models (HMMs). This framework presents the main aspects of the HMM-/GMM-based ASR concept: *a selection of the sub-word units and their quantitative and qualitative specification, the decoding algorithm for spontaneous speech, a language modeling* and *the adaptation techniques for a robust affective speech recognition.* In particular, the sub-word units selection for German ASR is described. Afterwards, a German phonetic pattern with a detailed specification of all consonants, vowels and diphthongs is presented. For specification of the vowels and diphthongs a vowel triangle is used. By generating vowels triangles for various speaker's emotional states we show the different characteristics of the affective and neutral speech. In this work, we prove that due to the pronunciation pattern similarity of affective and neutral speech, emotion-specific characteristics can be captured from existing emotional speech corpora within adaptive transformation of model parameters of the initial neutral speech model to obtain an emotional speech acoustic model. We investigate the potency of adapting emotional speech acoustic models for the German language and we obtain a considerable performance gain for the emotional speech recognition.

The second contribution of this thesis is to provide a detailed description of our various emotion-classification techniques. In Chapter 4 we present an overview of existing speech-based emotion-recognition techniques, and discuss acoustic feature sets, which are the most informative for emotional events determination. Two different emotion-classification techniques, namely, *static* (turn-level) and *dynamic* (frame-level) are presented. We use the main concepts of state-of-the-art speech recognition based on HMM/GMM models for developing our dynamic emotion-recognition techniques. In particular, we present various emotion-classification techniques with different units of analysis: *utterance, chunk,* and *phoneme.* Two different phoneme-level emotion-classification techniques, *emotional phoneme classes* and *vowel-level formants tracking*, are described in detail. Two possible combined emotion-classification methods, *two-stage processing and middle-level fusion*, are presented. Finally, we compare emotion-recognition performances for *unit-specific* (context dependent) and *general* (context independent) models. We show that the introduced unit-specific emotion-recognition models clearly outperform general models provided sufficient amount of training material per unit.

The above two contributions are evaluated on various speech corpora. For the experiments with affective speech corpora we use various types of evaluation strategies and recognition rate measures. With a vowel-level formants tracing technique we show that the unimodal acoustic features (average F1 values) extracted on a vowel-level are strongly correlated with the level of arousal of the speaker's emotional state. With these features, a straightforward Neyman-Pearson criterion and a small amount of training data (1-2 neutral utterances per speaker) we obtain comparable good emotion-recognition results. With our emotion-classification technique based on dynamic analysis we prove that only by using spectral features (Mel-frequency Cepstral coefficients (MFCC)) can we reach one of the best emotion-recognition performances for spontaneous emotional speech samples evaluated within the INTERSPEECH 2009 Emotion Challenge.

Some of the findings described in this thesis have been incorporated into a prototype dialog system specially developed by the author and colleagues within ongoing funded collaborations (since 2005) in order to demonstrate adaptation of the system to the user's emotional state. Within a usability experiment we find that during frustrating situations in HMI, the SDS with emotional user state adaptation successfully provides comprehensive help and exhaustive recommendations in the context of the current state of the task. The user-behavior-adaptive SDS built upon acoustic emotion recognition in combination with affective-speech-adapted ASR models significantly decreases interaction and vocabulary adaptation time, which shows that HMI becomes more friendly and user-centered.

Table of notations

General Notation:

s	a scalar is denoted by a plain lowercase letter
$\mathbf{v}$	a column vector is denoted by a bold lowercase letter
$\mathbf{A}$	a matrix is denoted by a bold uppercase letter
$Q(\cdot\|\cdot)$	an auxiliary function

Mathematical notation:

$p(\cdot)$	probability density function
$p(\cdot\|\cdot)$	conditional probability density function
$P(\cdot)$	probability mass distribution
$P(\cdot\|\cdot)$	conditional probability mass distribution

Standard HMM notation:

$\mathcal{M}$	parameter set of HMM
$\mathcal{W}$	hypothetical word sequence $\mathcal{W} = [w_1, w_2, \ldots, w_K]$
N	number of HMM's states
$\mathbf{o}_t$	observation vector at time t
$\mathbf{O}$	observation vectors sequence $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$
s_t	state at discrete time t
$\mathbf{s}$	state sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$
a_{ij}	discrete state transition probability from state i to j
$b_j(\mathbf{o}_t)$	state output distribution given state j at time t
$b_{jm}(\mathbf{o}_t)$	state output distribution given state j of m GMM component at time t
$\boldsymbol{\mu}$	mean vector
$\boldsymbol{\Sigma}$	covariance matrix
$\boldsymbol{\mu}_m$	mean vector of the m Gaussian component
$\alpha_j(t)$	forward variable in forward-backward algorithm at time t
$\beta_j(t)$	backward variable in forward-backward algorithm at time t

Acronyms

ABC	Airplane behavior corpus, [Schuller et al., 2009b]
ASR	Automatic speech recognition
ASU	Automatic speech understanding
AVIC	Audiovisual interest corpus, [Schuller et al., 2009b]
CMS	Cepstral mean substraction
CSDS	Conventional spoken dialog systems
DA	Dialog act
DCT	Discrete cosine transform
DES	Danish emotional speech corpus, [Engbert and Hansen, 1996]
DPP	Dynamic programming principles
EM	Expectation maximization
EMO-DB	Berlin emotional speech database, [Burkhardt et al., 2005]
F1	First formant
F2	Second formant
FFT	Fast Fourier transform
FSO	Features set optimization
G2P	Grapheme-to-phoneme
GBC	Global base class
GEW	Geneva emotion wheel
GMM	Gaussian mixture model
HMI	Human-machine interaction
HMM	Hidden Markov model
HNR	Harmonics-to-noise ratio
HTK	Hidden Markov model toolkit
IVR	Interactive voice response
LLD	Low-level descriptors
LOSO	Leave-one-speaker-out
LOSGO	Leave-one-speakers-group-out
MFCC	Mel-frequency cepstral coefficients
ML	Maximum likelihood
MLV	Maximum length vote
MSL	Maximum classifier prediction score multiplied with the length vote
MV	Majority vote
NIMITEK	Neurobiologically inspired multimodal intention recognition for technical communication systems
OOV	Out-of-vocabulary
PDF	Probability density functions
PE	Phoneme emotional

PLOI	Phoneme level of interest
PT	Phonetic transcription
RCT	Regression class tree
RHS	Right-hand side
SAL	Sensitive artificial listener corpus, [Wöllmer et al., 2008]
SCV	Stratified cross-validation
SD	Speaker-dependent
SDS	Spoken dialog systems
SER	Speech emotion recognition
SI	Speaker-independent
SN	Speaker normalization
SUSAS	Speech under simulated and actual stress, [Hansen and Bou-Ghazale, 1997]
SVM	Support vector machine
TASN	Textual associations semantic networks
TUM	Technische Universität München
UA	Unweighted average recall
UASDS	User-adapted spoken dialog systems
VAD	Valence-arousal-dominance
VAM	Vera-am-mittag corpus, [Grimm et al., 2008]
WA	Weighted average recall
WER	Word error rate
WHG	Word hypothesis graph
WOZ	Wizard of Oz

Glossary

- Explanation of terms as they are used in this thesis.
- **Bold** words refer to other entries in this glossary.

Acoustic model	*model which maps the acoustic observation vectors to the phonetic units.*
Adaptation	*model based compensation of acoustic mismatch. correction of user's commands set used during interaction with a system.*
Affective speech	*emotional* **speech.**
Annotation	*emotional specification of a* **speech** *sample.*
Arousal	*excitation level.*
Basic emotions	*primary or fundamental* **emotions** *defined by various psychologist.*
Behavior model	*a-priory information about user's emotional state.*
Boundary prosody	*phrasing, accentuation or focus of attention, sentence moods.*
Chunk	**context**-*independent acoustic signal segment obtained within emotional segment detection.*
Circumplex	*cone-shaped model (3D) or wheel model (2D) of* **emotion** *representation.*
Clustering of emotions	*clustering of* **emotions** *to a binary (positive/negative)* **arousal** *and* **valence** *or 4 quadrants discrimination task.*
Companion Technology	*a user-centred dialogic man-machine-interaction technology, based on fundamental technical, informational, psychological and neurobiological concepts. Investigated by an ongoing research project, the Transregional Collaborative Research Centre SFB/TRR 62.*

Context — *information about phonetic* **transcription**, *word or sentence (to be understood as emotional context).*

Corpus — *dataset of* **speech** *samples and corresponding* **transcriptions** *and/or* **annotations.**

Dialog — **speech** *based interaction between human and machine.*

Domain — *limited set of textual information which can be used for language modeling.*

Dominance — *apparent strength of the person. [Grimm et al., 2007]*

Dynamic analysis — **emotion** *processing on frame level.*

Emotion — *short time user's reaction bound to a specific stimulus.*

Emotion category — *word identifier which specifies an emotional user's state.*

Emotion descriptor — *user's emotional state specification with* **emotion categories** *or numeric values in an* **emotion space.**

Emotion space — *two- or three-dimensional (e.g.* **valence-arousal-[dominance(potency)]***) space, where each* **emotion** *can be defined as a point with corresponding coordinates.*

Formant — *the spectral peaks of the* **speech** *spectrum.*

Frame — *segment of acoustic* **speech** *signal (e.g. 25 ms length for automatic* **speech** *recognition and utterance-level* **emotion** *classification).*

Fusion — *combination of several classification techniques.*

Geneva emotion wheel — *wheel with* 20 *spokes (*__emotion__ *families), each spoke is associated with a type of* **emotion** *(*10 *negative and* 10 *positive* **emotions***) arranged on pleasure-*__dominance__ *space.*

Grammar — *specification of a possible word sequence with a predefined word occurrence order.*

Human-machine interaction *structured and thematic domain-dependent interaction between a user and a system.*

Intelligent spoken dialog system *system which accommodates the speaker's* **emotions** *in a proper way.*

Intention *user's operational goal.*

Language model *list of words that can follow each word included in the* **vocabulary** *with associated discrete probability.*

Lexicon *phonetical* **transcriptions** *for all words included in the* **vocabulary.**

Low-level descriptors *acoustic features applied for* **static analysis**, *see Table 4.2 on page 83.*

Mapping of emotions *functional mapping of categorical* **emotions** *on* **valence-arousal** *dimensional space.*

Middle-level fusion *combination of two classification techniques, which used classification scores of the first classifier as an additional feature set of the second classifier.*

Modality *an information channel used for classification.*

Motivation *process that initiates a reason or an interest that causes a specific action or certain behavior.*

Multimodal *based on several information channels.*

Phoneme *smallest acoustic component of* **speech** *to form meaningful* **utterances.**

Plutchik's emotional wheel *conceptualization of the primary* **emotions** *in a color-wheel fashion – placing similar emotions close together and opposites* 180 *degrees apart, like additional colors [Plutchik, 2001].*

Potency	*individual's sense of power or control, for example "concentrated vs. relaxed attention", "dominance vs. "submissiveness".*
Robust	*stable enough to be implemented in real-life application.*
Speech	*acoustic signal produced by a speaker.*
Static analysis	**emotion** *processing on turn level with* **statistical functionals.**
Statistical functionals	*functions which project uni-variate time series onto a scalar feature independent of the length of the turn (e.g. mean, standard deviation, etc.).*
Transcription	*phonetic specification of a* **speech** *sample.*
Turn	*word or word sequence within completed speaker's command.*
Unimodal	*based on single information channel.*
Unit specific	**context** *dependent.*
User-behavior adaptive	*adaptive to the current user's emotional state.*
Utterance	*word or word sequence within completed speaker's command.*
Valence	*represents the value – positive or negative – of the user's* **emotion.**
Vocabulary	*list of words which can be recognized by the system.*
Vowel triangle	*represents the extremes of vowel's* **formant** *location in the F1/F2 space.*
Wizard of Oz experiment	*experiments which are based on subjects' illusion that they are interacting with a computer driven system, while a human operator simulates a computing system.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation and aim

Currently, automatic recognition of emotions from speech, mimics and other modalities has achieved growing interest within the human-machine interaction research community and spoken dialog system designers. Emotion recognition is a guiding star on the path to making a communication between humans and computers more friendly and cooperative. With robust emotion recognition, we will be able to model a user's behavior within interaction with a computer. At the same time, automatic assessment of an affective speech will simplify speech understanding and intention detection tasks.

The importance of human-behavior-based dialog strategies in human-machine interaction (HMI) lies in an existing limitations of automatic speech-recognition (ASR) technology. The current state-of-the-art ASR approaches still cannot deal with flexible, unrestricted user's language [Lee, 2007], [Benzeghiba et al., 2007]. Therefore, problems caused by a misunderstanding of a user who refuses to follow a predefined, and usually restricting, set of communicational rules seems to be inevitable.

It has been shown in [Bosch, 2003], that the "linguistic content" of spoken utterance goes beyond its "text" content. During human-to-human communication, the listener extracts important information (*semantic boundaries, accents, sentence mood, focus of attention, and emotional state* of the user [Niemann et al., 1998]) out of prosodic cues. Detecting and utilizing such cues as a part of the user-behavior state descriptors is one of the major challenges in the development of reliable human-machine interfaces. Knowledge of the user's emotional states can help to adjust system responses so that the user

of such a system can be more engaged and have a more effective interaction with the system [Gnjatović and Rösner, 2008b].

The speech-recognition task becomes more and more difficult, and enormous challenging problems on acoustic modeling arise. One of the challenges is the diverse prosodic characteristics of the spontaneous speech data. For example, different non-lexical events, intonation variability, a speaker mood change. Most ASR systems are designed not to be receptive to intonation, user's emotional state, and loudness variability. It has been shown that ASR performance depends on speaking style and level of formality [Weintraub et al., 1996]. *Adaptation techniques* can be used to increase performance of affective spontaneous speech recognition. By adapting an ASR model trained on neutral speech on a sparse amount of affective speech samples, we can provide so-called 'statistical similarity' of training and test material [Ijima et al., 2009].

Research by neuroscientists and psychologists showed that a user's emotional state is closely related to the decision-making process during the human-to-human communication [Damasio, 1994], within a human-machine interaction and thus, emotion plays an important role in the sensible human actions. Realizing the importance of emotions in a human communication and a decision-making process, it is desirable for an intelligent human-machine interface to accommodate the human emotions in a proper way.

1.1.1 Applications of automatic emotion recognition

Emotions perform an important function in human communication and interaction, allowing people to express themselves beyond the bounds of the verbal communication. The ability to understand human emotions within human-machine interaction is desirable in several applications:

- Expressive speech synthesis, for a new generation of HMI systems which can be used to increase the naturalness of the human-machine interaction.
- Emotion recognition (e.g., for early miscommunication and frustration detection in spoken dialog systems, such as commercial telephone-based dialog systems)
- Safety drive assistance, automatic recognition and control of emotions for in-car interfaces,
- Opinion mining and level of interest classification which automatically tracks customer's attitudes regarding a product across blog comments (Web 2.0),
- Affective monitoring for "lie detection" systems like polygraph, fear detection for surveillance purposes or anger detection for conflict situations detection,

- Character design and interaction control for games and virtual-reality scenarios,
- Social robots, such as guide robots engaging with visitors (e.g., MEXI a Robot with Emotions, Fujitsu Service Robot "ENON"),
- Support for people with disabilities, such as educational programs for people with autism
- Automatic movie genre classification or episodes indexing (comedy, action, drama and etc.)

1.1.2 Variety of modalities

Humans-to-human interaction is mainly based on vocal communication, but also facial mimics and body gesture language. Both are used to emphasize a certain part of the speech and display of emotions. An analysis of a gaze, a posture, gestures, facial expressions, an eye contact, face and lip movements can support a user-behavior modeling. Likewise, the speech signal may convey linguistic as well as paralinguistic information. It has been shown that linguistic properties can be used as an indicator of miscommunication situations [Nöth et al., 2004]. Furthermore, it has been shown that sentence mood in German can be indicated by prosody, lexical content, word order, and morphology [Batliner et al., 2003]. Besides prosodic variation, speakers indeed employ a number of different linguistic features to express their emotions.

There are some physiological responses that can be used for the recognition of the user's emotional state. These include blood pressure, blood volume pulse, respiration rate, heart rate, galvanic skin response, ECG, EMG and others. It was proved that emotional states can be recognized automatically from generic, and efficient physiological feature set design for each physiological signal [Hönig et al., 2009].

It is well-known that using automatic lipreading in combination with automatic speech recognition leads to higher speech-recognition performance. In addition, comparable to the silent visual cues from a system, facial expressions of a user may indicate communication problems even when the person is not speaking, for instance when the user becomes aware of a miscommunication situation during the system's prompts.

Fusion of the user's speech and visual cues analysis is becoming an ordinary feature in advanced multimodal spoken dialog systems. Combined audio low-level descriptors and video low-level descriptors time series analysis approach to an audiovisual behavior modeling proved to be highly promising [Schuller et al., 2007c]. The visual information may provide a useful source for detecting miscommunication or frustration, next to existing sources such as linguistic and prosodic cues. Automatic facial tracking could be beneficial for improving

human-to-machine interactions in that audiovisual events indicate problematic dialog events and allow the system to monitor the level of frustration of a user [Barkhuysen et al., 2005].

1.1.3 Technical problems in realization

The main problem of user emotional states classification within speech is a difficulty of data collection. In most cases, actors simulate emotions according to some certain scenario usually in a perfect acoustic condition. These materials are good for emotion-classifier developing and the most informative acoustic feature set selection in the context of an emotion-recognition task. But this acted data is not applicable for training robust models for spontaneous emotion recognition.

An alternative to the prototypical expressions of "pure" emotions is to use experiments which simulate human-computer conversations with a so-called Wizard of Oz (WOZ) scenario. A questionnaire study conducted after some WOZ experiments showed that speakers may first be slightly frustrated, then become really annoyed, and as they believe they are talking to a computer, they do not attempt to display their emotional state to their artificial communication partner at all. In most cases, emotional data collected during WOZ is less emotionally intensive in comparison to acted material. As a result, in most publications related to emotional speech processing, performance of emotion classification on acted data outperforms evaluation results of spontaneous emotions. Acoustically based emotion classification works quite well for prompted affective speech [Schuller et al., 2009], but is not sufficient for the more realistic spontaneous emotions which occur in real systems or WOZ scenarios. It was demonstrated that spontaneous emotion-classification performance increases if we add more knowledge sources, for instance, syntactic-morphological parts of speech (POS) information. One can model and find miscommunication indicators better if one incorporates higher linguistic-pragmatic information, for instance, by recognizing repetitions [Batliner et al., 2003].

Another significant problem for the analysis of spontaneous emotional data is emotional chunks delimitation. The problem lies in defining the "reference" of a study; that is, determining which part of a user's utterance should be marked as emotional and which as neutral. In most cases, within human-machine interaction speakers do not display single, pure, emotions in their full intensity within one utterance. At the same time, correct detection of pure saturated anger will certainly be too late for the spoken dialog system to react in a way so as to fix a miscommunication problem. The main issue is not a detection of overflow anger, but classification of all forms of slight or

medium irritation indicating a critical phase in the dialog that may become real saturated anger if a wrong dialog strategy is applied.

There are ongoing debates in the affective-speech-processing community concerning how many emotion categories exist and which of them are applicable for intelligent spoken dialog systems, how to submit long-term (utterance, sentence, dialog act) properties, for example, moods with short-term affective events such as full-blown emotions. Research aimed at recognizing emotion requires databases that contain as many as possible of the indications by which a given emotion can be expressed. Most of the publications on acoustic-based emotion processing is underpinned by "datasets" rather than "databases". They are relatively small-scale collections of speech samples, usually established to examine a single case issue, and not publicly available [Douglas-Cowie et al., 2003].

One of the problems of automatic emotion-classification research is a non-standardized annotation methodology. Emotions annotation methodology needs to be standardized. Afterwards the speech-processing community can start a joint emotional speech data collection and annotation that solves the problem of a sparse amount of well-annotated affective speech data.

1.1.4 Research goals

The primary aim of this research is to present new affective speech-processing methods and their possible application for user-friendly spoken dialog systems. Recognition of prosodic cues such as emotional state and stress level of the speaker may be detected and used for an affective-behavior-adaptive dialog strategy.

An overview of existing affective-speech-processing methods is presented in this thesis. The advantages and disadvantages of different speech-based emotion-classification methods are discussed. Also, new methods of acoustic emotion classification and affective-speech-adapted ASR models are described. Robustness and usability of the above-mentioned methods have been proved by evaluations on well-known emotional speech corpora. Results of evaluations are presented in our publications and this thesis.

1.2 Practical implementation of the research

Within well-known projects like VERBMOBIL and SMARTKOM [Herzog et al., 2004] a framework for building integrated natural-language understanding with multimodal dialog systems was created. Both projects include the prosody module for boundary prosody analysis, sentence mood and phrase

accent classification. The prosody module integrated in the SMARTKOM demonstrator is based on the Verbmobil prosody module [Batliner et al., 2000a]. In contrast to the Verbmobil version, few major changes have been made concerning both implementation and classification models. The most noticeable is a user state classifier. All existing classification models for the recognition of prominent words, phrase boundaries, and questions have been retrained on the actual SMARTKOM Wizard of Oz data [Zeißler et al., 2006].

My research addresses aspects of design and implementation of user-behavior models in dialog systems for frustration detection and user-intention recognition, aimed to provide naturalness of human-machine interaction. For real-life evaluation, acoustic emotion-classification methods, robust affective automatic speech-recognition (ASR) methods, and user emotion correlated dialog management, a multimodal human-machine interaction system with integrated user-behavior model has been created within the project "Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems" (NIMITEK) [Wendemuth et al., 2008]. Currently the NIMITEK demonstration system provides a technical demonstrator to study user-behavior-modeling principles in a dedicated task, namely solving the game "Towers of Hanoi". The user-behavior model integrated in the NIMITEK demonstrator based on emotion-classification methods will be described in this thesis. Within a usability test [Vlasenko and Wendemuth, 2009a], we find that our system with user-behavior-adaptive dialog strategy provides more cooperative human-to-machine interaction and reduces interaction time required to complete the game.

1.3 Thesis structure

The thesis is organized as follows.

Chapter 2 presents the fundamental aspects of human-machine interaction including automatic spoken dialog systems, natural speech characteristics (boundary prosody, emotional prosody), user-behavior modeling during communication, affective speech collection and processing. Then, clustering of emotions and an adequate annotation strategy are described. Various evaluation strategies and recognition rate measures are discussed at the end of the chapter.

Chapter 3 reviews the fundamental issues of automatic speech recognition, namely, feature extraction, acoustic modeling with HMMs, maximum likelihood (ML) training, language modeling and search algorithms within recognition. Also, sub-word units selection and adaptation on affective speech samples are described.

Chapter 4 addresses various speech-based emotion-recognition techniques. An overview of existing emotion-classification methods, acoustic feature sets specification concepts and emotion descriptors characteristics are presented first. This chapter presents dynamic and static emotion-recognition methods with corresponding acoustic feature sets and possible optimization strategies. Our developed combined emotion-classification methods are also discussed in detail. Finally, context-dependent and context-independent models are evaluated.

Chapter 5 presents experimental results for affective speech recognition and speaker's emotional-state classification. Evaluation results for neutral and affective-speech-recognition experiments are presented first. Also, this chapter presents evaluation results of various emotion-classification methods described earlier in Chapter 4. Finally, evaluation results for our emotion-classification techniques within the INTERSPEECH 2009 Emotion Challenge [Schuller et al., 2009c] and cross-corpus acoustic emotion recognition are presented.

Chapter 6 describes a prototype of the user-friendly spoken dialog system integrated into a NIMITEK demonstrator. The system dynamically selects a dialog strategy according to the current user's emotional state. This system incorporate the findings described in previous chapters into a prototype dialog system especially developed by the author and colleagues to demonstrate emotional user state adaptation. In this chapter we describe the data collection strategy within the NIMITEK Wizard of Oz experiments, and the structure of the conventional and user-behavior-adaptive spoken dialog systems. Finally we discuss the results of an interactive usability test.

Chapter 7 addresses the conclusions and direction of future research.

Chapter 2

State of the art

Contents

2.1 Introduction

In this chapter we provide an overview of the several topics of interest in spoken dialog systems and human-machine interaction. We also provide a brief description of spontaneous speech characteristics, namely, boundary and emotional prosody. We also present an introduction to emotion theory, describe different emotion categorization approaches and survey existing sources of emotional speech. Finally, we describe different types of evaluation strategies and recognition-rate measures.

2.2 Human-machine interaction

Currently we live in the Age of Information. Information collection, searching, and structuring are usual activities in the everyday life of modern humans. We use electronic devices (computers, digital cameras, smartphones, mobile

phones etc.) for communication, multimedia data collection, entertainment, educational purposes, information access (Internet web resources, travel assistance, dictionary etc.), online shopping and other services.

Many existing human-machine interfaces within multimedia systems are far from being user-friendly, and only a few are based on a human-centered approach [Jaimes and Sebe, 2007]. Nowadays, computers are quickly becoming integrated into everyday devices, which implies that effective natural human-machine interaction is becoming critical. To make human-machine interaction more cooperative and productive, intelligent human-centered communication features have to be integrated into machine interfaces. The success of human-centered human-machine interfaces has to take into account two joint aspects [Jaimes et al., 2006]:

- *the way humans interact with such systems (speech, prosodic characteristics, mimic, gestures and etc) to express emotions, mood, attitude, and attention,*
- *the human factors that belong to multimedia data (human subjectivity, levels of interpretation).*

Whilst developing our intelligent human-centered human-machine interface we took into account the fact that human-to-human communication is usually socially situated and that humans use emotion to enhance their communication. However, since emotions are often expressed within communication, processing them is an important task for intelligent HMI. Our main aim is the creation of an HMI system that can "feel" the affective states of the human and is capable of adapting and adequately responding to these affective states.

2.2.1 Spoken dialog systems

Systems, in which human users use verbal communication to achieve a goal, are called spoken dialog systems (SDS). Such systems are some of the few realized examples of real-time, goal-oriented humans-to-computer interaction or humans-to-human communication with participation of computers (real-time spoken language translation systems, see VERBMOBIL). Commercial automatic spoken dialog systems are quite popular in English-speaking countries.

Still commercial automatic spoken dialog systems have just started to subjugate the German market. Some large projects like VERBMOBIL and SMARTKOM [Herzog et al., 2004] created a framework for building integrated natural-language understanding with multimodal dialog systems. Thematic domain restricted automatic dialog systems were created by Sympalog. *Sixt switchboard, Betri, Filtips* represent Sympalog's conversational dialogue

technology and today's standard IVR (Interactive Voice Response) technology [Nöth et al., 2004]. Unfortunately these dialog systems do not take into account most of the ideas of human-centered/human-initiative concepts. They are still task- and machine-centered.

VERBMOBIL is a speaker-independent speech-to-speech translation system. It provides users with a speech-to-speech translation service in mobile situations with simultaneous dialog interpretation services on restricted topics. The system processes dialogs in three thematic domains, namely appointment scheduling, travel planning, and remote PC maintenance, and it provides context-sensitive translations between three languages (German, English, Japanese) [Batliner et al., 2000a].

SMARTKOM is a mixed-initiative dialog system that provides full symmetric multimodality by combining speech, gesture, and facial expressions for both user input and system output [Reithinger et al., 2003]. The system aims to provide an anthropomorphic and affective user interface through its personification of an embodied conversational agent. The interaction metaphor is based on the so-called situated, delegation-oriented dialog paradigm [Zeißler et al., 2006].

The *Sixt switchboard* application handles all incoming calls (approx. 1000 per day) to the Sixt AG's central telephone number. 90% of the received calls by Sixt AG are redirected automatically to the correct person, the rest of the calls are handed over to a human operator. The system's knowledge database consists of more than 1000 employee names. *Berti*, which is a football Bundesliga information system, is now commercially operated by a large German media company on a pay-per-call basis. *Filmtips*, which is a movie information system, is operated by a cinema company in the Nuremberg region [Nöth et al., 2004].

2.2.2 Artificial communication advantages and disadvantages

Real-life and artificial communication are currently far away from being comparable. A natural communication system includes a natural verbal language with the huge prosodic variability combined with a nonverbal body and gesture language. On the one hand, a boundary prosody indicates a focus of attention, a sentence structure, a speaker intention. On the other hand, an emotional prosody shows a level of interest, a mood and a possible frustration during the human-to-human interaction. Artificial communications systems are those deliberately invented, usually to serve specific functions or tasks, such as booking tickets, controlling bank accounts, or searching for some information during human-machine interaction. User-behavior modeling by emotion classifica-

tion within the human-machine interaction received a great deal of attention during the last few years in the spoken dialog developers community.

It is highly desirable in most HMI applications such as computer-aided tutoring and learning, that the response of the computer takes into account the emotional or cognitive state of the human user. Emotions are displayed by mimics, body movements, speech, linguistic and paralinguistic means. More and more research in HMI confirmed that emotional skills modeling is an important part of the so-called intelligent system. Spoken dialog systems today can recognize much of what is said, and to some extent, who said it. Still, they are not able to process the affective channel of information [Jaimes and Sebe, 2007]. Intelligent systems with affective communication features consider how emotions can be classified and expressed during human-machine interaction. Three key points have to be applied during developing systems that process affective information: embodiment (experiencing physical reality), dynamics (mapping experience and emotional state with its label), and adaptive interaction (conveying emotive response, responding to a recognized emotional state) [Bianchi-Berthouze and Lisetti, 2002].

Nowadays, one takes a human-to-human interaction scenario, and replaces one of the humans with an automated dialog system, then the affective communication will disappear. It happens not because people stop communicating affectively – e.g., a person expresses anger at dialog systems during miscommunication situations. The problem arises because the human-machine interface has no ability to detect when a human is stressed, frustrated, pleased, interested, or bored. A person ignoring the non-verbal elements in human-to-human communication would be considered impolite or unintelligent. Detection and classification of emotions within artificial communication are key components of the intelligent system.

Research is therefore needed for new methods to communicate affectively through automated system controlled environments. Up-to-date spoken-dialog-system-driven communication almost always has less affective bandwidth than natural human-to-human machine interaction. The appearance of affective wearable dialog systems could change the nature and efficiency of human-machine interaction.

2.3 Prosodic characteristics of spontaneous speech

Linguists defined prosody as rhythm, stress and intonation of speech. Prosody reflects the following features of the speaker or the utterance: the emotional state; whether an utterance is a statement, a question, or a command; whether

the speaker is being cooperative or non-cooperative; the use of emphasis, contrast, and focus; or other elements of language such as paralinguistic events that may not be encoded by grammar or choice of vocabulary. In terms of an acoustic theory, the prosody of speech involves variation in syllable length, loudness, pitch, formant frequencies, poses and word length within the speech signal.

Prosodic information is encapsulated within vocalized phoneme, syllables, words, phrases, and whole turns of a speaker. To these units we ascribe perceived properties such as pitch, loudness, speaking rate, words and pause duration, voice quality, rhythm, etc. In human-to-human communication, the listener extracts multiple information from prosodic cues. Due to this fact, we can define certain functions of the prosody phenomena. The prosodic functions are the marking of boundaries, accents, the sentence mood, an intonation and the speaker's emotional state [Batliner and Nöth, 2003].

2.3.1 Boundary prosody

An application of prosody analysis is quite popular in automatic speech processing and dialog understanding. For example, many studies showed that prosodic information may influence listeners' analysis of an ambiguous phrase [Clifton et al., 2002]. In the real-life applications, spoken dialog system designers try to combine word hypothesis graphs (WHG) with the prosody analysis for accentuation or prosodic boundaries recognition.

The prosodic units can be very short – e.g. phoneme-level – or they can constitute a whole utterance. Dialog units are longer than those of semantics. The first prosody feature is *phrasing*, i.e., prosodic boundaries that reflect syntactic boundaries which, in turn, reflect dialog acts (DA) boundaries. As a second feature comes *accentuation or focus of attention*, the most important information in a semantic unit, e.g., in a sentence (focus). The third prosody feature is an ability to disambiguate between different *sentence moods/modalities*. For example, prosody can be used to decide whether a sentence is a statement or a question [Nöth et al., 2002].

In the case of miscommunication detection within human-machine interaction, Batliner et al. found that some boundary prosodic features indicate trouble in communication [Batliner et al., 2003]. These indicators are conducted in the following prosodic characteristics:

- *pause at phrases;*
- *strong articulation;*
- *strong emphasis;*
- *pause at words;*
- *contrastive accent;*

- *pause at syllable;*
- *lengthening of syllable;*
- *hyperarticulation;*
- *laughter/sighing.*

An evaluation of the SMARTKOM [Zeißler et al., 2006] prosody module, which is based on the Verbmobil prosody module [Batliner et al., 2000a], shows that boundary prosody analysis may provide higher performance of Automatic Speech Understanding (ASU).

2.3.2 Emotional prosody

While listening to speech, we rely on a variety of congruent prosodic and verbal-semantic cues upon which to base our interaction inference as to the communicative intention of others. To interpret the meaning of the speech, the way something is said may be as important as a linguistic content.

The paralinguistic decoding is an essential issue in the emotional prosody analysis. The emotion within speech may manifest itself on the semantic and acoustic levels. A variety of acoustic features were also explored in the context of speech-based emotion classification and emotional speech synthesis. These acoustic features are as follows:

- *pitch-related features;*
- *voice level features: signal amplitude, energy;*
- *formant frequencies;*
- *timing features: phrase, word, phoneme, and feature boundaries;*
- *voice-quality parameters;*
- *spectral features;*
- *articulation parameters.*

Emotion	**Speech Rate**	**Pitch Average**	**Pitch Range**	**Intensity**	**Voice Quality**
Anger	slightly faster	very much higher	much wider	higher	breathy
Joy	faster or slower	much higher	much wider	higher	blaring
Sadness	slightly slower	slightly lower	slightly narrower	lower	resonant
Fear	much faster	very much higher	much wider	normal	irregular
Disgust	very much slower	very much lower	slightly wider	lower	grumbled

Table 2.1: *Summary of human vocal characteristics variations of affective speech compared to neutral speech [Murray and Arnott, 1993]*

Murray and Arnott [Murray and Arnott, 1993] described emotional voice characteristics for Ekman (see 2.5.2) basic emotions. Table 2.1 describes mostly qualitative characteristics associated with the following fundamental emotions. These specifications are based on a comparison of the affective voice to the neutral voice characteristics.

Within our research we find out that only by using spectral features (Mel-frequency Cepstral coefficients (MFCC)) analysis we can reach a good performance of emotion recognition for acted and spontaneous emotions samples [Schuller et al., 2009, Vlasenko and Wendemuth, 2009b, Hübner et al., 2010]. In the case of spontaneous emotions, we have to extend our acoustic features set and use a multi-level processing paradigm to reach comparable classification performance on the acted data. Also a combination of acoustic, linguistic and conversational analysis yielded better results on spontaneous emotions classification than the pure acoustic analysis [Schuller et al., 2005b].

2.3.3 Interaction

In order to make spoken dialog systems more intelligent and user-friendly we have to combine automatic speech recognition with a reliable language model, boundary and emotional prosody analyzers, and a language-understanding module. In such a way they will be able to detect and classify user intentions. The basic structure of an intelligent SDS is shown in Figure 2.1.

The first stage of an intelligent spoken dialog system is to recognize the speech signal and provide a word hypotheses graph (WHG) and corresponding

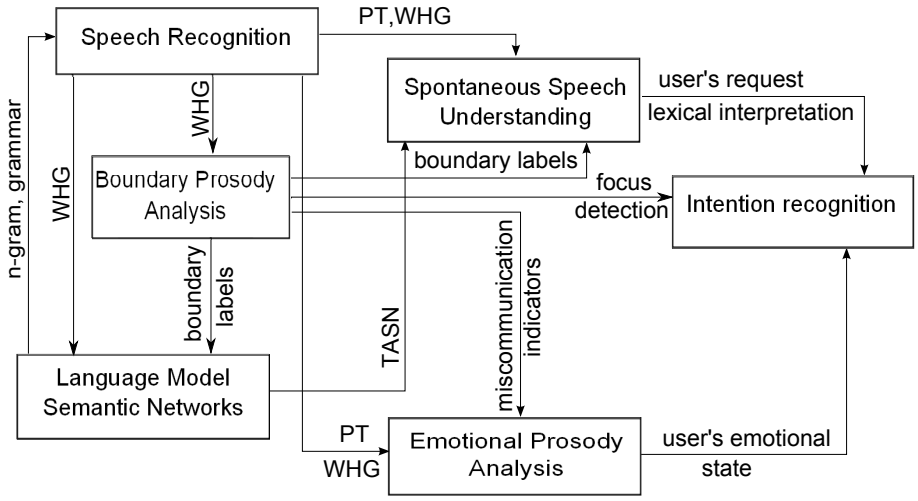


Figure 2.1: *General structure and modules interaction of an intelligent spoken dialog system*

phonetic transcription (PT). This process is known as automatic speech recognition. To attain an acceptable performance of speech recognition, the module requires language models, for example, *n-grams*. The WHG is directed to the boundary prosody analysis module, which later generates the boundary labels, detects the focus of attention and miscommunication indicators. Semantic network modules based on the boundary labels and the WHG generate textual associations semantic networks (TASN). Taking into account TASN, PT, WHG and boundary labels, the spontaneous-speech-understanding module estimates the user's request's lexical interpretation. The emotional prosody analysis module based on speech signals, PT, WHG and miscommunication indicators classifies the current user's emotional state. An intention recognition module takes into consideration the detected focus of attention, the user's emotional state and lexical interpretation of the user's request, and provides user's intention classification.

2.4 Emotion theory

In summary an emotion is a transitory, valenced experience that is felt with some intensity as happening to the self, generated in part by a cognitive appraisal of situations, and accompanied by both learned and innate physical responses. Through emotion, people communicate their internal states and intentions to others. Emotion often disrupts thought and behavior, but it also triggers and guides cognitions and organizes, motivates, and sustains behavior and social relations (adopted from [Bernstein et al., 1997]).

In recent years considerable research was carried out, both theoretical and empirical, on the perception and production of affective speech. Most of this research effort is now being made in a field called "affective computing" [Picard, 1997]. The main goal in affective computing is to design automatic speech-recognition and text-to-speech algorithms that understand and react to the human emotions.

Klaus Scherer distinguished the following *affective* phenomena: *emotions, feelings, moods and attitudes* [Scherer, 2005]. Also, he suggested that "feelings integrate the central representation of appraisal-driven response organization in emotion" [Scherer, 2004]. The affective states caused by a salient attitude can be labeled using terms such as desiring, respecting, hating, and loving. In most cases, attitude is a long-term affective event which can make the occurrence of a short-term emotion episode more likely. For example, people in love usually express positive emotions more often. Generally, mood is considered a diffuse affect state, characterized by subjective feelings that affect the behavior of a person. Moods are generally low-intensity affect states which can

last for days, weeks, or months. Within our research we use the term affect in its *short-term* nature, namely, emotional state. Also, from our point of view, affective phenomena like feelings, moods and attitudes are the usual cases for human-to-human communication; in the case of human-machine interaction most of these phenomena do not occur. As a consequence, terms such as "affective computing" [Picard, 1997] and "affective speech recognition" are quite popular in the speech-processing community and human-machine interaction research groups. In this thesis, *"affective"* will in general refer to any non-neutral short-term expression.

A uniform definition of emotion in psychology is very controversial. Basically, emotions describe subjective sensations of shorter periods which are related to certain events, persons or objects. The word "emotion" comes from Latin and means to move or to stir up. Generally, psychologists use the word "emotion" to refer to the show of feelings that are produced when important things happen to us.

Four different theoretical approaches to the origins and nature of human emotions have primarily crystallized:

- **Darwinian approach**: According to the Darwinian perspective [Darwin, 1872] emotions are a result of general human evolution. They have an essential importance for the species survival. As a consequence, certain behaviors are directly linked to the associated emotional feelings. Universal facial expression, infants, and basic emotions are evidence supporting this theory.
- **Jamesian approach**: This approach is well grounded owing to the work by James [James, 1884]. James believed that the human perception of feelings are in response to events. Thus, an emotion appears through the stimulation of sensory organs by an object. The self-perception takes place through afferent impulses leading to the brain until they reach the cortex. As a consequence, internal organs and muscles are stimulated by efferent impulses. With the return in the form of re-afferent impulses from the organs and muscles to the cerebral cortex, eventually it appears in the described perception of physical change in the form of an emotion. An emotional feeling is possible only in combination with a succeeding physical response. Emotion is inferred or constructed from instinctive peripheral physiological responses. The following is evidence in support of James:
  - patterns of autonomic changes vary with different emotional states;
  - people reliving emotional experiences show different patterns of autonomic activity;
  - spinal cord injuries reduce peripheral responses – less intense emotion

(following Hohmann [Hohmann, 1996]).

- **Cognitive approach**: This theory is similar to the Jamesian theory as people label emotions using perceptions of their own somatic activity. But labeling is a cognitive process that reflects the person's beliefs about a situation. If people believe they have a reason to be angry they will perceive their bodily symptoms as anger. The representatives of this theory Schachter [Schachter and Signer, 1962] and Arnold [Arnold, 1960] assumed that emotions are the cause of body reactions to certain circumstances and that they are traceable.
- **Social constructivist approach**: Averill [Averill, 1980] and Harré [Harré, 1986] argued that feelings reflect the result of learned social rules of behavior. The decisive factor is the underlying culture, because it implies significantly the assessment of the circumstances which lead to an emotion. Hence, the triggers for anger differ inter-culturally and even interpersonally. Following this model, the cultural context plays an important role for the assessment of emotions. The social-constructivist approach is one of the youngest and most controversial psychological theories about human feelings. It shows that some syndromes in different cultures can be detected as unambiguous emotions, while in other circles this can be only conditionally true. This approach is in conflict with the others, but within it the emotions are considered a product of evolution.

After all, we need to establish that although these theories in parts can be accumulated, not one of them was examined correctly. Besides, a lot of efforts were made to combine them. Within our research we applied the basic ideas of Jamesian, cognitive and social constructivist approaches.

2.5 Emotion categorization

An annotation of emotional episodes within affective speech is a non-trivial task. An essential problem for the analysis of spontaneous emotional speech is to determine what an emotional episode is, where it starts and where it ends. Afterwards we have to provide a reference for the following episodes. Quite often, several emotions can be present at the same episode. There are two possible emotional annotation approaches based on multi-dimensional representation and classical emotion categories.

2.5.1 Multi-dimensional representation

There are few ways of representing emotions in a multi-dimensional emotion space. Emotions can be distinguished by the numeric values in two- or three-dimensional *valence-arousal-(potency or dominance)* spaces [Wundt, 1897], [Kehrein, 2002], [Grimm et al., 2007] or by meaning of their basic entities within *circumplex* models [Plutchik, 2001], [Scherer, 2005]. This chapter will describe in detail the most popular existing dimensional *valence-arousal-(potency or dominance)* space and *circumplex* models.

The first multi-dimensional representation of emotions was proposed by the German psychologist Wilhelm Wundt [Wundt, 1897]. He proposed to describe an emotional experience in terms of three dimensions: valence, arousal, and potency. These dimensions can be interpreted as three orthogonal axes. Each emotion can be characterized be the three numerical values which correspond to the coordinates within the *valence-arousal-potency* space. Valence

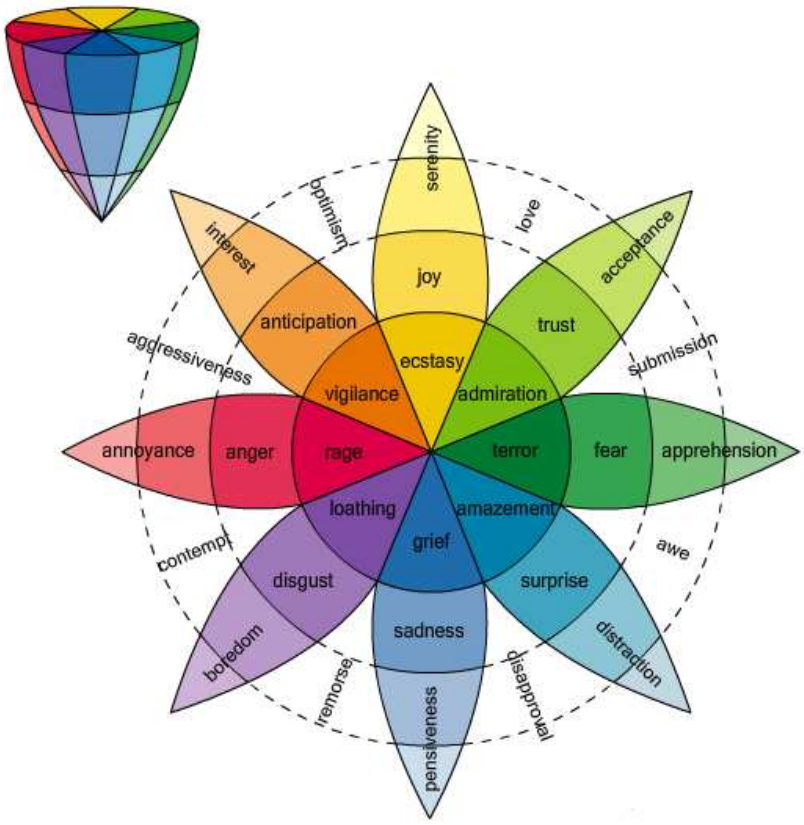


Figure 2.2: *Plutchiks's two- and three-dimensional circumplex emotional wheel model describes the relations among emotional classes. Adopted from [Plutchik, 2001]*

represents the value – positive or negative – of the user's emotion. Arousal/activation represents the user's degree of excitation – from high to low, like "active vs. passive", "high vs. low excitation". "Potency" refers to the individual's sense of power or control, for example "concentrated vs. relaxed attention", "dominance vs. submissiveness".

The multi-dimensional description benefits from a higher-level of generality. It provides a possibility for describing the intensity of emotions. In the case of mixed emotions within the same semantic unit (dialog act (DA), sentence, utterance, word), which is quite often the case in spontaneous affective speech, the emotion space concept allows for a more adequate description of these affective samples. Nowadays, annotation of emotional events within speech has led to the multi-dimensional emotion descriptor becoming more and more popular. Kehrein [Kehrein, 2002] and Grimm et al. [Grimm et al., 2007] proposed the use of the following dimensions: *appraisal* (or *valence, evaluation*), *activation* (or *arousal, excitation*) and *dominance* (or *power*).

Another quite popular multi-dimensional representation of emotion is based on the so-called *circumplex* model. In 1980, Robert Plutchik created a wheel of emotions which consisted of 8 basic emotions: *joy, acceptance, fear, surprise, sadness, disgust, anger,* and *anticipation.* Plutchik found that the

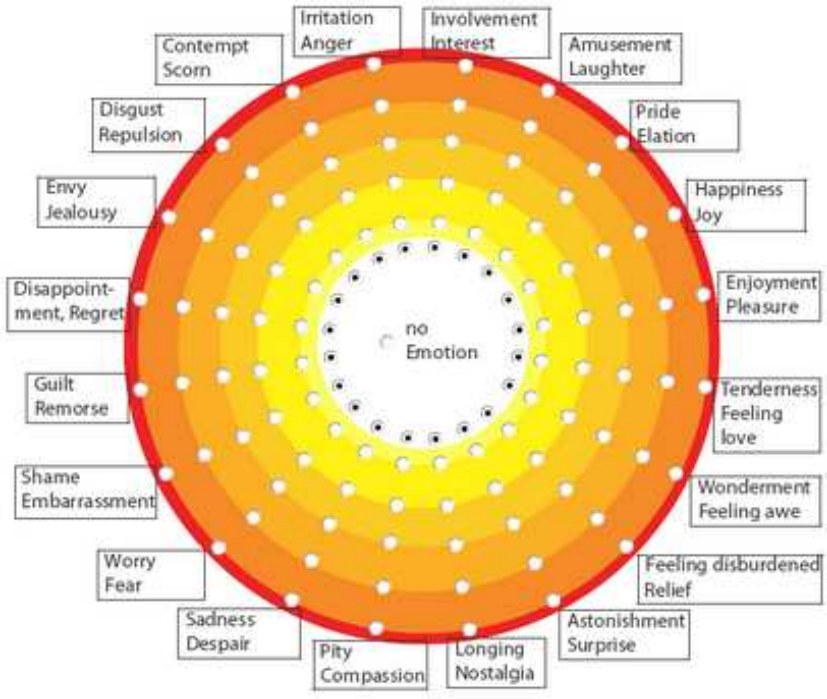


Figure 2.3: *Sample of the adapted Geneva emotion wheel applied for annotation purposes within the SEAT project. Adopted from [GEW, 2008]*

primary emotions can be conceptualized in a color-wheel fashion – placing similar emotions close together and opposites 180 degrees apart, like additional colors. This so-called *circumplex* model can be used as a tool for representation of relation and nature of emotional categories. Plutchik extended the *circumplex* model into a third dimension, modifying the intensity of emotions (see, different color intensity in Figure 2.2), so that the complex so-called structural model of emotions is shaped like a cone.

An alternative circular representation of emotions appears nowadays, see Figure 2.3. This adapted Geneva emotion wheel (GEW) was applied for the digital questionnaire within the SEAT (http://www.wearable.ethz.ch/research/groups/context/seat/) project by the ETH Zurich research group [GEW, 2008]. The GEW was developed in 2005 by Klaus Scherer. The GEW is a wheel with 20 spokes. Each spoke is associated with a type of emotion (10 negative and 10 positive emotions). The spokes of the wheel are made up of five labels which allow the annotator to choose the intensity for which they felt that selected emotion.

2.5.2 Classical emotion categories

In English there is an enormous amount of emotion words, some of them tend to fall into families based on similarity and some can be classified as opposites [Plutchik, 2001].

How many emotions are present in human-to-human communication? The

Author	Basic Emotions	Basis
McDougall (1926)	anger, disgust, elation, fear, subjection, tender-emotion, wonder	relation to instincts
Arnold (1960)	anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness	relation to action tendencies
Plutchik (1980)	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	relation to adaptive biological processes
Ekman, Friesen& Ellsworth (1982)	anger, disgust, fear, joy, sadness, surprise	universal facial expressions
Tomkins (1984)	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	distinctive set of bodily and facial reactions
Oatley& Johnson-Laird (1987)	anger, disgust, anxiety, happiness, sadness	do not require propositional content

Table 2.2: *Basic emotions sets, presented by different emotion psychology researches [Ortony and Turner, 1990]*

proposers of "classical" discrete emotion theories, inspired by Darwin, have suggested from 3 to 14 of *basic emotions*. Those emotions are also called *primary* or *fundamental*. A wide range of research on identification of basic emotions [Ortony and Turner, 1990] was presented to the emotion research community, see Table 2.2.

The discrepancy of opinion about the quantity of primary emotions is matched by the divergence of opinion about their identity. Some of the lists of basic emotions include categories that are not included in other lists. Only Arnold included *courage*, Plutchik included *acceptance* and *anticipation*, also McDougall proposed that *subjection* and *"tender-emotion"* are fundamental emotions. Still, most of the lists include *anger, disgust, fear, joy, sadness,* and *surprise* categories. Currently there is no standard basic emotions list acknowledged by all emotion psychology researchers. Still, all of these basic emotions are dialing with *"full-blown"* [Scherer, 1999] emotions, in contrast to low emotional saturation events within real-life communication. In this thesis we do not specify our own basic emotions set. Within our evaluations presented in Chapter 5 we use different set of emotions presented in public available emotional corpora.

2.6 Emotional speech data

Collecting and annotating emotional speech corpora is quite a difficult and expensive task. As a result, we decided to train and test our affective-speech-processing models on selected well-known emotional corpora. The chosen set of emotional corpora covers a broad variety of models reaching from acted (DES, EMO-DB) over induced (ABC, eNTERFACE) to natural emotion (AVIC, SmartKom, SUSAS, VAM) ranging from strictly limited textual content (DES, EMO-DB, SUSAS) over more variation (eNTERFACE) to full variance (ABC, AVIC, SAL, SmartKom, VAM). Further human-to-human (AVIC, VAM) as well as human-to-computer (SAL, SmartKom) interaction are contained. References for the earlier-mentioned databases will be given in section 2.6.2. Three languages (English, German, and Danish) are comprised. However, these languages belong to the same family of Germanic languages. The speaker's ages and backgrounds vary strongly, and so do of course microphones used, room acoustics, and coding (e.g., sampling rate reaching from 8 kHz to 44.1 kHz) as well as the annotators.

2.6.1 Data collection

Our main goal is a sufficient modeling of the spontaneous speech of common human beings in real-life human-computer interaction. Extracting data in real-life scenarios, usually, faces two main problems: Firstly, it is quite difficult to control and record such real-life conditions because of ethical restrictions and due to the point that automatic dialog systems are quite rare in everyday human life. Secondly, if we change the thematic domain of our dialog system, this can influence the linguistic and emotional behavior of the user.

To simulate a real-life situation we can use the Wizard of Oz scenario. In such a scenario, subjects believe they are interacting with a real automated system while the system's interaction interface is manipulated by a human 'wizard'. For such kind of simulation, we need 'naive' users. But still, we do not know the range of the user's emotional behavior variation in a real-life scenario. Also, human 'wizards' usually are not able to predict all possible miscommunication situations in real-life conditions which can provoke frustration and/or affective user's behavior. As a result, collected data does not cover all possible situations, where a dialog strategy can be implemented which is adaptive to the user's behavior.

2.6.2 Affective speech corpora

One of the major needs of the community – perhaps even more than in many related pattern recognition tasks – is the constant need for datasets [Douglas-Cowie et al., 2003], [Ververidis and Kotropoulos, 2003]. In the late 1990s, the early days of emotion recognition, there were only a few datasets available, which were small (500 turns) with few subjects (10), uni-modal, recorded in studio noise conditions, and acted. Furthermore, the spoken content was mostly predefined (DES [Engbert and Hansen, 1996], Berlin Emotional Speech-Database [Burkhardt et al., 2005], SUSAS [Hansen and Bou-Ghazale, 1997]). These were seldom made public and few annotators – if any at all – usually labeled exclusively the perceived emotion. Additionally, these were partly not intended for analysis, but for quality measurement of synthesis (e.g., DES, Berlin Emotional Speech-Database). However, any data is better than none. Today we are happy to see more diverse emotions covered, more elicited or even spontaneous sets of many speakers, larger amounts of instances (5k -10k) of more subjects (up to more than 100), multimodal data that is annotated by more labelers (4 (AVIC [Schuller et al., 2009b]) - 17 (VAM [Grimm et al., 2008])), and that is made publicly available. Thereby it lies in the nature of collecting acted data that equal distribution among classes is easily obtainable. In more spontaneous sets this is not given, which

Corpus	Content	# Emotion							# Arousal -	# Arousal +	# Valence -	# Valence +	# All	hh:mm	# Sub	Type	Freq [kHz]
ABC	German	agr	che	int	ner	neu	tir	-	104	326	213	217	431	01:15	4 m	acted	16
	fixed	95	105	33	93	79	25								4 f	stud	
AVIC	English	bor	neu	joy	-	-	-	-	553	2449	553	2449	3002	01:47	11 m	spon	44.1
	variable	553	2279	170											10 f	norm	
DES	Danish	ang	hap	neu	sad	sur	-	-	169	250	169	250	419	00:28	2 m	acted	20
	fixed	85	86	85	84	84									2 f	norm	
EMO-	German	ang	bor	dis	fea	hap	neu	sad	248	246	352	142	494	00:22	5 m	acted	16
DB	fixed	127	79	38	55	64	78	53							5 f	stud	
eNTER-	English	ang	dis	fea	hap	sad	sur	-	425	852	855	422	1277	01:00	34 m	acted	16
FACE	fixed	215	215	215	207	210	215								8 f	norm	
SAL	English	q1	q2	q3	q4	-	-	-	884	808	917	779	1692	01:41	2 m	spon	16
	variable	459	320	564	349										2 f	norm	
Smart-	German	ang	hel	joy	neu	pon	sur	uni	3088	735	381	3442	3823	07:08	32 m	spont	16
Kom	variable	220	161	284	2179	643	70	266							47 f	noisy	
SUSAS	English	hst	mst	neu	scr	-	-	-	701	2892	1616	1977	3593	01:01	4 m	mixed	8
	fixed	1202	1276	701	414										3 f	noisy	
VAM	German	q1	q2	q3	q4	-	-	-	501	445	875	71	946	00:47	15 m	spon	16
	variable	21	50	451	424										32 f	norm	

Table 2.3: *Overview of the selected emotion corpora*

Content: language, fixed/variable (spoken text). Number of turns per emotion category (# Emotion), binary arousal/valence, and overall number of turns (All). hh:mm : total duration. Number of subjects (Sub), number of female (f) and male (m) subjects. Type of material (acted/natural/mixed) and recording conditions (studio/normal/noisy) (Type). Freq [kHz]: discretization frequency. Abbreviations: agr - aggressive, ang - angry, bor - boredom, che - cheerful, dis - disgust, hap - happy, hel - helplessness, hst - high stress, int - intoxicated, joy - joyful, mst - medium stress, ner - nervous, neu - neutral, pon - pondering, q1-q4 - quadrants in the arousal-valance plane, sad - sadness, sur - surprise, tir - tired, uni - unidentifiable

forces one to either balance data in the training or to shift from reporting of simple recognition rates to F-measures or unweighted recall values, best per class (e.g., FAU AIBO [Batliner et al., 2008], and the AVIC databases). However, some acted and elicited datasets with pre-defined content are still seen (e.g., eNTERFACE [Martin et al., 2006]), yet these also follow the trend of more instances and speakers. The positive fact is, that transcription is becoming richer: additional annotation of spoken content and non-linguistic interjections (e.g., FAU AIBO, AVIC databases), multiple annotator tracks (e.g., VAM corpus), or even manually corrected pitch contours (FAU AIBO database) and additional audio tracks in different recordings (e.g., close talk and room microphone), syllable boundaries and manual syllable labeling (e.g., EMO-DB database), different chunking (e.g., FAU AIBO database) levels. At the same time, these are partly also recorded under more realistic conditions (or taken from the media). However, in future sets multilinguality and subjects of diverse cultural backgrounds will be needed in addition to all named positive trends.

For our evaluations, we chose nine corpora amongst the most popular. Only these available to the research community were considered. These should cover a broad variety reaching from acted speech (the Danish (DES, [Engbert and Hansen, 1996]) and the Berlin Emotional Speech (EMO-DB, [Burkhardt et al., 2005]) databases), over story guided as the eNTERFACE corpus [Martin et al., 2006] with fixed spoken content and the Airplane Behaviour Corpus (ABC, [Schuller et al., 2009b]), to spontaneous with fixed spoken content represented by the Speech Under Simulated and Actual Stress (SUSAS, [Hansen and Bou-Ghazale, 1997]) database, to more modern corpora with respect to the number of subjects involved, spontaneity, and free language covered by the Audiovisual Interest Corpus (AVIC, [Schuller et al., 2009b]), the Sensitive Artificial Listener (SAL, [Wöllmer et al., 2008]), the SmartKom [Steininger et al., 2002], and the Vera-Am-Mittag (VAM, [Grimm et al., 2008]) datasets.

An overview on properties of the chosen datasets can be found in Table 2.3. Next, we will briefly introduce the datasets.

2.6.2.1 AIBO

It is a corpus with recordings of children interacting with Sony's pet robot called Aibo [Batliner et al., 2008]. The corpus consists of spontaneous, German speech which is emotionally colored. The data was collected at two different schools, MONT and OHM, from 51 children (age 10 - 13, 21 male, 30 female; about 9.2 hours of speech without pauses). Speech was transmitted with a high quality wireless head set and recorded with a DAT recorder (16 bit, 48 kHz down-sampled to 16 kHz). Five annotators (advanced students of

set	A	E	N	P	R	NEG	IDL	Σ
train	881	2,093	5,590	674	721	3,358	6,601	9,959
test	611	1,508	5,377	215	546	2,465	5,792	8,257

Table 2.4: *Number of instances for 2-class and 5-class annotation schema within AIBO corpus*

linguistics) listened to the turns and annotated each word as neutral or as belonging to one of ten other classes. The data is labeled on the word-level. We resort to majority voting (MV): if three or more labelers (five labelers in all) agreed, the label was attributed to the word. The number of cases with MV is given in parentheses: joyful (101), surprised (0), emphatic (2,528), helpless (3), touchy, i. e. irritated (225), angry (84), "motherese" (1,260), bored (11), reprimanding (310), rest, i. e. non-neutral, but not belonging to the other categories (3), neutral (39,169); 4,707 words had no MV; all in all, there were 48,401 words.

The whole corpus consisted of 18,216 emotional chunks. The five-class annotation schema covers the classes **A***nger* (subsuming angry, touchy, and reprimanding) **E***mphatic*, **N***eutral*, **P***ositive* (subsuming motherese and joyful), and **R***est* and they are to be discriminated. The two-class annotation schema consists of the covered classes **NEG***ative* (subsuming angry, touchy, reprimanding, and emphatic) and **IDL***e* (consisting of all nonnegative states).

The classes within the whole corpus are highly unbalanced. The transcriptions of spoken content within the training set are provided allowing for ASR training and linguistic feature computation.

2.6.2.2 Danish Emotional Speech

The Danish Emotional Speech (DES) [Engbert and Hansen, 1996] database has been chosen as the first set as one of the 'traditional representatives' for our study, because it is easily accessible and well-annotated. The data used in the experiments are nine Danish sentences, with two words and chunks that are located between two silent segments of two passages of the fluent text. For example: *"Nej" (No), "Ja" (Yes), "Hvor skal du hen?" (Where are you going?).* The total amount of data adds up to more than 500 speech utterances (i. e., speech segments between two silence pauses) which are expressed by four professional actors, two males and two females. All utterances are equally separated for each gender. Speech is expressed in five emotional states: *anger, happiness, neutral, sadness,* and *surprise.* Twenty judges (native speakers from 18 to 58 years old) verified the emotions with a score rate of 67 %.

2.6.2.3 Berlin Emotional Speech Database

A further well-known set chosen to test the effectiveness of emotion classification is the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [Burkhardt et al., 2005], which covers *anger, boredom, disgust, fear, joy, neutral,* and *sadness* speaker emotions. The spoken content is again pre-defined by ten German emotionally neutral sentences, such as *"Der Lappen liegt auf dem Eisschrank" (The cloth is lying on the fridge.)*. As with DES, it thus provides a high number of repeated words in diverse emotions. Ten (five female) professional actors speak ten German emotionally undefined sentences. While the whole set comprises of around 800 utterances, only 494 phrases are marked as a minimum 60 % natural and minimum 80 % assignable by 20 subjects in a listening experiment. 84.3 % mean accuracy is the result of this perception study for this limited "more prototypical" set.

2.6.2.4 eNTERFACE

The eNTERFACE [Martin et al., 2006] corpus is a further public, yet audiovisual emotion database. It consists of induced *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* speaker emotions. 42 subjects (eight female) from 14 nations are included. It consists of office environment recordings of pre-defined spoken content in English. Each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion.

They then had to react to each of the situations by uttering previously read phrases that fit the short story. Five phrases are available per emotion, such as *"I have nothing to give you! Please don't hurt me!"* in the case of fear. Two experts judged whether the reaction expressed the emotion in an unambiguous way. Only if this was the case, the sample was added to database. Overall, the database consists of 1,170 samples.

2.6.2.5 Airplane Behaviour Corpus

Another audiovisual emotion database is the Airplane Behaviour Corpus (ABC) [Schuller et al., 2009b], crafted for the special target application of public transport surveillance. In order to induce a certain mood, a script was used, which led the subjects through a guided storyline: prerecorded announcements by five different speakers were automatically played back and controlled by a hidden test-conductor. As a general framework a vacation flight with return flight was chosen, consisting of 13 and 10 scenes as the start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The general setup consisted of an airplane seat for the subject, positioned in front of a blue screen. 8 subjects in gender

balance from 25–48 years (mean 32 years) took part in the recording. The language throughout the recording is German. A total of 11.5 hours video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set. The average length of the 396 clips in total is 8.4 seconds.

2.6.2.6 Speech Under Simulated and Actual Stress

The Speech Under Simulated and Actual Stress (SUSAS) database [Hansen and Bou-Ghazale, 1997] serves as a first reference for spontaneous recordings. As an additional challenge, speech is partly masked by field noise. We decided for the 3,663 actual stress speech samples. Seven speakers, three of them female, in roller coaster and free fall actual stress situations are contained in this set. Next to *neutral* speech and *fear* two different stress conditions have been collected: *medium stress*, and *high stress*, and *screaming*. SUSAS is also restricted to a pre-defined spoken text of 35 English air commands, such as *"brake"*, *"help"* or *"no"*. Likewise, only single words are contained similar to DES where this is also mostly the case.

2.6.2.7 Audiovisual Interest Corpus

To add spontaneous emotion samples of non-restricted spoken content, we decided to use the Audiovisual Interest Corpus (AVIC) [Schuller et al., 2009b], another audiovisual emotion corpus. In its scenario setup, a product presenter leads one of 21 subjects (10 female) through an English commercial presentation. The level of interest is annotated for every sub-speaker turn reaching from *boredom* (subject is bored with listening and talking about the topic, very passive, does not follow the discourse), over *neutral* (subject follows and participates in the discourse, it cannot be recognized, if she/he is interested or indifferent in the topic) to *joyful* interaction (strong wish of the subject to talk and learn more about the topic). Additionally, the spoken content and non-linguistic vocalisations are labeled in the AVIC set. For our evaluation we use the 996 phrases as, e.g., employed in [Schuller et al., 2009b].

2.6.2.8 Sensitive Artificial Listener

The Belfast Sensitive Artificial Listener (SAL) data is part of the final HUMAINE database [Douglas-Cowie et al., 2007]. We consider the subset used, e.g., in [Wöllmer et al., 2008] which contains 25 recordings in total from 4 speakers (2 male, 2 female) with an average length of 20 minutes per speaker. The data contains audio-visual recordings from natural human-computer conversations that were recorded through an interaction interface designed to let

users work through a range of emotional states. The data was labeled continuously in real-time by four annotators with respect to valence and activation using a system based on FEELtrace [Cowie et al., 2000]: the annotators used a sliding controller to annotate both emotional dimensions separately whereas the adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. To compensate linear offsets that are present among the annotators, the annotations were normalized to zero mean globally. Furthermore, to ensure common scaling among all annotators, each annotator's labels were scaled so that 98 % of all values are in the range from -1 to +1. The 25 recordings have been split into turns using an energy-based Voice Activity Detection. A total of 1,692 turns is accordingly contained in the database. Labels for each turn are computed by averaging the frame-level valence and activation labels over the complete turn. Apart from the necessity to deal with continuous values for time and emotion, the great challenge of the SAL database is the fact that one must deal with all data – as recorded – and not only manually pre-selected 'emotional prototypes' as in practically any other database [Schuller et al., 2009c].

2.6.2.9 SmartKom

We further included a second audiovisual corpus of spontaneous speech and natural emotion in our tests: the SmartKom [Steininger et al., 2002] multi-modal corpus consists of Wizard of Oz dialogs in German. For our evaluations we use German dialogs recorded during a public environment technical scenario. As with SUSAS, noise is overlaid (street noise). The database contains multiple audio channels and two video channels (face, body from side). The primary aim of the corpus was the empirical study of human-computer interaction in a number of different tasks and technical setups. It is structured into sessions which contain one recording of approximately 4.5 minutes length with one person. Utterances are labeled in seven broader emotional states: *neutral*, *joy*, *anger*, *helplessness*, *pondering*, *surprise* are contained together with *unidentifiable* episodes.

2.6.2.10 Vera-Am-Mittag

The Vera-Am-Mittag (VAM) corpus [Grimm et al., 2008] consists of audiovisual recordings taken from a German TV talk show. The corpus contains 947 spontaneous and emotionally coloured utterances from 47 guests of the talk show which were recorded from unscripted, authentic discussions. The topics were mainly personal issues such as friendship crises, fatherhood questions,

or romantic affairs. To obtain non-acted data, a talk show in which the guests were not paid to perform as actors was chosen. The speech extracted from the dialogs contains a large amount of colloquial expressions as well as non-linguistic vocalisations and partly covers different German dialects. For annotation of the speech data, the audio recordings were manually segmented to the utterance-level, whereas each utterance contained at least one phrase. A large number of human labelers was used for annotation (17 labelers for one half of the data, six for the other).

The labeling bases on a discrete five-point scale for three dimensions mapped onto the interval of [-1,1]: the average results for the standard deviation are 0.29, 0.34, and 0.31 for valence, activation, and dominance. The averages for the correlation between the evaluators are 0.49, 0.72, and 0.61, respectively. The correlation coefficients for activation and dominance show suitable values, whereas the moderate value for valence indicates that this emotion primitive was more difficult to evaluate, but may partly also be a result of the smaller variance of valence.

2.7 Clustering of emotions

Although the ability to recognize a large variety of emotions is attractive, it may not be necessary or practical in the context of developing algorithms for conversational interfaces. Based on this assumption, some research groups favor the notion of an application-dependent reduced space of emotions. In particular, *negative and non-negative* emotions can be used for miscommunication detection tasks within automated spoken dialog systems [Lee and Narayanan, 2005].

It is possible to map the diverse emotion groups onto the most popular general dimensions (valence, arousal) borrowed from the dimensional emotion model: arousal and valence, see Figure 2.4. The chosen mappings [Schuller et al., 2009] are depicted in Table 2.5. Notably, these mappings are not straight forward. This would only be exactly true for the neutral emotion, which could have been chosen as a third state. Sadly, however, not all databases provide such a state. Thus, the mapping can be seen as a compromise in favor of better balance amongst the target classes. We further discretized emotion values in the arousal-valence plane for the emotional corpora with multi-dimensional annotation (SAL and VAM). We consider only four quadrants obtained by discretizing into binary tasks as described above, but now handling the problem as a four-class problem, see Figure 2.4. The according quadrant's q1–q4 (counterclockwise, starting in positive quadrant, assuming valence as ordinate and arousal as abscissa) can also be assigned emotion tags: "happy / excited"

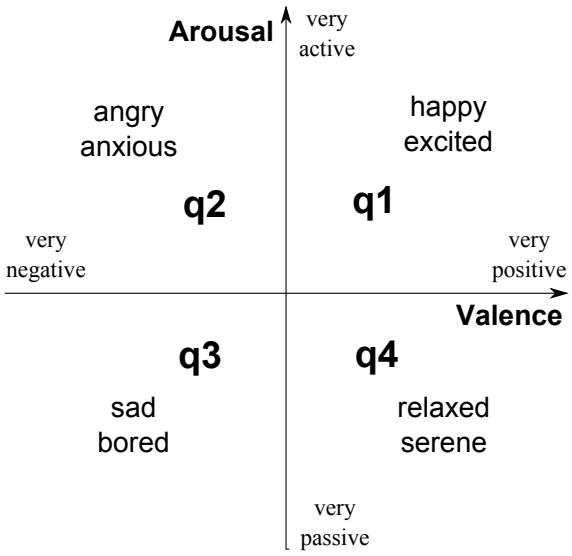


Figure 2.4: *Specification of the quadrant's q1–q4 in arousal-valence space*

Corpus	Arousal		Valence	
	Negative	Positive	Negative	Positive
ABC	neutral, tired	aggressive, cheerful, nervous, intoxicated	aggressive, nervous, tired	cheerful, intoxicated, neutral
AVIC	boredom	neutral, joyful	boredom	neutral, joyful
DES	neutral, sad	angry, happy, surprise	angry, sad	happy, neutral, surprise
EMO-DB	boredom, disgust, neutral, sadness	anger, fear, happiness	anger, boredom, disgust, fear, sadness	happiness, neutral
eNTER-FACE	disgust, sadness	anger, surprise, fear, happiness	anger, disgust, fear, sadness	happiness, surprise
SAL	q2, q3	q1, q4	q3, q4	q1, q2
Smart-Kom	neutral, pondering, unidentifiable	anger, helplessness, joy, surprise	anger, helplessness,	joy, pondering, neutral, surprise, unidentifiable
SUSAS	neutral	high stress, medium stress, screaming	high stress, screaming	medium stress, neutral
VAM	q2, q3	q1, q4	q3, q4	q1, q2

Table 2.5: *Mapping of emotions for the clustering to a binary (positive/negative) arousal and valence discrimination task. Abbreviations: q - quadrants*

(q1), "angry / anxious" (q2), "sad / bored" (q3), and "relaxed / serene" (q4).

2.8 Data assessment

Four main issues need to be considered in acquiring an emotional corpora; the scope, the level of naturalness and context of the content; and the type of corresponding descriptors [Douglas-Cowie et al., 2003].

- *Scope*
  It covers the amount of speakers presented in corpora; language spoken; gender variability of speakers; types of emotional state considered; level of annotation (word-level, utterance-level, context-independent time alignment); social/cultural setting (human-to-human interaction, task-oriented human-machine interaction). Real-life emotions in general are controlled by strong cultural influences [Harré, 1986]. Since speech is a cultural human activity, emotional events within speech may be related to cultural influences. Usually, within real-life verbal interaction, humans show less expressive emotions rather than full-blown.
- *Level of the naturalness*
  The simplest way to collect affective speech is to ask actors to simulate emotions within pronounced utterances. The main problem with this approach is that no in-depth research about relationships between acted material and spontaneous emotional speech has been done. It is of course true that preselected actors can generate speech that listeners classify reliably within a perception test. Still it is hard to measure how closely the prompted affective speech reflects spontaneous expression of emotion.
  From the other side, the price of high-level naturalness is a lack of control on the lexical and phonetic content of the material. For induced or spontaneous emotions it is difficult to collect samples in a target emotional state due to the unpredictability of the collecting process (users are able to use natural language for system interaction). A lot of applications (emotional speech synthesis, phoneme-level emotion modeling, etc.) require phonetically balanced datasets, which is hard to achieve within a truly natural speech interaction session.
- *Context*
  Three different types of context can be discriminated [Douglas-Cowie et al., 2003].
  - **Semantic context:**
  Sincere emotional speech is likely to contain words with a different level of emotionality. And this level of emotionality has a semantic nature. An example of emotionally significant words are emotive words (like *"good", "lovely", "aggression"* , etc.) that are part of some utterance.
  - **Structural context:**

Emotional events depend on the syntactic structure of the utterance: focus of attention, sentence stress, intonation variability, etc. Structural characteristics of the utterances (repetitions, rephrasing, interruptions and long pauses) can be used as indicators of change in the emotional state of the user. The sentence *"I really, really like this"* is an example of contextual amplification by repetition the word "really".

- **Temporal context:**

Spontaneous speech contains distinctive characters of change as emotion ebbs and flows during time. Due to their temporal nature, some words within an utterance can be more expressive in comparison with their neighbor words. By interpreting nearby utterances and words we can resolve local ambiguity in emotional state classification. The sentence *"This was a great failure"* contains positive in general but negative in context the word "great".

- *Descriptors*

Describing the para-linguistic and emotional content on one hand, and transcribing the speech on the other is an important issue of constructing a high-standard database. The requirements for correct labeling of emotional events may be a concern to the level of naturalness. Acted emotions can be adequately described with emotion category labels from a basic emotions list. Corpora with spontaneous emotions, though, can require a gradation of the emotion (cold angry, hot angry, etc.) and indication of the most expressive peaks within an utterance.

There are two main issues in terms of speech descriptors: First, the full range of features responsible for the vocal expression of emotion should be taken into account. This range of features should include at least the prosodic description, non-linguistic features like *breathing, clatter, laughter,* and *crying.* Second, it is important to describe the attributes that define emotional states and their dynamic specification (intensity variability in the time domain). As discussed in section 2.5.1 and section 2.5.2, emotions can be described with emotion categories or numeric values within a two- or three-dimensional space, namely *valence-arousal-(dominance) VA(D).*

Providing "ground truth" measures within emotional content annotation is an important issue. Defining "ground truth" measures for emotions described in numerical values in VA(D) space is a non-trivial task. It can also be problematic to measure "ground truth" for real-life emotions defined with discrete emotion categories which have a mixed nature or low-intensity.

To be able to measure the quality of the emotional annotation, inter-

rater reliability measures as an alternative to the "ground truth" have been introduced. To estimate the inter-rater agreement, it is common to use the *Kappa* coefficient κ [Carletta, 1996]:

$$\kappa = \frac{P_A - P_0}{1 - P_0} \tag{2.1}$$

where P_A corresponds to the proportion of the raters that assigned the same class label, P_0 is an estimation of the proportion where raters agree by chance.

A description of our annotation strategy with an example of the adequate annotation of spontaneous emotions will be given in section 2.8.1.

2.8.1 An adequate annotation strategy

An annotation process is the most expensive and time-consuming part within prosodic speech corpora development. Two of the key points identified in the previous section – scope and level of naturalness – are described in Table 2.3. This table is designed to provide some brief information about existing emotional speech corpora. Scope describes the language specification, number of speakers, and emotions considered. Under level of naturalness, we consider several categories: acted, spontaneous, mixed (contain both acted and spontaneous samples); and the type of material (e.g., sentences, utterances, short commands).

As one can see, just five (AVIC, EMO-DB, ENTERFACE, SmartKom, VAM) from nine datasets contain the sufficient amount of speakers. To be able to model inter-subject variability, corpora should contain enough female and male speakers (at least 5 speakers for each gender).

A good example of a reliable and close to "natural" acted emotional speech database is the Berlin Emotional Speech Database [Burkhardt et al., 2005]. The emotion recognizability level, and the level of naturalness estimated within a perception test for each utterance, are presented in this database. To provide reliable measures, twenty perception-test evaluators took part in this test. Each "rater" heard all of the utterances in a random order. They were allowed to listen to each utterance only once before the perception-test evaluator had to decide in which emotional state the speaker had been and how persuasive the performance was. Within our recognition evaluations, see Chapter 5, we used utterances with a minimum 60 % level of naturalness and minimum 80 % recognizability level. In practice, the perception test implemented for evaluation of the Berlin Emotional Speech Database [Burkhardt et al., 2005] with estimation of the levels of naturalness and recognizability

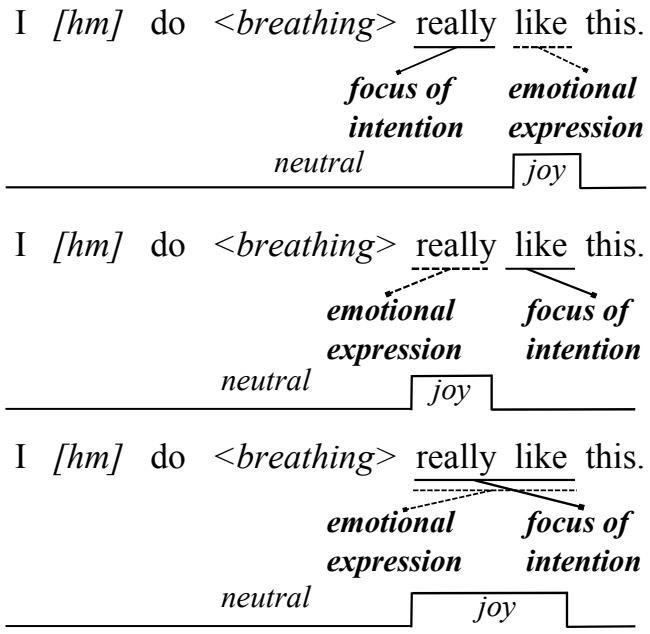


Figure 2.5: *An example of reliable spontaneous affective speech annotation*

for each emotional utterance can be used as a "ground truth" measure of the level of naturalness.

In case of applicable affective speech annotation, two issues stand out: Firstly, transcription needs to acknowledge the full range of features involved in the acoustic expression of emotion, including voice quality, boundary prosody and non-linguistic features such as *laughter, crying, clatter*, and *breath*. Secondly, it needs to describe the attributes (e.g., linguistic, dialog acts specification) that are relevant to emotion. An example of reliable affective spontaneous speech annotation is presented in Figure 2.5.

As one can see, the structural context (focus of attention, sentence stress, intonation variability, etc.) should be carefully annotated. There is a high correlation between boundary and emotional prosody. Annotators should be extremely careful with distinguishing between these two different events. An example of a possible conflict between focus of attention (boundary prosody event) and emotional events is presented in Figure 2.5. Each of these utterances have slightly different semantic accents which should be taken into account by human-distinguishable boundary prosody and emotional prosody events. Afterwards, we will be able make annotation process faster and reach higher quality of annotation within the spontaneous affective speech description task [Siegert et al., 2011].

In real-life communication humans use a number of different variations to denote emphasis in speech. Speakers may render emphasis with different

combinations and even individuals may change their strategies for various prosodic cues (boundary prosody and emotional prosody). In the first two sentences presented in Figure 2.5 the words the "really" and "like" can be pronounced with emphasis. At the same time only one word is pronounced emotionally. In the first sentence the speaker points out that his emotions are real, not simulated. In the second sentence the speaker places an accent on the action ("like"). It is quite important to distinguish emphasis which represents two different paralinguistic phenomena. As one can see, from the third sentence these phenomena can be mixed. In this case both words "really" and "like" are pronounced with emphasis and emotional prosody cue. Correct interpretation of those sentences can provide system information about the speaker's intentions.

Most datasets evaluated in our recognition experiments and described in Table 2.3 used a description of emotion with defined emotion categories list. Only two databases (VAM, SAL) implemented the VA(D) dimensional approach. From our point of few, both types of emotional state descriptors have advantages and disadvantages. In a case of emotion-categories-based descriptors, we can model different dialog strategies for different emotional state subsets in contrast to the emotions defined by VA(D) dimensions. Also, it is much easier to organize perception evaluation with a defined or "open" emotion categories list in contrast to emotion perception evaluation with the VA(D) space, where "raters" should be preliminarily trained to be able to make reliable emotional annotations. As described earlier, it is easier to provide "ground measures" for acted emotions annotated with a set of emotion categories. From the other side, VA(D) dimension-based annotation provides a higher-level of discrimination. As a consequence, mixed emotions and emotions with light exclusivity can easily be defined with numeric values in VA(D) space. Of course, standard mapping of categorical emotions on VA(D) dimensional space will be appreciated. Due to the huge variability of "rater"-dependent measures of categorical emotions within VA(D) space, no standard mapping technique exist. Grimm et al. in [Grimm et al., 2007] proposed *evaluator weighted estimator* (EWE). They introduced evaluator-dependent weights which measure the correlation between the listener's responses, and the average ratings of all evaluators. These weights can be used as a possible normalization technique for the variable "rater"-dependent measures.

Our emotion-classification engine, integrated into the NIMITEK (Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems) demonstrator [Wendemuth et al., 2008], has been trained on the EMO-DB database which is annotated with emotion categories descriptors. A detailed introduction to various types of speech-based emotion-classification techniques will be given in Chapter 4. A NIMITEK demon-

strator's dialog module supports different strategies based on an actual user's emotional state. More details on this can be found in Chapter 6.

2.9 Evaluating recognition results

Once the test material has been processed by the recognizer, the next step is to analyze the results. The main aim of this analysis is a representation of recognition performance of evaluated classifiers. Also, this analysis can be used for comparison of recognition performances during iterative classifier parameters tuning. Within our research we use different measures to characterize performance of ASR and emotion recognition from speech. These measures will be described in this section.

2.9.1 Automatic speech recognition

For estimating the performance of automatic speech recognition we use standard measures included in the HTK tool [Young et al., 2009]. The HResults tool has been used to estimate ASR performance. It compares the transcriptions output from the ASR engine with the original reference transcriptions and then generates various statistical measures. HResults matches each of the recognized and reference label sequences by retrieving an optimal string match using dynamic programming.

Once the optimal alignment has been found, the number of deletion errors (D), substitution errors (S) and insertion errors (I) can be estimated [Young et al., 2009]. The percentage of correct recognized labels is called *correctness* and is given by

$$Corr = \frac{N - D - S}{N} \times 100\,\% \tag{2.2}$$

where N is the total number of labels presented in the reference transcriptions. This measure ignores insertion errors. Taking into consideration insertion errors, the percentage of so-called *accuracy* is defined as

$$Acc = \frac{N - D - S - I}{N} \times 100\,\% \tag{2.3}$$

which is a more representative figure of ASR performance. For the evaluations of our ASR engine we will use both measures.

2.9.2 Emotion recognition

As classes are often unbalanced in the emotional speech datasets, see Table 2.3, we decided to use two different evaluation measures for presenta-

tion of emotion-recognition performances: *unweighted average recall (UA)* and *weighted average recall (WA)*.

Unweighted average recall (*UA*) is the sum of all class accuracies, divided by the number of classes, without considering the number of instances per class. Weighted average recall (*WA*), also known as *accuracy*, is the accuracy per class, including consideration of the number of instances per class. In other words *WA* (*accuracy*) is the number of instances with correctly classified classes, divided by the total number of classified instances. For estimating *WA* we simply calculate *Acc* presented in equation 2.3. For this purpose we use the HResults tool.

To show the difference between *UA* and *WA* measures, let's consider an example. We have an emotional speech dataset with 99 joy samples and 1 anger sample. That is to say, we have heavily unbalanced class distributions within our dataset. If our classifier recognizes all 100 samples as joy, an accuracy of emotion recognition $WA = 99\,\%$, which is a really good result. At the same time, our classifier was not able to classify an anger sample. To show "real" emotion-recognition performance of our classifier it is better to use *UA* rate. For our example it can be calculated as

$$UA = \frac{\frac{99}{99} + \frac{0}{1}}{2} \times 100\,\% = 50\,\% \tag{2.4}$$

Now we can resume that our classifier has $WA = 99\,\%$ which is a really good performance from one side, and has $UA = 50\,\%$ which is equal to selection "by chance" of a possible emotional state for two emotional classes recognition task.

While tuning our classifiers we should use the most reliable measures. If we have balanced class distributions within an emotional speech dataset we can use *WA*, in the other case it is better to use *UA*. If the classifier parameters are optimized on the measure of *WA* (number of accurately classified samples by total number of tested samples), it will likely recognize only a few of the dominant emotional classes accurately. Unweighted average recall provides a method for estimating the performance of a classifier in emotionally biased datasets. For the estimation of *UA* we use our own Perl script which provides a detailed comparison of recognized and reference emotional labels.

2.10 Evaluation strategies

The most general parameters for evaluating the performance of a classifier are its general recognition rates (*UA*, *WA*, *Acc*, *Corr*), and they have to be estimated on the source dataset S. Usually the number of class instances

in dataset S is quite small. Limited availability of the data source or high expenses of data collection are the main reasons for a sparse amount of the data.

A common methodology for evaluating the recognition rates is to split the source dataset into two subsets: training and test set. The training set is used for training purposes and the test set is applied to estimate the recognition rate of the earlier trained classifier. This process is usually repeated multiple times (with different random or preselected subunits of the dataset into training and test sets), and the average of all estimated recognition rates gives an estimation of the general recognition rate.

2.10.1 Speaker-dependent evaluation

Within a N-fold cross-validation strategy, a dataset S is first randomly divided into n disjoint subsets $S_1, S_2, \ldots, S_N$, which have an equal or quasi-equal amount of instances per class. Each of the n subsets is then one after another applied as the test set, while the remaining $n-1$ subsets are applied as the training set. A classifier is then trained on the training set material, and its accuracy is estimated on the test set material. This process is repeated n times, with a different subset applied as the test set. The evaluated general recognition rates by this method is the average over the n subsets. An extension to cross-validation is a stratified cross-validation. Within a N-fold stratified cross-validation strategy, a dataset S is divided into n subsets in such a way that each class is uniformly distributed among the n subsets [Zeng and Martinez, 2000].

For our speaker-dependent evaluations we applied a 10-fold stratified cross-validation (SCV) strategy. Such strategy is used for datasets which have a small amount of data per class instance and/or per speaker presented in a corpus (SUSAS, DES).

2.10.2 Speaker-independent evaluation

To address speaker independence (SI) within our evaluations we applied *leave-one-speaker-out* (LOSO) or *leave-one-speakers-group-out* (LOSGO) strategies. In such a way we simulate close to real-life application conditions. For these strategies, evaluation material should contain a sufficient amount of instances (emotional samples, utterances) per each speaker presented in the dataset. Within LOSO strategy the number of folds n presented in the previous section is equal to the number of speakers presented in corpora. In the case of LOSGO strategy n is a number of speaker groups. In contrast to a random partitioning process within a cross-validation strategy we divided a dataset S

into n folds in such a way that each fold contains samples of only one speaker (within LOSO) or only one speaker group (LOSGO). An additional advantage of these methods is a possibility to concentrate on inter-speaker variation and not to deal with acoustic channel changes. For presentation of the recognition performance within an evaluation based on LOSO strategy we estimate the average evaluation measures (*UA*, *WA*, *Corr*, *Acc*). For this purpose we developed a Perl script which analyzes the recognition results for each speaker (leave-one-speaker-out trial) within the complete evaluation cycle.

2.10.3 Cross-corpora evaluation

Within the previously described strategies we conclude a simplification that characterizes that most of the current speech-processing research is that classifiers are usually trained and tested using the same datasets. By using two different datasets for training and testing we can simulate that, in particular development tasks, corpora may not be available which cover all emotions of speaker in a given application domain. This type of experiments called *cross-corpora evaluation*. Speaker-independent evaluations (LOSO, LOSGO) have become quite common, still other mismatches between training and test datasets, such as different recording conditions (including different acoustic environment, acoustic channel characteristics, microphone types, signal-to-noise ratios, etc.), are often not considered. Addressing such typical sources of mismatch, however, we believe that an impression about the generalization ability of speech-based emotion recognition and automatic speech-recognition engines can be obtained by cross-corpora evaluations. A considerably more realistic impression can be gathered by *interset evaluation*: We therefore use a cross-corpora evaluation experiment, which could also be helpful for learning about chances to add resources for training and overcoming the typical sparseness in the field. By using cross-corpora evaluation for emotion-recognition experiments we want to estimate emotion-recognition performance in conditions which are close to real-life development tasks.

2.11 Summary

This chapter reviews the fundamentals of the user-centered human-machine interaction. The variety of existing spoken dialog systems with German interaction language is described first. Characteristics of the natural human speech, namely boundary and emotional prosody, are then presented. The emotion theory and existing emotion-categorization schemes are presented in detail. Different sources of emotional speech data are then introduced. Also,

a possible emotion clustering technique is then introduced. Then, the main issues of adequate annotation of the affective speech are presented. Finally, a variety of recognition rate measures and evaluation strategies are discussed.

In the next chapter we will describe the general architecture of the automatic speech-recognition (ASR) system. Some ASR methods will be used for our phoneme-level emotion-recognition methods. Methods described in the next chapter have been used to create an ASR module integrated in our NIMITEK demonstration prototype of a spoken dialog system (SDS). Also, we need the ASR system for time alignment within phoneme-level emotion classification. Finally, the ASR module can be used for semi-automatic transcription of the data collected during a Wizard of Oz scenario.

CHAPTER 3

Spontaneous affective speech recognition

Contents

3.1 Introduction

In this chapter an introduction to automatic spontaneous speech-recognition system with acoustic model based on hidden Markov models (HMMs) is given. Main aspects of the concept presented in Figure 3.1 are described in this chapter, namely feature extraction, the mathematical description of an HMMs-based algorithm, a selection of the sub-word units and their quantitative and qualitative specification, the decoding algorithm for spontaneous speech, a language modeling and the adaptation techniques for a robust affective speech recognition.

3.2 General ASR models/architecture

Automatic speech recognition (ASR) is a task of converting acoustic waveform automatically to a word sequence. The basic structure of an ASR system is presented in Figure 3.1.

Converting of an acoustic speech signals into stream of acoustic features, referred to as *observations* is the first stage of speech recognition. So-called, *front-end processing* or *feature extraction* have to generate compact acoustic observation vectors with sufficient information applicable for efficient

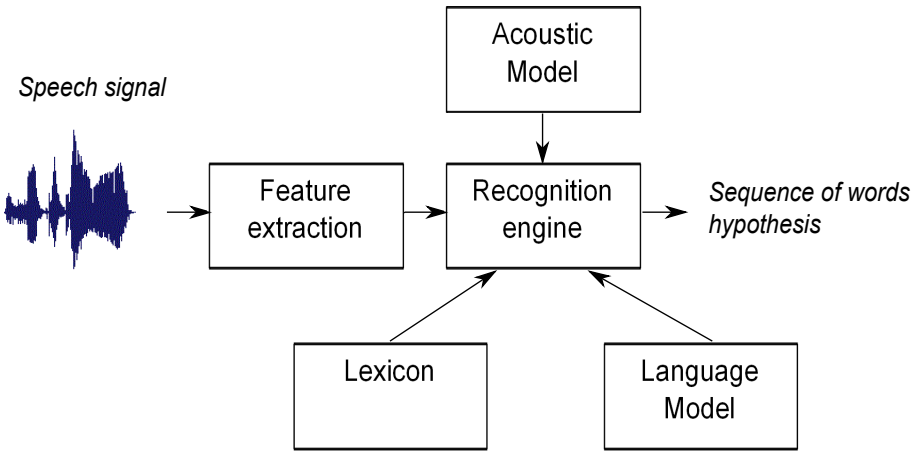


Figure 3.1: *General structure of a standard ASR system*

recognition. Three types of components are required for a standard speech-recognition system: the *lexicon* (or dictionary), *language model* and *acoustic model*. The lexicon is usually used to map phonetic units (monophones, triphones, etc), from which the acoustic models are built, to the hypothesis word present in the lexicon and language model. The language model represents a-priory information about syntactic and semantic structure of the uttered sentences, which include the possibility of each possible word sequence. The acoustic model maps the acoustic observation vectors to the phonetic units. A detailed description to various components in Figure 3.1 will be given later in Chapter 3.

Statistical analysis is the most popular speech-recognition algorithms to determine word sequence hypothesis given the information presented in Figure 3.1. The main decision criterion to find the most likely word sequence hypothesis $\hat{\mathcal{W}}$ for the sequence of observation vectors $\mathbf{O} = [\mathbf{o}_1 \dots \mathbf{o}_T]$ is the Bayesian decision rule [Young, 1995]:

$$\hat{\mathcal{W}} = \arg\max_{\mathcal{W}} P(\mathcal{W}|\mathbf{O}) = \arg\max_{\mathcal{W}} \left\{ \frac{p(\mathbf{O}|\mathcal{W})P(\mathcal{W})}{p(\mathbf{O})} \right\} \tag{3.1}$$

Take into account that the most likely word sequence is independent of the likelihood of the observation

$$\hat{\mathcal{W}} = \arg\max_{\mathcal{W}} \{p(\mathbf{O}|\mathcal{W})P(\mathcal{W})\} \tag{3.2}$$

where $P(\mathcal{W})$ is the prior probability of a particular sequence of words presented by a language model. $p(\mathbf{O}|\mathcal{W})$ is estimated by the acoustic model which is in most cases implemented as hidden Markov models (HMMs).

3.2.1 Feature extraction

For effective speech recognition, the speech signal is usually converted into a series of discrete time acoustic features. These acoustic features are supposed to present speech variability in a compact form. In the speech-processing community these features are often referred to as *feature vectors* or *observations*. The most widely used feature extraction scheme applied in ASR systems is a Mel-frequency Cepstral coefficient (MFCC).

The MFCC extraction is based on cepstral analysis. Firstly, the acoustic signal is split into discrete frames usually with a 10 ms shifting step and a 25 ms window length. These parameters were estimated based on the quasi-stationarity property of the speech signals [Rabiner and Juang, 1993]. These discrete fragments are usually referred to as *frames*. The feature extraction is applied for each frame. A first-order pre-emphasizing technique in combination with a Hamming smoothing window are used. The pre-emphasizing is implemented with high-frequency amplification to compensate for the attenuation produced by the radiation from the lips [Young, 1995]. Using a window function like Hamming, is useful for a boundary effect reduction. A fast Fourier transform (FFT) is performed on the time-domain acoustic signal for each individual frame, generating speech representation in complex frequency domains. Afterwards, the frequency warping methods are used [Young et al., 2009]:

- *Mel-frequency warping:*
  Within psychophysical experiments it has been shown that human perception of the frequency content of acoustic signals does not follow a linear scale. Therefore the frequency is warped using the Mel-frequency scale, with following frequency axis scaling. Estimation of the magnitude of each FFT complex value will be processed in a scaled magnitude-frequency domain.
- *Down-sampling with triangular filter bank:*
  By using the mel triangle filter bank we can down-sample the warped magnitude-frequency domain. The magnitude coefficients are multiplied by filter gains, afterwards the results are accumulated as the amplitude value, see Figure 3.2. As a consequence, one amplitude value was calculated for each filter. As a next step the logarithm of each filter amplitude value is calculated, later referred as m_j, where j is a filter number. For our evaluations we used the *lower cut-off* equal to 300 Hz and the *upper cut-offs* equal to 3,400 Hz.

- *Discrete Cosine transform (DCT):*
  A DCT is conducted on the log filter-bank amplitudes, to reduce the spa-

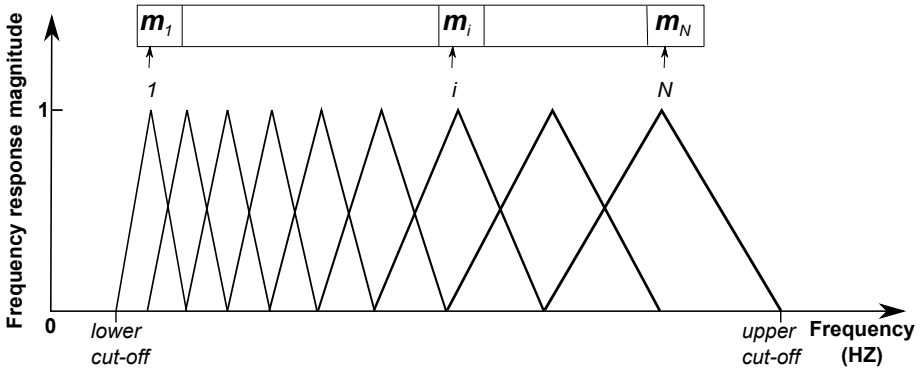


Figure 3.2: *Triangular mel-scale filter bank*

tial correlation within filter bank amplitudes. The DCT coefficients calculated by equation 3.3 are referred as *Cepstral coefficients*, also known as *MFCC* coefficients.

$$c_i = \sqrt{\frac{2}{N_{ch}}} \sum_{j=1}^{N_{ch}} m_j \cos\left(\frac{\pi i}{N_{ch}}(j - 0.5)\right) \tag{3.3}$$

where N_{ch} is the number of triangle filter bank channels.

Within our evaluations the 12 coefficients and the zero-order Cepstral coefficient are used. Hence a 13-dimensional feature vector is constructed for each frame.

By adding dynamic coefficients the performance of ASR system can be greatly enhanced. These time derivative features represent the correlation within static features for the different time instances. The *delta coefficients*, $\Delta \mathbf{o}_t$, are computed using the following linear regression formula:

$$\Delta c_t = \frac{\sum_{k=1}^{K} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2} \tag{3.4}$$

where Δc_t is a *delta* coefficient at the discrete time t with respect to the static coefficients c_{t-k} and c_{t+k}; K is the width over which delta coefficients are calculated. Within our evaluations we applied $K = 2$. The *delta-delta* coefficient $\Delta(\Delta c_t)$, or so-called *acceleration* features or second-order delta coefficients, is defined in equation 3.4. In this case the static coefficients c_{t-k} and c_{t+k} in equation 3.4 are replaced by the first-order delta coefficients Δc_{t-k} and Δc_{t+k}. For our evaluations we used both: delta and acceleration coefficients in addition to the 13-dimensional MFCC feature vector. As a result a 39-dimensional acoustic feature vector is constructed for each window of analysis.

3.2.2 Acoustic model

Now to be able to evaluate on observation vectors sequences, we need an acoustic model. The most robust and general acoustic technique in automatic speech recognition are hidden Markov models (HMM). The first applications of HMMs for the acoustic modeling were used in the mid-1970s [Baker, 1975]. Currently, the HMMs-based acoustic models are presented in the HTK toolkit [Young et al., 2009] an extremely popular in speech-processing community. For our evaluations we used this toolkit, to create and test our German acoustic models.

The main goal of the acoustic model is to supply a method of estimation of the likelihood of any observation feature vectors sequence $\mathbf{O}$ given a hypothetical word sequence $\mathcal{W}$. For small vocabulary speech-recognition tasks, HMMs can be used to model single words. However, for speech-recognition application with large vocabularies, it is impossible to acquire sufficient training material for each word included in the vocabulary. One possible solution to this problem is to use HMMs to model *sub-word (phonetic)* units, instead the words themselves. More details about this decomposition and type of the sub-word unit selection can be found in section 3.3.2.

The HMM is a generative statistical model where each sub-word unit is supposed to be generated by a finite state machine. This state machine, could change an active state at some discrete time with a predefined probability. When an emitting state is activated, an observation vector is generated at that discrete time instance with a defined probability function. A *left-right* HMM with three emitting and two non-emitting states is the most popular topology applied for monophone-based ASR system, see Figure 3.3. The entry and exit states are produced to facilitate sub-word models connections. The exit state of one sub-word model can be joined with the entry state of the

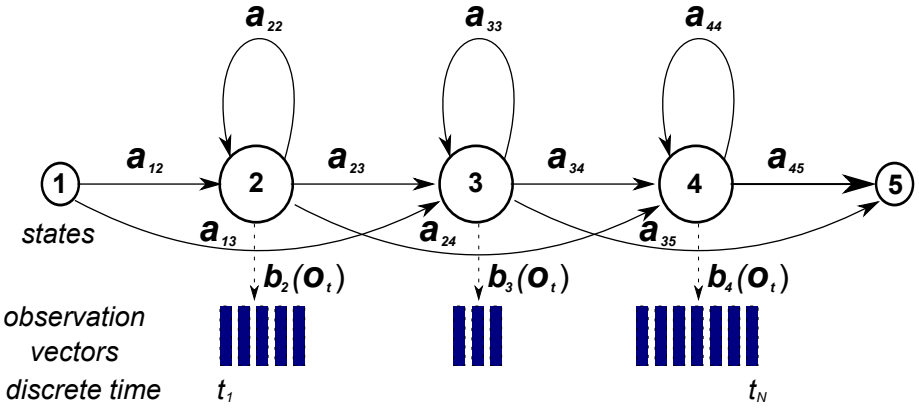


Figure 3.3: *Simple left-right HMM with five-state topology*

next sub-word model to arrange composite HMM.

To be able to use a HMM, two assumptions should be true:

- *The stationarity assumption:*
  The speech waveform can be divided into stationary fragments, which correspond to the same hidden states. It is required that observation vectors within the same fragments have similar acoustic characteristics. Transactions from one state to another are supposed to be instantaneous.

- *The observation independence assumption:*
  A generation of a current observation is statistically independent of the previous and following generated observations. From that assumption the following equation can be formed:

$$p(\mathbf{O}|s_1, s_2, \ldots, s_T, \mathcal{M}) = \prod_{t=1}^{T} p(\mathbf{o}_t|s_t, \mathcal{M}) \tag{3.5}$$

  where $\mathbf{O}$ is an observation sequence $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$, s_t is an active state at the discrete time t, $\mathcal{M}$ is an HMM's parameter set.

Suppose $\mathbf{O}$ is an observation vectors sequence $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$ corresponding to some sample of a particular phonetic unit (monophone, triphone, etc), where T is the length of the vector sequence or in other words the duration in discrete time samples. The generation begins from the first non-emitting state. At each discrete time, an active state can be switched with the probability given by the model. The transition probability, is defined as a discrete distribution a_{ij} for the possible transitions from state i to state j. During the emitting state activation process, an observation vector is generated at the discrete time with either discrete or continuous density $b_j(\mathbf{o}_t)$, where j is an active state number. Let's assume that $\mathbf{s} = [s_1, s_2, \ldots, s_T]$ is the state sequence associated with the observation vectors sequence. Within modeling, only the observation vector sequence can be observed and the corresponding state sequence $\mathbf{s}$ is unknown. This is the reason why the model is called the *hidden* Markov model.

The HMM's parameter set $\mathcal{M}$ consists of the following parameters [Rabiner, 1989]:

- π - *Initial state distribution*
  The initial state distribution is expressed as:

$$\pi_i = P(s_1 = i), \quad \sum_{i=1}^{N} \pi_i = 1, \quad \pi \geq 0 \tag{3.6}$$

where N is the number of sates, s_t is an active state number at the discrete time t.

- **A** - *State transaction probability matrix*
  The state-transaction probability matrix **A** includes the following elements:

$$a_{ij} = P(s_{t+1} = j | s_t = i), \quad \sum_{j=1}^{N} a_{ij} = 1, \quad a_{ij} \geq 0 \tag{3.7}$$

- **B** - *Observation generation probability distribution*
  Every emitting state k is associated with an output probability distribution, which is responsible for the observation vectors generation at each discrete time instance. The following distribution is expressed as

$$b_k(\mathbf{o}_t) = p(\mathbf{o}_t | s_t = k) \tag{3.8}$$

  The state output probability distribution can be defined with a discrete distribution or a continuous density distribution function. For our evaluations we use the continuous density distribution case.

In context of the ASR task, there are three following basic problems for HMMs [Rabiner and Juang, 1993]:

- *Probability evaluation*
  Given the observation vectors sequence $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$, and a HMM's model $\mathcal{M} = (\pi, \mathbf{A}, \mathbf{B})$, how can we estimate $p(\mathbf{O}|\mathcal{W}, \mathcal{M})$. This problem can be solved with the forward-backward algorithm.
- *Optimal state sequence decoding*
  Given the observation vectors sequence $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$, and the model $\mathcal{M}$, what is the optimal state sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$. The Viterbi algorithm can be used to solve this problem [Viterbi, 1967].
- *Parameters Estimation*
  How do we estimate the model parameters $\mathcal{M} = (\pi, \mathbf{A}, \mathbf{B})$ which maximize $p(\mathbf{O}|\mathcal{W}, \mathcal{M})$? The *Baum-Welch* re-estimation algorithm can be used as a solution for the following problem [Baum et al., 1970].

3.2.3 Probability evaluation

Let's say we have the observation vector $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$ which corresponds to some hypothetical word sequence $\mathcal{W}$. We wish to calculate the likelihood of the observation vector $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$, for the given HMM model $\mathcal{M} = (\pi, \mathbf{A}, \mathbf{B})$. As mentioned earlier the state sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$ is hidden. As a consequence, the most straightforward way of likelihood $p(\mathbf{O}|\mathcal{W}, \mathcal{M})$ estimation is through enumerating all possible

state sequences, which can generate an observation vectors sequence $\mathbf{O}$ of length T. We should take into account N^T possible state sequences.

Take into account the observations independence assumption (see equation 3.5), the likelihood of the observation vectors sequence $\mathbf{O}$ generation by the given state sequence $\mathbf{s}$ may be expressed as:

$$p(\mathbf{O}|s_1, s_2, \ldots, s_T, \mathcal{W}, \mathcal{M}) = \prod_{t=1}^{T} b_{s_t}(\mathbf{o}_t) \tag{3.9}$$

The likelihood of such a state sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$ can be estimated by:

$$p(s_1, s_2, \ldots, s_T|\mathcal{W}, \mathcal{M}) = \pi_{s_1} \prod_{t=2}^{T} a_{s_{t-1}s_t} \tag{3.10}$$

By using equations 3.9, 3.10 the likelihood $p(\mathbf{O}|\mathcal{W}, \mathcal{M})$ may be estimated by accumulating the joint likelihood of $\mathbf{O}$ and $\mathbf{s}$ over all possible state sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$

$$\begin{aligned} p(\mathbf{O}|\mathcal{W}, \mathcal{M}) &= \sum_{\forall \mathbf{s}} p(\mathbf{O}, \mathbf{s}|\mathcal{W}, \mathcal{M}) \\ &= \sum_{\forall \mathbf{s}} p(\mathbf{s}|\mathcal{W}, \mathcal{M}) p(\mathbf{O}|\mathbf{s}, \mathcal{M}) \\ &= \sum_{\forall \mathbf{s}} \pi_{s_1} \prod_{t=1}^{T} b_{s_t}(\mathbf{o}_t) a_{s_{t-1}s_t} \end{aligned} \tag{3.11}$$

where $a_{s_0 s_1}$ is an initial transition probability from the first non-emitting state to the emitting state, is equal to 1.

To estimate the likelihood expressed in equation 3.11, we should be able to model the distribution $b_j(\mathbf{o}_t)$. One of a possible continuous density HMM technique is based on a multivariate *Gaussian mixture model* (GMM). Besides, the $b_j(\mathbf{o}_t)$ can be represented as a multivariate GMM [Yu, 2006]:

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M_j} c_{jm} b_{jm}(\mathbf{o}_t) \tag{3.12}$$

where M_j is the number of Gaussian mixture components related to the state j , c_{jm} is a weight coefficient of m component of the state j. Each component $b_{jm}(\mathbf{o}_t)$ is the D-dimensional multivariate Gaussian distribution with the following parameters $\mathcal{N}\left(\mathbf{o}_t|\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right)$:

$$b_{jm}(\mathbf{o}_t) = \frac{1}{\sqrt{(2\pi)^D|\mathbf{\Sigma}_{jm}|}}\exp\left\{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jm})^T{\mathbf{\Sigma}_{jm}}^{-1}(\mathbf{o}_t - \boldsymbol{\mu}_{jm})\right\} \quad (3.13)$$

where $\boldsymbol{\mu}_{jm}$ is a mean vector of m component and j HMM's state, and $\mathbf{\Sigma}_{jm}$ is a covariance matrix of m component and j HMM's state.

3.2.3.1 The forward process

Consider the *forward* variable $\alpha_j(t)$, is defined as the joint likelihood of the partial observation vectors from corresponding discrete time interval from 1 to t with the final active state $s_t = j$:

$$\alpha_j(t) = p(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, s_t = j|\mathcal{W}, \mathcal{M}) \quad (3.14)$$

The forward variable of the partial observation vectors sequence $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t$ and an active state i at the discrete time t can be efficiently calculated using a recursive formula:

$$\alpha_j(t+1) = b_j(\mathbf{o}_{t+1})\sum_{i=1}^{N}\alpha_i(t)a_{ij} \quad (3.15)$$
$$1 \leq t \leq T-1,\ 1 \leq j \leq N$$

where N is the total number of HMM's states (emitting and non-emitting). The initialization condition for equation 3.15 is:

$$\alpha_j(1) = \pi_j b_j(\mathbf{o}_1), \quad 1 \leq j \leq N \quad (3.16)$$

By using the forward variable, equation 3.11 in section 3.2.3 can be rewritten as:

$$p(\mathbf{O}|\mathcal{W}, \mathcal{M}) = \sum_{i=1}^{N}\alpha_i(T) \quad (3.17)$$

Calculation of the forward variable is based on the lattice tracking. The general model of the lattice an N state HMM is presented in Figure 3.4. At the initial discrete time $t = 1$, we need to compute forward variables $\alpha_j(1)$, $1 \leq j \leq N$. Afterwards, we need only compute forward variables $\alpha_j(t)$, $1 \leq j \leq N$ at the discrete time $2 \leq t \leq T$. Each calculation uses just the N previous forward variables $\alpha_j(t-1)$ because each of N lattice nodes can be reached from only the N lattice nodes at the previous discrete time slot [Rabiner and Juang, 1993]. Calculation of all $\alpha_j(t)$ forward variables requires on the order of N^2T calculation, in comparison with $2TN^T$ calculations required by the direct computation method.

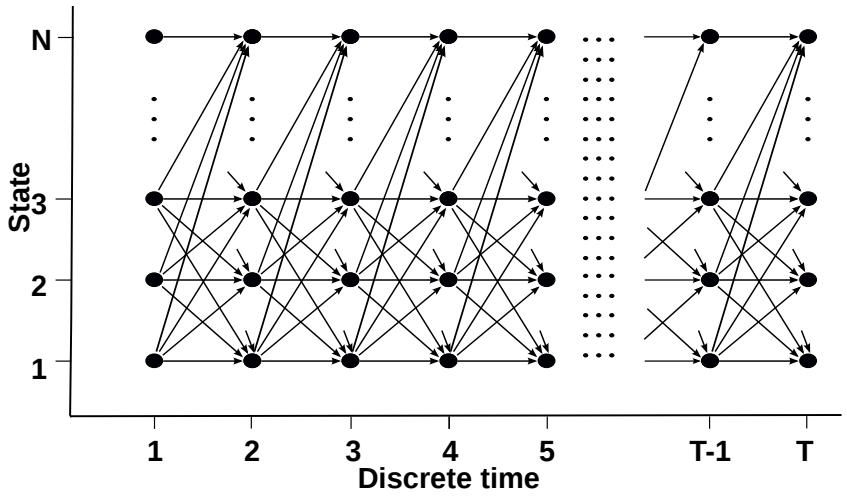


Figure 3.4: *General representation of the series of operations required for estimation forward variable* $\alpha_i(t)$

3.2.3.2 The backward process

In a similar way, we can can define a *backward* variable, $\beta_t(j)$, as

$$\beta_j(t) = p(\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, \ldots, \mathbf{o}_T | s_t = j, \mathcal{W}, \mathcal{M}) \tag{3.18}$$

that, is the probability of the partial observation vectors sequence from discrete time $t+1$ to the end, with an active state j at the discrete time t.

The backward variable can be calculated using the following recursion:

$$\beta_j(t) = \sum_{i=1}^{N} a_{ji} b_i(\mathbf{o}_{t+1}) \beta_i(t+1) \tag{3.19}$$
$$1 \leq t \leq T-1,\ 1 \leq j \leq N$$

An initial condition of recursion 3.19 is:

$$\beta_j(T) = 1, \quad 1 \leq j \leq N \tag{3.20}$$

Hence the conditional probability $p(\mathbf{O}, s_t = j | \mathcal{W}, \mathcal{M})$ can be calculated as:

$$p(\mathbf{O}, s_t = j | \mathcal{W}, \mathcal{M}) = \alpha_j(t)\beta_j(t) \tag{3.21}$$

3.2.4 An optimal state sequence decoding

The second basic problem for the HMM is to find an optimal state sequence associated with the given observation vectors sequence. There are several

possible optimality criteria: A simple possible optimality criterion is to choose the states s_t, which are the most likely at each discrete time t. This criteria might be applicable for some simple tasks, but the most suitable criterion is to find the one optimal state sequence $\mathbf{s}$ that is, to maximize $p(\mathbf{s}|\mathbf{O}, \mathcal{M})$, which can be interpreted to maximizing $p(\mathbf{O}, \mathbf{s}|\mathcal{M})$. The Viterbi algorithm [Viterbi, 1967] is one of the possible techniques for finding one optimal state sequence. It is based on dynamic programming methods. A detailed discretion of the Viterbi algorithm applied for isolated word recognition will be discussed in this section. A description of the Viterbi decoding within continuous speech recognition will be given in section 3.2.8.

3.2.4.1 Viterbi algorithm

To find one optimal state sequence $\mathbf{s} = [s_1, s_2, \ldots, s_T]$, for some observation vectors sequence $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$, we have to define the maximum likelihood variable $\chi_j(t)$ of the partial observation vectors sequence $[\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t]$ and an active state j at the discrete time t:

$$\chi_j(t) = \max_{\forall s_1, s_2, \ldots, s_{t-1}} p(s_1, s_2, \ldots, s_{t-1}, s_t = j, \mathbf{O}|\mathcal{W}, \mathcal{M}) \tag{3.22}$$

Take into account dynamic programming principles (DPP) [Bellman, 1957], [Bertsekas, 2000], to find the optimal state sequence from discrete time 1 to discrete time $t+1$ any intermediate state must be the optimal state (local optima) within the optimal partial state sequences before and after that state. As the result of the DPP, we can express $\chi_j(t+1)$ by the induction:

$$\chi_j(t+1) = \left\{ \max_{1 \leq i \leq N} \chi_j(t) a_{ij} \right\} b_j(\mathbf{o}_{t+1}) \tag{3.23}$$

To determine an optimal state sequence we need an additional variable $\psi_t(j)$ to store the argument that maximized equation 3.23. The algorithm of finding an optimal state sequence can be presented as follows:

- *Initialization*

$$\chi_i(1) = \pi_i b_i(\mathbf{o}_1), \quad 1 \leq i \leq N \tag{3.24a}$$
$$\psi_1(i) = 0, \quad 1 \leq i \leq N \tag{3.24b}$$

- *Recursion*

$$\chi_j(t) = \max_{1 \le i \le N} \{\chi_i(t-1)a_{ij}\}\, b_j(\mathbf{o}_t) \tag{3.25a}$$
$$2 \le t \le T, \quad 1 \le j \le N$$
$$\psi_t(j) = \arg\max_{1 \le i \le N} \{\chi_i(t-1)a_{ij}\} \tag{3.25b}$$
$$2 \le t \le T, \quad 1 \le j \le N$$

- *Termination*

$$\hat{s}_T = \arg\max_{1 \le i \le N} \{\chi_i(T)\} \tag{3.26}$$

- *State sequence backtracking*

$$\hat{s}_t = \psi_{t+1}(\hat{s}_{t+1}) \tag{3.27}$$
$$t = T-1, T-2, \ldots, 1$$

The Viterbi algorithm is almost similar (backtracking step is an exception) in realization to the forward variable estimation 3.15 - 3.17 within forward-backward algorithm. The main difference is the maximization in equation 3.25a instead the summing in equation 3.15.

3.2.5 Maximum likelihood training

Maximum likelihood (ML) training is the most often used approach for estimation of the HMM parameters. The main task is to compute the model parameters that maximize the likelihood of the observation vectors sequence given the defined transcriptions and the model parameters. The general ML criterion can be expressed as:

$$\hat{\mathcal{M}}_{ML} = \arg\max_{\mathcal{M}} p(\mathbf{O}|\mathcal{W}, \mathcal{M}) \tag{3.28}$$

Where $\mathcal{W}$ is the defined training word sequence (or sub-word unit level transcription), $\mathcal{M}$ is the HMM parameter set.

It is often more convenient to maximize the logarithm of the likelihood function in order to decrease required computational power. In this case equation 3.28 can be expressed as:

$$\hat{\mathcal{M}}_{ML} = \arg\max_{\mathcal{M}} \log p(\mathbf{O}|\mathcal{W}, \mathcal{M}) \tag{3.29}$$

One possible solution for maximum likelihood training task is an *expectation maximization* (EM) algorithm.

3.2.5.1 Expectation maximization algorithm

The *expectation maximization* (EM) is a general statistic method of finding the maximum likelihood estimate of the parameters of an underlying distribution from a given data set when the data is incomplete or has missing values.

The EM algorithm has two main applications: The first takes place when the data has some missing values, due to problems with or restrictions of the observation process. The second takes place when optimizing the likelihood function is analytically quite difficult but when the likelihood function can be simplified by assuming the presence of and values for additional but hidden or missing parameters. The second case is more common in the computational pattern recognition field [Bilmes, 1998].

The EM algorithm is a well-known method of finding maximum likelihood estimates of parameters in various statistical models. The Baum-Welch algorithm [Baum et al., 1970] is a prominent instance of Expectation Maximization algorithm.

The basic idea of the algorithm is to iteratively compute the maximum likelihood estimation when the observations can be considered as incomplete data. Each iteration of the algorithm includes an expectation step followed by a maximization step. The term "incomplete data" implies the existence of two sample spaces X and Y. We assumed that observation feature vectors x are realization from X. The corresponding state sequences y in Y are not observed directly, but only indirectly through observation feature vectors x. We suppose that a complete data set exists $Z = (X, Y)$. Then the joint density function $p(z|\mathcal{M})$ can be specified as:

$$p(z|\mathcal{M}) = p(x, y|\mathcal{M}) = p(y|x, \mathcal{M})p(x|\mathcal{M}) \tag{3.30}$$

First, the EM algorithm finds the expected value of the complete data set log-likelihood $log\, p(X, Y|\mathcal{M})$ with respect to the hidden data Y given the observed data X and the actual parameters estimates. We can define the following auxiliary function $Q(\mathcal{M}, \hat{\mathcal{M}}_{k-1})$:

$$Q(\mathcal{M}, \hat{\mathcal{M}}_{k-1}) = E\left[log\, p(X, Y|\mathcal{M})|X, \hat{\mathcal{M}}_{k-1}\right] \tag{3.31}$$

Where $\hat{\mathcal{M}}_{k-1}$ are the actual parameters estimates that we used to estimate the expectation and $\hat{\mathcal{M}}_k$ are the new parameters that we optimize to increase the auxiliary function Q.

To find the optimal parameters estimates, two main steps are taken:

- **Expectation:** The evaluation of the auxiliary function $Q(\mathcal{M}, \hat{\mathcal{M}}_{k-1})$. The first argument $\mathcal{M}$ represents the parameters estimates that will be optimized in an attempt to maximize the likelihood [Bilmes, 1998].

The second argument $\hat{\mathcal{M}}_{k-1}$ represents the current parameters estimates that have available to estimate the expectation.

- **Maximization:** The next step of the EM algorithm is to maximize the expectation we computed in the previous step:

$$\hat{\mathcal{M}}_k = \arg\max_{\mathcal{M}} Q(\mathcal{M}, \hat{\mathcal{M}}_{k-1}) \tag{3.32}$$

This is the reason why the algorithm is called *expectation maximization* (EM) algorithm.

3.2.6 Parameters re-estimation

To describe the iterative process for re-estimation of HMM parameters we first define variables $\xi_{ij}(t)$ and $\gamma_j(t)$. The variable $\xi_{ij}(t)$, is defined the probability being an active state i at the discrete time t, and state j at the discrete time $t+1$:

$$\xi_{ij}(t) = p(s_t = i, s_{t+1} = j | \mathbf{O}, \mathcal{W}, \mathcal{M}) \tag{3.33}$$

From the definitions of forward and backward variables, we can express $\xi_{ij}(t)$ as:

$$\begin{aligned} \xi_{ij}(t) &= \frac{p(s_t = i, s_{t+1} = j, \mathbf{O} | \mathcal{W}, \mathcal{M})}{p(\mathbf{O} | \mathcal{W}, \mathcal{M})} \\ &= \frac{\alpha_i(t) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_j(t+1)}{p(\mathbf{O} | \mathcal{W}, \mathcal{M})} \\ &= \frac{\alpha_i(t) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_j(t+1)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_j(t+1)} \\ &= \frac{\alpha_i(t) a_{ij} b_j(\mathbf{o}_{t+1}) \beta_j(t+1)}{\sum_{i=1}^{N} \alpha_i(t) \beta_i(t)} \end{aligned} \tag{3.34}$$

The variable $\gamma_j(t)$, is defined as:

$$\gamma_j(t) = p(s_t = j | \mathbf{O}, \mathcal{W}, \mathcal{M}) \tag{3.35}$$

It is the probability of being active state j at the discrete time t, given observation vectors sequence $\mathbf{O}$, the word sequence hypothesis $\mathcal{W}$, and the model $\mathcal{M}$. We can calculate $\gamma_j(t)$ in such a way:

$$\begin{aligned}\gamma_j(t) &= p(s_t = j|\mathbf{O}, \mathcal{W}, \mathcal{M}) \\ &= \frac{p(\mathbf{O}, s_t = j|\mathcal{W}, \mathcal{M})}{p(\mathbf{O}|\mathcal{W}, \mathcal{M})} \\ &= \frac{p(\mathbf{O}, s_t = j|\mathcal{W}, \mathcal{M})}{\sum_{i=1}^{N} p(\mathbf{O}, s_t = i|\mathcal{W}, \mathcal{M})}\end{aligned} \tag{3.36}$$

Bu using equation 3.21, we can express $\gamma_t(j)$ as:

$$\gamma_j(t) = \frac{\alpha_j(t)\beta_j(t)}{\sum_{i=1}^{N} \alpha_i(t)\beta_i(t)} \tag{3.37}$$

Re-estimation formulas for HMM parameters $\hat{\mathcal{M}} = (\hat{\pi}, \hat{\mathbf{A}}, \hat{\mathbf{B}})$ can be derived by evaluating equation 3.32. By using variables $\xi_{ij}(t)$ and $\gamma_j(t)$, we can express re-estimation formulas as:

$$\hat{\pi}_j = \gamma_j(1) \tag{3.38a}$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_{ij}(t)}{\sum_{t=1}^{T-1} \gamma_i(t)} \tag{3.38b}$$

GMM is the most popular type of continuous density function within continuous HMM. To calculate parameters of the observation generation continuous density function $b_{jm}(\mathbf{o}_t)$, expressed in equation 3.13, we should define a variable $\gamma_{jm}(t)$. The Gaussian component posterior variable $\gamma_{jm}(t)$ is related to the m-th Gaussian component, and the active state j can be estimated by:

$$\gamma_{jm}(t) = \frac{c_{jm} b_{jm}(\mathbf{o}_t)\beta_j(t) \sum_{i=1}^{N} \alpha_i(t-1)a_{ij}}{\sum_{i=1}^{N} \alpha_i(t)\beta_i(t)} \tag{3.39}$$

where $b_{jm}(\mathbf{o}_t)$ is the D-dimensional multivariate Gaussian distribution with the following parameters $\mathcal{N}\left(\mathbf{o}_t, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}\right)$.

The re-estimation equation for GMM parameters for an active state j are given by [Yu, 2006]:

$$\hat{c}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t)}{\sum_{m=1}^{M_j} \sum_{t=1}^{T} \gamma_{jm}(t)} \tag{3.40a}$$

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \mathbf{o}_t}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{3.40b}$$

$$\hat{\boldsymbol{\Sigma}}_{jm} = \begin{pmatrix} \frac{\sum_{t=1}^{T} \gamma_{jm}(t)(o_t^1 - \hat{\mu}_{jm}^1)^2}{\sum_{m=1}^{M_j} \gamma_{jm}(t)} & 0 & \dots & 0 \\ 0 & \frac{\sum_{t=1}^{T} \gamma_{jm}(t)(o_t^2 - \hat{\mu}_{jm}^2)^2}{\sum_{m=1}^{M_j} \gamma_{jm}(t)} & 0 \dots & 0 \\ & \dots\dots\dots\dots\dots\dots\dots\dots & & \\ 0 & \dots\dots\dots & 0 & \frac{\sum_{t=1}^{T} \gamma_{jm}(t)(o_t^{M_j} - \hat{\mu}_{jm}^{M_j})^2}{\sum_{m=1}^{M_j} \gamma_{jm}(t)} \end{pmatrix} \tag{3.40c}$$

The calculation of the full covariance matrix $\hat{\boldsymbol{\Sigma}}_{jm}$ requires a lot of computation power and memory for the second-order statistics. Take into account, that most ASR systems are using a large number of Gaussian components, only the estimation of the diagonal elements of covariance matrices are done in equation 3.40c.

3.2.7 Language modeling

A language model is an important source of priory information, namely, the probability of a hypothesized sequence of K words, $\mathcal{W} = w_1, w_2, \dots, w_k$. For each word presented in the vocabulary, the language model defines the list of words that can follow it with associated discrete probability. Those prior discrete probabilities can be factorized into a product of conditional probabilities:

$$\begin{aligned} P(\mathcal{W}) &= P(w_1, w_2, \dots, w_k) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_2, w_1) \dots P(w_k|w_{k-1}, \dots, w_1) \\ &= \prod_{k=1}^{K} P(w_k|w_{k-1}, \dots, w_1) \end{aligned} \tag{3.41}$$

where w_k is the k-th word of the hypothesized word sequence. The estimation of the discrete probability of any word sequence using equation 3.41 demands estimating the probability of all of it is possible complete sequences. In the case of large vocabulary tasks, the number of possible complete sequences is too big. As a result it is hard to provide an accurate estimate of every possible word sequence. N-gram language models is a possible solution for this

problem. This type of language model restricts the length of the complete sequence required to calculate the conditional probability. This method is the most widely used for statistical language modeling in automatic speech recognition. The following simplification of probability estimation of the hypothesized sequence of K words can be expressed as:

$$\begin{aligned} P(\mathcal{W}) &= P(w_1, w_2, \ldots, w_k) \\ &= \prod_{k=1}^{K} P(w_k | w_{k-1}, \ldots, w_1) \\ &\approx \prod_{k=1}^{K} P(w_k | w_{k-1}, \ldots, w_{k-N+1}) \end{aligned} \tag{3.42}$$

where N is the fixed size of word history. N usually has a small value, for example: $N = 2$ so it is called a bigram language model, $N = 3$ is a trigram language model. Taking into account this assumption, it is easy to use the ML estimate for N-gram by using the word sequence frequency counts with length N

$$P(w_k | w_{k-1}, \ldots, w_{k-N+1}) = \frac{f(w_k, w_{k-1}, \ldots, w_{k-N+1})}{f(w_{k-1}, \ldots, w_{k-N+1})}; \tag{3.43}$$

where $f(w_k, w_{k-1}, \ldots, w_{k-N+1})$ indicates the number of times the N-gram word sequence $w_k, w_{k-1}, \ldots, w_{k-N+1}$ appears in the training dataset and $f(w_{k-1}, \ldots, w_{k-N+1})$ is the number of times the $(N-1)$-gram word sequence $w_{k-1}, \ldots, w_{k-N+1}$ appears.

Since the vocabulary of datasets we consider in this thesis is sufficiently limited, we use back-off bigram language models for evaluation of our ASR engine. The bigram language model is a table which includes the probability of a given word being followed by another word. This able is estimated based on a training dataset.

So-called zero-gram model is the simplest language model, which assumes $P(w_k|w_i) = 1$ for all k and i, so that every word from the vocabulary is supposedly capable of being followed by any other word from the vocabulary. Zero-gram language models can be performed as finite state networks, so-called *word networks*. In such a form they can be integrated simply into a recognition decoding process.

For construction of a word network from a specified recognition grammar we used *HParse* tool from HTK 3.4 [Young et al., 2009]. HParse format grammars are an easy way of defining a specific thematic domain grammar for IVR technologies. An example of a recognition grammar in HParse format

```
$simple_object = Ring | Scheibe;
$articles = die | der | den | sil;

$type1 = kleinste | mittlere | mittelgrosse | grosse |
groesste | naechste;
$type2 = kleinsten | mittelgrossen | mittleren |
grossen | groessten | naechsten;

$li_rech = links | rechts;

$num = eins | zwei | drei;
$num2 = erste | zweite | dritte;

$object = $articles $type1 $simple_object |
        $articles $type2 $simple_object |
        $articles $type1 | $articles $type2;

$direction = auf die Nummer $num | auf Nummer $num |
        auf die $num | auf $num | auf Position $num |
        nach $li_rech | $li_rech | nach ganz $li_rech |
        zu $num | in die Mitte | auf die Mitte | zur Mitte;

$action = lege | legen | bewege | setzen |
hinlegen | runterlegen | positionieren ;

$input = $object | $direction | $commands | X ;
(< $input | sil >)
```

Listing 3.1: *Simple Tower of Hanoi task (with 3 disks) grammar*

is presented in listing 3.1. This grammar is suitable for an ASR system for speech-based control within solving a simple logic game "Tower of Hanoi" with 3 disks.

Listing 3.1 shows an example of a grammar for "Tower of Hanoi" game with 3 disks. As can be noticed, the grammar contains the following word groups: *object specification* (simple_object, type1, type2, num, num2, articles, object), *direction specification* (li_rech, num, direction), *action specification* (action) and a so-called *"garbage" model* (X).

The dictionary entry for X would reference out-of-vocabulary (OOV) words or a so-called "garbage" model. The simplest way of "garbage" modeling is to include phonetic transcriptions of the most frequently used task-unrelated words to the X word-related-lexicon entries.

3.2.8 Viterbi decoding and continuous speech recognition

Within recognition, the acoustic score is computed with equation 3.11 which is presented in section 3.2.3. As described in sections 3.2.3.1 and 3.2.3.2, the likelihood $p(\mathbf{O}|\mathcal{W},\mathcal{M})$ can be estimated using the forward-backward algorithm [Baum et al., 1970]. However, it is unpractical for the real-time continuous speech recognition since:

- *backward iteration is needed, hence the whole utterance has to be buffered first*
- *the sum over states takes a lot of time and computational recourses, hence it is approximated by the maximum*

The Viterbi algorithm [Viterbi, 1967], described in section 3.2.4.1, is the most widely used approach in the continuous speech recognition applied to find the single best state sequence that has the highest probability to generate the observation vectors sequences. In such a way, the maximum likelihood of the observation vectors sequence uses only one hidden state sequence to approximate the marginal likelihood over all possible state sequences [Yu, 2006].

$$\begin{aligned} p(\mathbf{O}|\mathcal{W},\mathcal{M}) &= \sum_{\forall \mathbf{s}} p(\mathbf{O},\mathbf{s}|\mathcal{W},\mathcal{M}) \\ &\approx \max_{\forall \mathbf{s}} p(\mathbf{O},\mathbf{s}|\mathcal{W},\mathcal{M}) \end{aligned} \tag{3.44}$$

Taking into account equation 3.25a, the maximum likelihood of the observation vectors sequence can be expressed as:

$$p(\mathbf{O}|\mathcal{W},\mathcal{M}) \approx \chi_N(T) = \max_{1\leq i\leq N} \{\chi_i(T-1)a_{iN}\}\, b_j(\mathbf{o}_T) \tag{3.45}$$

where T is the length of the observation vectors sequence. As one can notice, in equation 3.45 the backward processing is not applied. Hence, real-time processing becomes possible.

The Viterbi algorithm can be applied for isolated word recognition tasks. Continuous speech recognition is a complex task. Since an average continuous-speech-recognition system deals with a huge number of possible word sequences, it is not applicable for such a system to construct a single composite HMM for each potential word sequence. In this case, a Viterbi-beam search with a token passing algorithm [Young, 1995] is usually used.

To understand the complexity of the continuous speech-recognition task, suppose that a branching word network tree is built such that at the start there is a branch to every possible start word. All start words are linked to all

possible following words and so forth. At the end, this branching word network tree will be quite big and represents all of the possible word sequences within a closed thematic domain. After construction of the word network tree, let each word be replaced by the sequence of corresponding phonetic models. In a case of multiple phonetic transcriptions for the same word, these models can be combined in parallel. As one can notice, the constructed branch network is very large. As a consequence, a pruning of the search space is required.

Any path from the start point to some node in the network tree can be presented as a movable *token* placed in the node at the end of the path [Young, S. J. et al., 1989]. The token is characterized by the likelihood of the partial path $\chi_j(t)$ (token score) and a path history. As a starting point of the token passing algorithm, a single token is set in the start node of the network tree. At each discrete time, tokens are duplicated in connected HMM states or connected network tree nodes and their scores are re-estimated. Within the words transaction, the language model score is added to the corresponding token score. When the last observation vector is processed, the token with the highest score is traced back to show the most likely sequence of HMMs and corresponding lexical interpretation.

3.2.9 Adaptation techniques in ASR

The training approaches described earlier use an assumption that training and test datasets have similar acoustic characteristics (speaking rate, acoustic environment, vocal tracts variability, emotional speech, etc.). However, in real-life applications, it is usually not the case. The acoustic characteristics mismatch may significantly decrease the recognition performance compared to the ASR systems build on data with matched acoustic characteristics. To compensate the mismatch of acoustic characteristics between test and training datasets, adaptation techniques are usually applied. A simplified schema of the speaker adaptation technique as used in HMM-based speech-recognition models is presented in Figure 3.5.

As one can see from Figure 3.5, adaptation techniques use information provided in an adaptation material to adjust the HMM/GMM parameters (i.e. mean and diagonal elements of the covariance matrix (variance) of the multivariate Gaussian mixture models) of the basic model to reflect specific acoustic characteristics (acoustical environment, speaker-dependent modeling, etc.). In our research we use adaptation approaches for compensation the mismatch of acoustic characteristics between neutral speech samples and affective speech material.

One of the most popular adaptation techniques applied within ASR systems are model-based transforms: *Maximum Likelihood Linear Regression*

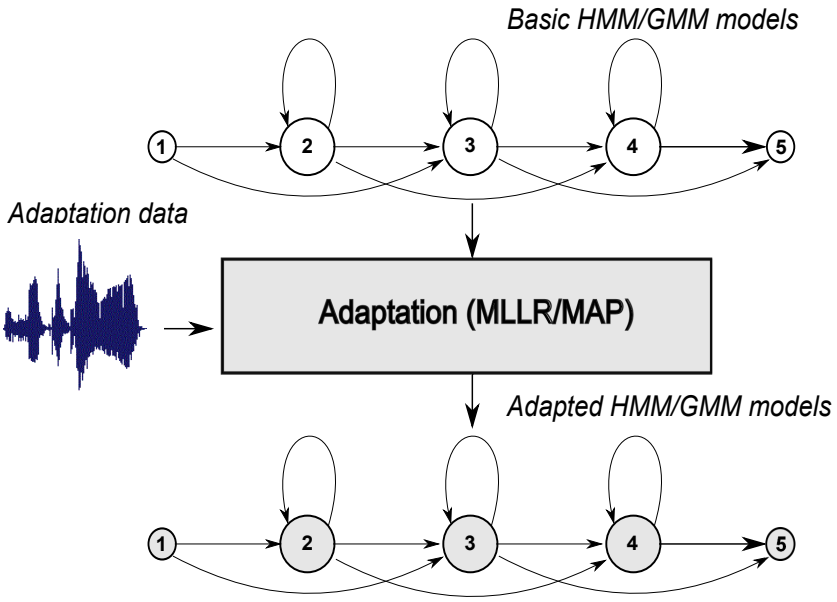


Figure 3.5: *General structure of an adaptation ASR models*

(MLLR) and *Maximum a Posteriori (MAP)*. The Maximum Likelihood Linear Regression (MLLR) and Maximum a Posteriori (MAP) adaptation techniques will be described in this section.

3.2.9.1 Maximum a Posteriori (MAP) Adaptation

The Maximum a Posteriori (MAP) [Gauvain and Lee, 1994] approach (sometimes referred as the Bayesian adaptation) maximizes the posteriori probability using a prior HMM parameter distribution.

$$\hat{\mathcal{M}}_{MAP} = \arg\max_{\mathcal{M}} \{\, p(\mathbf{O}|\mathcal{W}, \mathcal{M})\, p(\mathcal{M}|\mathbf{O}_{trn}, \mathcal{W}_{trn})\} \tag{3.46}$$

where $p(\mathcal{M}|\mathbf{O}_{trn}, \mathcal{W}_{trn})$ is the prior distribution of the HMM models parameters estimated on training data $\mathbf{O}_{trn}$ and $\mathcal{W}_{trn}$.

To evaluate the HMM model parameter estimate using the MAP transformation, an iterative EM algorithm is applied. If the prior mean estimate for state j and Gaussian mixture component m is $\tilde{\boldsymbol{\mu}}$, then the MAP estimate for the adapted mean of the m Gaussian mixture component $\hat{\boldsymbol{\mu}}_{jm}$ can be expressed as:

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{\tau\tilde{\boldsymbol{\mu}}_{jm} + \sum_{t=1}^{T} \gamma_{jm}(t)\mathbf{o}_t{}^{ad}}{\tau + \sum_{t=1}^{T} \gamma_{jm}(t)} \tag{3.47}$$

where τ is a hyper-parameter which regulates the balance between the maximum likelihood estimate of the mean value and its prior value; ${\mathbf{o}_t}^{ad}$ is the adaptation observation feature vector at the discrete time t; $\gamma_{jm}(t)$ is the m Gaussian component of the probability of being active state j at the discrete time t. Usually the hyper-parameter is in the range $2 \leq \tau \leq 20$.

The MAP adaptation requires more adaptation data to be present. When the amount of adaptation data increases, so the MAP estimate converges to the maximum likelihood estimate. If sufficient amount of adaptation data become available, the MAP approach begins to perform better than the MLLR.

3.2.9.2 Maximum Likelihood Linear Regression (MLLR)

The Maximum Likelihood Linear Regression (MLLR) is the best-known linear transformation method applied for speaker adaptation. It uses the ML criterion to estimate a linear transformation which may be applied to adapt Gaussian parameters of HMMs.

$$\hat{\boldsymbol{\mu}}_m = \mathbf{A}\boldsymbol{\mu}_m + \mathbf{b} = \mathbf{W}\boldsymbol{\xi}_m \tag{3.48}$$

where $\hat{\boldsymbol{\mu}}_m$ is the MAP estimate for the adapted mean of the m Gaussian mixture component; ξ_m is an extended mean vector $\xi_m = [1\, {\boldsymbol{\mu}_m}^T]$ and $\mathbf{W} = [\mathbf{b}\,\mathbf{A}]$

Equation 3.48 can be deconstructed as follows:

$$\hat{\mathbf{w}}_d = {\mathbf{G}_d}^{-1}\mathbf{k}_d \tag{3.49a}$$

$$\mathbf{G}_d = \sum_{m=1}^{M_j}\sum_{t=1}^{T}\frac{\gamma_m(t)}{\sigma_{m,dd}}\boldsymbol{\xi}_m\,{\boldsymbol{\xi}_m}^T \tag{3.49b}$$

$$\mathbf{k}_d = \sum_{m=1}^{M_j}\sum_{t=1}^{T}\frac{\gamma_m(t)\,o_{t,d}}{\sigma_{m,dd}}\boldsymbol{\xi}_m \tag{3.49c}$$

where matrix elements $\hat{\mathbf{w}}_d$ construct the matrix $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_D]^T$, $o_{t,d}$ is the d-th feature value from observation feature vector $\mathbf{o}_t$; $\sigma_{m,dd}$ is the d-th diagonal element of covariance matrix $\boldsymbol{\Sigma}_m$.

3.2.9.3 Base class specifications

In the previous section we described the MLLR adaptation technique. Specifying the set of the HMMs which share the same transformation is the first requirement to allow adaptation. One of the possible specifications is achieved

```
~b ''global''
<MMFIDMASK> Kiel*
<PARAMETERS> MIXBASE
<NUMCLASSES> 1
 <CLASS> 1 {*.state[2-4].mix[1-18]}
```

Listing 3.2: *Global base class (GBC) specification*

using a base class. For base class definitions, the HMMs must always be specified. A global transformation for all HMMS is the simplest form of transformation used for adaptation. An example of a base class specification for the global transformation can be found in listing 3.2

The base class specified in listing 3.2 defines a global transformation for HMMs which contain up to 3 emitting states and up to 18 Gaussian mixture components per state.

With base classes specification it is possible to define several classes of HMMs. An example of a base class specification with three classes can be found in listing 3.3

```
~b ''global''
<MMFIDMASK> Kiel*
<PARAMETERS> MIXBASE
<NUMCLASSES> 1
 <CLASS> 1 {(sil,sp).state[2-4].mix[1-18]}
 <CLASS> 2 {(a,ai1,at,au1,e,er,e1,i,i1,o,oe,o1,o1y,u,u1,y).
 state[2-4].mix[1-18]}
 <CLASS> 3 {(b,c1,d,f,g,h,j,k,l,m,n,n1,p,r,s,s1,t,v,x,z).
 state[2-4].mix[1-18]}
```

Listing 3.3: *Three base classes specification*

The base class specified in listing 3.3 defines three different classes: *class1* which represents long and short pauses, *class2* which represents vowels, and *class3* which represents consonants. Also, the HMMs could be grouped into the broad phone classes: *silence, vowels, stops, glides, nasals* and *fricatives,* etc. [Gales, 1996].

These base classes can be used to define which HMMs share a separate transformation. A more general approach based on a regression class trees will be described in the next section.

3.2.9.4 Regression classes tree scheme

To make an adaptation process more flexible it is possible to specify the convenient set of base classes according to the amount of adaptation material that is obtainable. The global adaptation transformation presented in the previous section can be used when a small amount of adaptation material is available. As more adaptation material becomes available, increasing the number of base classes for advanced adaptation is possible. For each base class we use a different transformation.

Instead defining static HMMs classes, it is possible to use a dynamic method for the generation of further transformations as more adaptation material becomes available. A *regression class tree* [Gales, 1996] is used to group Gaussian components so that the number of transformations to be estimated can be dynamically selected according to the amount of available adaptation material. Automatic clustering of Gaussian components which are similar in acoustic space is used for constructing the regression class tree. The regression class tree should be extracted before adaptation.

3.3 Construction of robust ASR models for German spontaneous affective speech

In this section we present the main aspects of developing German spontaneous affective speech-recognition methods: *sub-word units selection and lexicon construction, German phonetic pattern, spontaneous speech variability, comparison of affective and neutral speech* and *Emotional speech acoustic modeling.*

3.3.1 Emotional neutral German speech dataset

For a natural speech corpus we used part of The Kiel corpus of Read Speech [KIE, 2002]. The Kiel Corpus is a growing collection of read and spontaneous German speech which has been collected and labeled segmentally since 1990. For our evaluation, we used speech samples from 6 female and 6 male speakers. The list of speakers is k01,...,k10, k61 (also defined as kko), k62 (also defined as rtd). To reach a qualitative acoustic parameters estimation, selected material from Kiel's read speech corpus were manually freed from technical noise and breathing. 1041 utterances for female speakers and 1033 utterances for male speakers were used for our experiments presented in this chapter. The number of vowel instances presented in selected material from the Kiel dataset can be found in Table 3.3 on page 74.

3.3.2 Sub-word units selection and lexicon construction

In the real-life application, it is not possible to obtain sufficient training data for each individual word which can occur during a natural human-machine interaction. The possible solution to this problem is to use HMMs to model sub-word units, rather than the whole word included in the vocabulary. The phoneme is the smallest acoustic component of speech and it is widely used as the sub-word unit for an automatic speech-recognition task. The main advantage of using phonemes as the sub-word unit is that there is a standard set of phonetic rules to map words to phonemes. In such a way, words can be represented as a sequence of phonemes. The number of phonemes is usually considerably smaller than the number of words in a vocabulary. In a state-of-the-art ASR system used in this work, we use 39 distinct German phonemes (modified compact SAM-PA list). German phonetic pattern used in our ASR system will be described in detail in the next section.

To map the word sequence to a phonetic sequence we require a lexicon. The lexicon, also referred to as the dictionary, is a standard part in an ASR system. The dictionary maps phonetic units, from which the acoustic models are built, to the present words included in the vocabulary and language model. The training and recognition processes are executed at the phonetic units level. Finally, within the recognition process, the phonetic units sequence is transformed back to the word sequence. It is common to use two different lexicons within he same ASR system. The first is responsible for mapping the word sequence to the unique phonetic sequence within the training process, and it contains only one possible phonetic transcription for each word. The second extracts the word sequence from phonetic sequences within the recognition process, and it supports variable phonetic transcriptions for each word included in the vocabulary.

Two main types of phoneme unit sets are widely used in modern ASR systems: context-independent phonemes, namely *mono-phones*, and context-dependent phonemes, such as: *bi-phones, tri-phones*, and *quin-phones*. With a mono-phones set, we do not take into account the context of each particular phoneme. Still, due to the co-articulation effect, the articulation of most phonemes is highly dependent on their neighboring phonemes. The most common context-dependent phoneme unit sets are tri-phones. For example, with 39 phonemes there are $39^3 = 59319$ possible tri-phones, but not all of them can have a place due to the phonotactic constrains of the German language. To train robust tri-phones-based ASR models we need more data in comparison to the mono-phones. Also this data should be well-annotated, because each annotation error will have a triple effect in comparison to the mono-phone-based model. To the best of our knowledge, to date there is no

publicly available corpus for the German language which can provide a sufficient amount of training material with a high-standard phonetic transcription which can be used for effective tri-phones-based HMM modeling. For example, Kiel, SmartKom, Verbmobil databases do not provide detailed transcription of paralinguistic cues, also lexicons attached to these corpora contain a lot of incorrect phonetic transcriptions and do not provide lists of all possible pronunciation forms. An example of incomplete phonetic transcription of German word "Abend" will be described in the next section. In the case of tri-phone HMM models each incorrect phoneme will cause us threefold incorrect modeling. Take into account sparse amount of instances for some tri-phones this threefold error could be crucial. As a result, we use the mono-phone set for our ASR system.

3.3.2.1 German phonetic pattern

The number of phonetically distinguishable phonemes in a language is often a matter of judgment. Table 3.2 and Table 3.1 present lists of German vowels and consonants, their corresponding IPA and SAM-PA symbols [SAM, 1996]. There are 39 phonemes in the German language, including 13 *unreduced vowels*, 2 *reduced vowels*, 3 *diphthongs*, 6 *plosive consonants*, 9 *fricative consonants*, 3 *nasal consonants*, and 2 *liquid consonants*.

The German language contains a standard set of strict phonetic rules to map words to phonemes. The amount of these rules and exceptions are significantly smaller in comparison with English. Still there is no rule-based grapheme-to-phoneme (G2P) open-source toolkit available for the German speech processing research community. There is a data-driven G2P open-source toolkit [Bisani and Ney, 2008] available, but this method requires a huge amount of training material to train reliable models. Also, it is not able to generate reliable phonetic transcription alternatives for words which can be pronounced in different ways.

It is also possible to use existing German lexicons included in publicly available corpora (Kiel, SmartKom, Verbmobil). Still, there are some oversights in existing German lexicons. For example, in phonetics transcriptions dictionary Duden 6 "Das Aussprachewörterbuch" [Mangold, 1990] the word "Abend" is transcribed as ['a:bnt]. It is the so-called "hochdeutsch" pronunciation standard. On the other hand, Kiel lexicon contains slightly different transcription ['a:bənt]. Both versions are acceptable for colloquial German language. Adequate lexicons included in corpora should contain both variations of transcription, which is not the case with current publicly available German speech databases. Hence, even existing lexica need further refinement before they can be used.

However, it is possible to determine the actual pronunciations used in the utterances used to train ASR model with *forced alignment*. Force alignment is presented in HTK [Young et al., 2009] toolkit. It is a technique which can generate the words and phonemes boundaries on utterance-level based on textual transcriptions of the corresponding utterance and reliable mono-phone HMM models.

3.3.2.2 Consonants

There are few classes of consonant present in German language: *plosives, fricatives, nasals, liquids* [Pompino-Marschall, 1992]. Those classes specify physical characteristics of the generation process. The list of all German consonants with their corresponding class description are presented in Table 3.1.

IPA name	IPA symbol	SAM-PA symbol	IPA name	IPA symbol	SAM-PA symbol
Plosives					
Lower-case P	**p**	**p**	Lower-case B	**b**	**b**
Lower-case T	**t**	**t**	Lower-case D	**d**	**d**
Lower-case K	**k**	**k**	Lower-case G	**g**	**g**
Fricatives					
Lower-case F	**f**	**f**	Lower-case V	**v**	**v**
Lower-case S	**s**	**s**	Lower-case Z	**z**	**z**
Esh	**ʃ**	**S**	Yogh	**ʒ**	**Z**
C Cedilla	**ɕ**	**C**	Lower-case J	**j**	**j**
Lower-case X	**x**	**x**	Lower-case H	**h**	**h**
Nasals					
Lower-case M	**m**	**m**	Lower-case N	**n**	**n**
Eng	**ŋ**	**N**			
Liquids					
Lower-case L	**l**	**l**	Lower-case R	**r**	**r**

Table 3.1: *German Consonants*

For our ASR engine based on mono-phones HMM we used all of the consonants presented in Table 3.1. Some of the SAM-PA IDs have been changed to enable the use of the HTK [Young et al., 2009] toolkit for ASR modeling. Converting non-acceptable SAM-PA IDs will be described later in this section.

3.3.2.3 Vowels

Most existing ASR systems rely heavily on robust vowel recognition to reach a high performance. The vowels acoustic segments are usually long in duration (in comparison to consonants) and are spectrally well represented. As such,

IPA name	IPA symbol	SAM-PA symbol	IPA name	IPA symbol	SAM-PA symbol
Unreduced					
Lower-case A	**a (aː)**	**a (a:)**	Slashed O	**ʌ (ʌː)**	**2 (2:)**
Lower-case E	**e (eː)**	**e (e:)**	O-E Digraph	**ɘ**	**9**
Epsilon	**ɛ (ɛː)**	**E (E:)**	Lower-case U	**uː (uː)**	**u (u:)**
Lower-case I	**i (iː)**	**i (i:)**	Upsilon	**ʊ**	**U**
Small Capital I	**ɪ**	**I**	Lower-case Y	**y (yː)**	**y (y:)**
Lower-case O	**o (oː)**	**o (o:)**	Small Capital Y	**ʏ**	**Y**
Open O	**ɔ**	**O**			
Reduced					
Schwa	**ə**	**@**	Turned A	**ɒ**	**6**
Diphthongs					
Lower-case A, Small Capital I	**aɪ**	**aI**	Open O, Small Capital Y	**ɔʏ**	**OY**
Lower-case A, Upsilon	**aʊ**	**aU**			

Table 3.2: *German vowels. The symbol ":" corresponds to the Length Mark*

they are generally reliably and easily recognized by human beings and by ASR systems [Rabiner and Juang, 1993].

There are 18 vowels in the German phonetic alphabet [Pompino-Marschall, 1992]. Three different classes of vowels (*unreduced, reduced, diphthongs*) and their representatives SAM-PA and IPA symbols can be found in Table 3.2

For our ASR engine based on mono-phones HMM we used all of the vowels (*unreduced*, *reduced*, *diphthongs*) presented in Table 3.1. Some of the SAM-PA IDs have been changed to enable the use of HTK [Young et al., 2009] toolkit for ASR modeling. Converting non-acceptable SAM-PA IDs will be described later in this section.

There are several ways to classify and characterize vowels, including the typical articulatory configuration required to produce the sounds, typical spectral representation, etc. In 1952, Gordon Paterson and Harold Barney [Paterson and Barney, 1952] created a classic plot of measured values of the first (F1) and second (F2) formant for 10 English vowels spoken by a wide range of male and female talkers. They proposed to represent each vowel by a centroid in the formant space.

Instead of representing of each vowel by a centroid, we represent each vowel by the means of the average F1 and F2 values. In Figure 3.6 one can see German vowels mapped into F1/F2 space and the outline of the general vowel triangle for male and female speakers which are included in selected material from Kiel read speech corpus [KIE, 2002]. To reach a qualitative acoustic parameters estimation, selected material from Kiel read speech corpus were

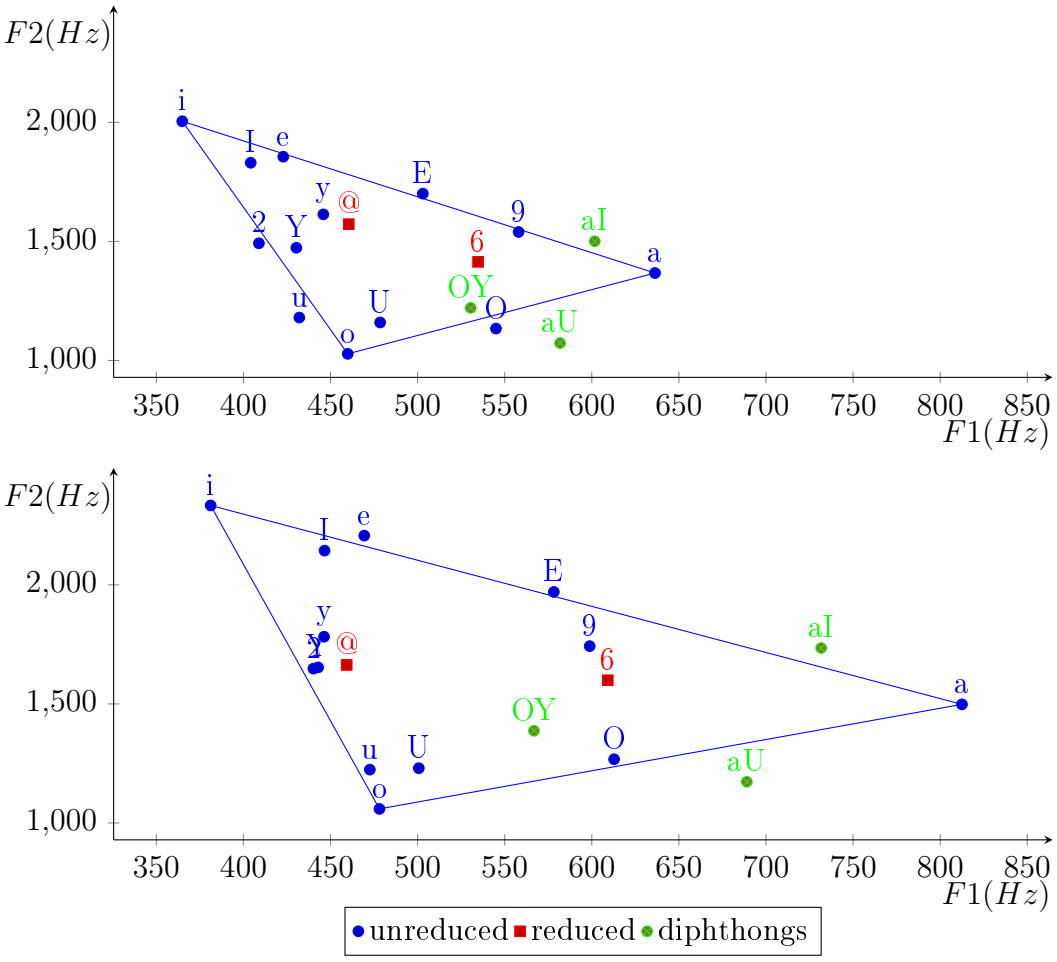


Figure 3.6: *The vowel triangle with mean values positions of the all German vowels. Male speakers (top), female speakers (bottom)*

manually freed from technical noise and breathing.

On the vowel triangles presented in Figure 3.6, one can see an absolute and relative position of 13 unreduced, 2 reduced and 3 diphthongs in the first ($F1$) and second ($F2$) formants space. The vowel triangle represents the extremes of formant location in the F1/F2 space, as represented by [i] (low F1, height F2), [o] (low F1, low F2), [a] (height F1, middle F2), with the other vowels appropriately disposed with respect to the triangle sides and vertices. As one can see the relative position of vowels within the vowel triangle are relatively stable for both genders. Still, female speakers use the larger frequency scale intervals during vowels articulation $381.2\,\text{Hz} \leq F1 \leq 812.5\,\text{Hz}$ and $1,059.1\,\text{Hz} \leq F2 \leq 2,333.5\,\text{Hz}$ in contrast to the male speakers $364.9\,\text{Hz} \leq F1 \leq 636.3\,\text{Hz}$ and $1,073.0\,\text{Hz} \leq F2 \leq 2,004.7\,\text{Hz}$.

3.3.2.4 Diphthongs

A diphthong is a gliding monosyllabic speech sound, and it refers to two adjacent vowel sounds occurring within the same syllable. There are three diphthongs in German, namely [aI] (as in "zwei"), [aU] (as in "Bauch"), [OY] (as in "neun"). Diphthongs are generated by varying the vowel tract shape smoothly between vowel shapes that are appropriate to the diphthong. This non-trivial smoothing produces a new set of vocalized phonemes. In support of the complexity of smoothing one can see that a diphthong could not be represented as a linear combination of compound vowels, see Figure 3.6.

3.3.2.5 HTK format lexicon generation

To be able to use lexicon encoded in extended SAM-PA symbols for an HTK-base [Young et al., 2009] ASR system we should provide some modification of the lexicon files. First of all, HTK do not allow the use of symbols like [@], ['] for the HMM specification. Also, vowels with an additional symbol [:] (Length Mark) can be replaced with corresponding vowels without a length mark. It can be done due to the robust dynamic HMM modeling of the temporal characteristics of phonemes.

The transformed HTK compatible lexicon format will be used for our speech-recognition experiments and for our ASR system integrated in NIMITEK [Wendemuth et al., 2008] demonstrator. More details about NIMITEK demonstrator can be found in Chapter 6.

3.3.3 Spontaneous speech variability

The speech signal not only represents the linguistic content but also a lot of additional information about the speaker: *age, gender, social status, accent (foreign accent, dialects, etc.), emotional state, health, level of reliability,* etc. Characterization of the influence of some of these speech signal variations, together with related methods to improve ASR performance, is an important research field [Benzeghiba et al., 2007].

It is possible to assign three main classes of effects caused by the spontaneous speech variability. The first is the modification of the voice quality by *physiological* or *behavioral* factors. The second is the long-term modulation of the voice for transmission of non-emotional high level information events like *emphasizing* or *questioning*. The third is *pronunciation* variability like foreign accents, dialects, and colloquial speech. A detailed description of spontaneous speech characteristics has been presented in section 2.3.

3.3.3.1 Comparison of affective and neutral speech

For the comparison of affective and neutral speech, vowel triangles have been estimated for selected EMO-DB's [Burkhardt et al., 2005] utterances. We used utterances which represent low-arousal emotions (*boredom*, *sadness*), *neutral*, and high-arousal emotions (*anger*, *fear*, and *joy*).

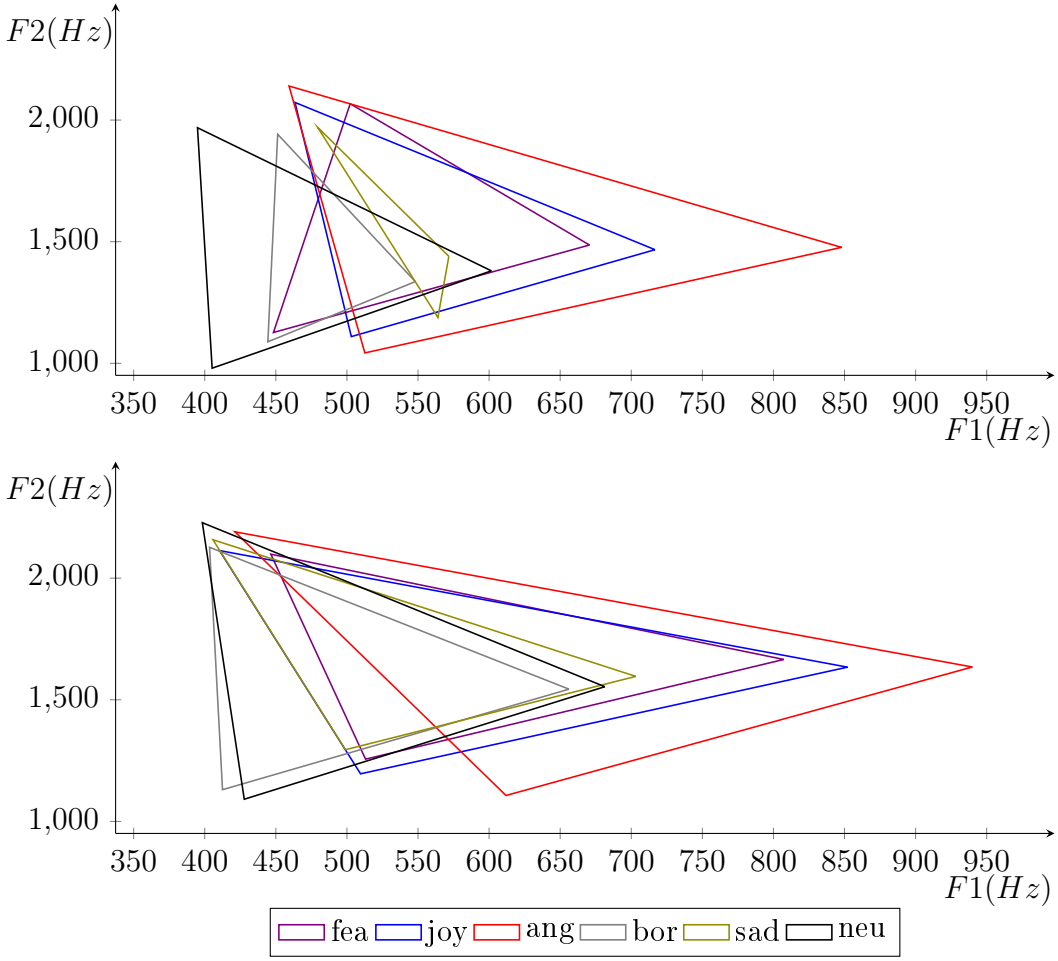


Figure 3.7: *Classical vowel triangle form for different speaker's emotional states. Male speakers (top), female speakers (bottom)*

As one can see from Figure 3.7, the vowel triangles form and their position are different for different emotional states of the speaker. This variability is one of the reasons why ASR models trained on neutral speech are not able to provide a reliable performance in affective speech recognition. Adaptation on affective speech samples of the acoustic model will be presented in the next section.

3.3.4 Emotional speech acoustic modeling

The simplest way to achieve emotional speech acoustic modeling for reliable ASR performance is to train acoustic models for each possible user's emotional states. Training emotional speech acoustic models for each possible user emotional state is not feasible because collecting affective speech in large enough amounts to train a robust ASR acoustic model is quite an expensive and time-consuming process. Nevertheless, due to the pronunciation pattern similarity of affective and neutral speech, emotion-specific characteristics can be captured from existing emotional speech corpora within adaptive transformation of model parameters of the initial neutral speech model to obtain an emotional speech acoustic model.

For the neutral speech ASR model we used mono-phones HMM trained on selected material from Kiel read speech corpus. For adaptation on affective speech samples we used material from the EMO-DB [Burkhardt et al., 2005] database. Vowels can be reliably and easily recognized by human beings and by ASR systems [Rabiner and Juang, 1993]. The total amount of vowel instances presented in selected speech datasets are presented in Table 3.3. An interpretation of the emotional class name abbreviations can be found in Table 2.3 on page 24.

#	EMO-DB						Kiel
	fear	**joy**	**anger**	**boredom**	**sadness**	**neutral**	**read**
a	144	172	348	211	148	207	3357
e	74	80	166	100	59	105	1239
E	42	55	98	64	46	58	1403
i	73	68	159	89	54	101	1323
I	115	125	244	171	124	146	2315
o	24	24	52	34	22	33	535
O	15	17	40	25	22	24	767
u	4	6	11	9	9	7	674
U	33	42	73	48	31	45	1273
y	12	18	22	14	4	14	363
Y	10	14	30	18	12	16	290
2	0	0	0	0	0	0	188
9	5	7	14	7	6	6	209
@	177	222	436	274	201	254	4340
6	66	66	138	85	49	91	3462
aI	22	25	43	36	22	33	1313
aU	16	15	36	23	15	26	528
OY	5	7	14	7	6	6	289

Table 3.3: *Number of instances per vowel in EMO-DB and Kiel datasets*

As one can see from Table 3.3 acoustic form of the vowel [**2**] is not presented in EMO-DB recordings. Also, EMO-DB speech material contains quite small number of instances for some vowels [**u, 9, OY**]. For adaptation of affective speech samples we used MAP and MLLR adaptation techniques. Within MLLR adaptation we used the following HMMs groups specifications:

- Regression class tree
- Two Base classes: phonemes, silence
- Three Base classes: vowels, consonants, silence

Consequently, we investigated the potency of adapting emotional speech acoustic models for German language and we obtained a considerable performance gain as will be discussed in section 5.2.3.

3.4 Summary

This chapter reviews the automatic speech-recognition methods based on hidden Markov models (HMMs). The feature extraction approach, namely, MFCC is discussed first. The hidden Markov models (HMMs), the most frequently used acoustic models, are then presented. The maximum likelihood (ML) training of HMM parameters and the expectation maximization (EM) algorithm are discussed. In this chapter we presented detailed description of German phonetic patterns which will be used later for detailed phoneme-level emotion recognition. N-gram language models and generation word networks with *HParse* grammar format are described. Extensively used Viterbi decoding for spontaneous speech is presented in detail. Standard adaptation approaches like MAP and MLLR are presented. Results of the evaluation of our German ASR models will be presented in Chapter 5. Methods described in this chapter have been used to create an ASR module integrated in our NIMITEK spoken dialog system prototype.

In the next chapter we will describe different classification techniques applied for automatic emotion recognition from speech. The HMM/GMM models presented in this section will be used for our phoneme-level emotion-recognition methods. Force alignment presented in section 3.3.2.1 will be used in the next chapter for time alignment within phoneme-level emotion classification.

CHAPTER 4

Emotion recognition from speech

Contents

4.1 Introduction

To be able to design a user-centered spoken dialog system, we set up a framework that should be robust enough to detect emotional events within human-machine interaction. In this chapter we offer an overview of existing speech-based emotion-recognition techniques, and discuss acoustic feature sets which are the most informative for emotional events determination. Two different techniques of emotion classification, namely, *static* (turn-level analysis) and *dynamic* (frame-level analysis) are presented. Afterwards, two possible combined emotion-classification methods: *two-stage processing and middle-level fusion* are described. Finally, we compare emotion-recognition performances for unit-specific (context-dependent) and general (context-independent) models.

4.2 An overview of existing methods

Since the beginning of emotional speech processing [Scripture, 1921], [Skinner, 1935], [Fairbanks and Pronovost, 1939], [Williams and Stevens, 1972], [Scherer, 1986], [Whissell, 1989], the usefulness of automatic recognition of emotion in speech seems increasingly agreed given the large amount of applications for

user-centered human-machine interfaces. Most of these expect sufficient robustness, which may not be given yet [Picard, 1997], [Cowie et al., 2001], [Shriberg, 2005], [Lee and Narayanan, 2005], [Schröder et al., 2007], [Wendemuth et al., 2008], [Schröder et al.,], [Zeng et al., 2009]. When evaluating the accuracy of emotion-recognition engines, attainable performances are usually overrated since usually acted, prompt or elicited emotions are considered instead of spontaneous, real-life case emotions, which are harder to recognize.

Speech-based emotion classifiers used in the research publications include a broad variety [Ververidis and Kotropoulos, 2006]. Depending on the type of acoustic feature extraction level, either dynamic analysis [Fernandez and Picard, 2003] for processing on a frame-level or static analysis for higher-level statistical functionals [Ververidis and Kotropoulos, 2004] are established.

Among dynamic analysis, hidden Markov models are dominant (cf., e. g., [Nwe et al., 2003], [Schuller et al., 2003], [Lee et al., 2004], [Vlasenko et al., 2007a]). Also, a "bag-of-frames" approach for multi instance learning is used within dynamic analysis [Shami and Verhelst, 2006]. A rarely used alternative is a dynamic time warping, supporting easy adaptation. Also, dynamic Bayesian network architectures [Lee et al., 2009a] could help to combine features on different time levels as spectral on a frame-level basis and supra-segmental prosodic.

Relative to static analysis, the list of possible classification techniques seems endless: Bayes classifier [Ververidis and Kotropoulos, 2004], multi-layer perceptrons or other types of neural networks [Schuller et al., 2004], Baysian networks [Fernandez and Picard, 2003], [Cohen et al., 2003], Gaussian mixture models [Slaney and McRoberts, 1998], [Lugger and Yang, 2007], random forests [Iliou and Anagnostopoulos, 2009], decision trees [Lee et al., 2009b], k-nearest neighbor distance classifiers [Dellaert et al., 1996], and support vector machines (SVM) [Fernandez and Picard, 2003], [Batliner et al., 2006], [Eyben et al., 2009] are applied most often.

Also, a selection of ensemble techniques [Schuller et al., 2005a], [Morrison et al., 2007] has been used, as bagging, boosting, multi-boosting, and stacking with and without confidence scores. New developing techniques as hidden conditional random fields [Wöllmer et al., 2008], long-short-term-memory recurrent neural networks [Wöllmer et al., 2008], tandem Gaussian mixture models with support vector machines [Kockmann et al., 2009] could further be seen more frequently in near future. Table 4.1 presents the most popular existing classification techniques with representative research publication references.

In the past, within the speech emotion-classification research community, the focus was lain on prosodic features extracted on the turn-level. In particular, these feature sets (from 10–100 features) include durations, intensity

Classifier	Selected reference
Naive Bayes	[Dellaert et al., 1996] , [Batliner et al., 2010], [Metze et al., 2010], [Schuller et al., 2010], [Yildirim et al., 2011]
Bayesian logistic regression	[Lee et al., 2009b]
Decision tree	[Yacoub et al., 2003], [Litman and Forbes, 2003]
Support vector machine	[McGilloway et al., 2000], [Yu et al., 2001], [Yacoub et al., 2003], [Lee et al., 2009b], [Polzehl et al., 2009], [Metze et al., 2010], [Seppi et al., 2010], [Schuller et al., 2009a], [Yildirim et al., 2011]
Linear discriminant classifier	[McGilloway et al., 2000], [Batliner et al., 2000b], [Litman and Forbes, 2003], [Lee and Narayanan, 2005]
K-nearest neighborhood	[Dellaert et al., 1996], [Yu et al., 2001], [Yacoub et al., 2003], [Lee and Narayanan, 2005], [Yildirim et al., 2011]
Gaussian mixture models	[Breazeal and Aryananda, 2002], [Kockmann et al., 2009], [Dumouchel et al., 2009], [Kim et al., 2010]
Hidden Markov model	[Nogueiras et al., 2001], [Schuller, 2002], [Schuller et al., 2010], [Metallinou et al., 2010]
Artificial neural networks	[McGilloway et al., 2000], [Yu et al., 2001], [Yacoub et al., 2003], [Polzehl et al., 2009]

Table 4.1: *Classification techniques applied for speech emotion classification*

and pitch, etc. [Cairns and Hansen, 1994], [Banse and Scherer, 1996], [Li and Zhao, 1998], [Zhou et al., 1998], [Nwe et al., 2003], [Schuller et al., 2003], [Lee et al., 2004]. Only a few studies applied low-level feature modeling on a frame-level as an alternative: usually by hidden Markov models (HMM) or Gaussian mixture models (GMM) [Schuller et al., 2003], [Nwe et al., 2003], [Vlasenko and Wendemuth, 2007]. The higher success of static feature vectors derived by mapping of the low-level contours like energy or pitch by descriptive statistical functional application like lower order moments (mean, standard deviation) or extremal values specification [Ververidis and Kotropoulos, 2004] is probably proved by the supra-segmental nature of the phenomena appearing with respect to emotional content within a speech signal [Schuller et al., 2009b], [Schuller et al., 2009c]. In current speech emotion-classification research, voice quality features such as shimmer, jitter or harmonics-to-noise ratio (HNR) and spectral and cepstral features such as formants and MFCC have become the "new standard" feature sets [Barra et al., 2006], [Schuller et al., 2007a], [Lugger and Yang, 2007], [Schuller et al., 2009d]. Traditionally prosodic acoustic features, which can be classified in different ways, have been applied for affective speech processing. One of the possible emotional prosody features categorization was proposed by Anton Batliner in [Batliner et al., 2011].

The first categorization criterion lies in the feature set selection ap-

proach. The *'selective'* approach is based on phonetic and linguistic knowledge, [Kießling, 1996]; it is also well-known as *'knowledge-based'*. It has a strict systematic strategy for generating the features; a constant set of functions, which are applied to time series of different acoustic features. This approach normally results in more than 1 k features per set. Another approach is based on *brute-forcing* of features (1,000 up to 50,000) by analytical feature generation, partly also in combination with evolutionary generation [Schuller et al., 2008]. The difference between the two approaches lies in the feature selection step: in the *selective* approach, the selection takes place on an empirical level before putting the features into the classification process; in the *brute-force* approach an automatic feature selection is required.

The second categorization criterion is related to feature extraction staging. There is a *"two-layered"* approach, where firstly features are computed on the words level; secondly, functionals such as mean values and the average value are computed for all words within one utterance. An alternative is a *"single-layered"* approach, where features are computed for the complete utterance. In [Batliner et al., 2006], authors combined for the first time features extracted at different sites. By combining features from all sites, authors achieved up to 2.1 % absolute improvement for emotion-classification accuracy. These results will be discussed in more detail in section 4.4.4.

4.3 Emotion descriptors

One of the most important problems for the analysis of emotional speech is the selection on optimal unit of analysis. It is quite important to segment spontaneous speech signal into units that are discriminative for emotions [Vogt et al., 2008]. These are usually linguistically completed speech segments such as words, turns and/or utterances. However, the approval of the selected unit of analysis is an open research topic within the emotion-recognition research community. In most prototypical acted emotional speech datasets, subjects have to pronounce a complete utterance with some prompted emotional state. Most emotion-recognition experiments have been realized on datasets which contain acted emotions. As a result, the choice of an optimal unit of analysis is obviously just one utterance, a linguistically completed unit with no change of speaker's emotional state within this case. However, in spontaneous affective speech this kind of linguistically completed middle-length unit (utterance) is quite rare. Even the straight-forward extraction of linguistically completed segments like utterances do not guarantee a constant emotional state within the same utterance. An optimal unit of analysis of emotional speech has to fulfill certain requirements:

- *long enough to provide a sufficient amount of material for the calculation of acoustic features based on statistical functions*
- *short enough to provide stable acoustic properties with respect to emotions within the same unit*

For most acoustic features calculated from global statistics over an extracted speech signal, these units should have a minimum length. The emotion units analysis become more explicit as it is used more statistical acoustic features. On the other hand, all changes of the emotional state within one speech segment should be distinguishable, so the unit of analysis should be short enough that no alteration of emotion is likely to occur. Also, it should be so short that the acoustic properties of the unit of analysis with respect to speaker's emotional state are stable, so that informative acoustic features can be extracted. This is important for the extraction of acoustic features based on statistical measures, since, e.g., the mean value of a non-uniform unit of analysis induces an insufficient description. So the length of the optimal unit of analysis for emotional speech has to be chosen for these two conflicting requirements.

Just a few research evaluations have been performed to compare different types of units of analysis of emotional speech. Comparisons of utterances, words, words in context and fixed time intervals have been presented in [Vogt and Andre, 2005]. Authors have found that longer, linguistically completed segments tended to be better. Batliner et al. [Batliner et al., 2003] establish their acoustic features on words with a different number of context words. Further to simple word-level emotion recognition, they also mapped word-level results onto utterances and on chunks within the utterances. Within their evaluation authors found both advantages and disadvantages of shorter units than utterances, but they have not further quantitatively analyzed this aspect of emotional speech processing. In [Vogt et al., 2008] authors pointed out that the selection of the unit of analysis strongly depends on the type of emotional speech data. Most commonly dialog acts, utterances and turns as, e.g., in [Devillers et al., 2005], [Fernandez and Picard, 2005], [Oudeyer, 2003], [Schuller et al., 2005b] have been used as unit of analysis of emotional speech, but also words [Batliner et al., 2003], [Nicholas et al., 2006]. In the paper of Fragopanagos [Fragopanagos and Taylor, 2005] et. al. it is pointed out that most research efforts were made in order to investigate the affective speech processing on complete utterance, word-level or context-independent chunks. Only a few research groups provided a vowel- or syllable-level analysis during emotional speech processing. Goudbeek and others [Goudbeek et al., 2009] presented their investigation of the effect of emotion dimensions on formant placement in individual vowels. In affective speech synthesis,

Inanoglu [Inanoglu and Young, 2009] developed a set of fundamental frequency (F0) conversion methods on a syllable-level which utilized a small amount of expressive training data (approximately 15 minutes) and which had been evaluated for three target emotions: anger, surprise and sadness. Furthermore, an emotion-classification test showed that converted utterances with either F0 generation technique were able to convey the desired emotion above chance level. Research of Busso and others [Busso et al., 2007] showed that the mean and the variance of the likelihood score for emotional speech differ from the results observed in neutral speech, especially for emotions with a high level of arousal and observed in some broad phonetic classes (front vowels and mid/back vowels) which present stronger differences than others. Lee and others [Lee et al., 2004] showed quite a good speech-based emotion-recognition performance by using phoneme-class-dependent HMM classifiers with short-term spectral features. It has been shown by Vlasenko [Vlasenko and Wendemuth, 2009a] that a combination of a robust emotion-classification engine with a user-behavior-adaptive dialog model can make a spoken dialog system more friendly and user-centered.

4.4 Developed emotion-classification techniques

In this section we describe two pre-dominant paradigms of emotion classification: modeling on a frame-level by means of hidden Markov models and suprasegmental modeling by systematic feature brute-forcing. The second paradigm which can also be classified as static analysis has been introduced by our research partner Björn Schuller from Technische Universität München (TUM). In this section we will provide a detailed description of the classifiers which have been used for evaluations presented in our common publications [Vlasenko et al., 2007a], [Schuller et al., 2007], [Vlasenko et al., 2008b], [Vlasenko et al., 2008a], [Schuller et al., 2008], [Schuller et al., 2009], [Schuller et al., 2010].

4.4.1 Acoustic features

Within static analysis state-of-the-art emotion recognition we use a set of 1406 systematically generated acoustic features based on 37 low-level descriptors (LLD) as seen in Table 4.2 and their first-order delta coefficients. These 37×2 descriptors are then smoothed by low-pass filtering with a simple moving average filter. Statistics have been estimated on the turn-level by a projection of each uni-variate time series of the low-level descriptors onto a scalar feature

Low-level descriptors	Functionals
(Δ) Pitch	mean, centroid, standard deviation
(Δ) Energy	Skewness, Kurtosis
(Δ) Envelope	Zero-Crossing-Rate
(Δ) Formant 1–5 amplitude	quartile 1/2/3
(Δ) Formant 1–5 bandwidth	quartile 1 – min., quart. 2 – quart. 1
(Δ) Formant 1–5 position	quartile 3 – quart. 2, max. – quart. 3
(Δ) MFCC 1–16	max./min. value,
(Δ) HNR	max./min. relative position
(Δ) Shimmer	range max. – min.
(Δ) Jitter	position 95 % roll-off-point

Table 4.2: *Overview of low-level descriptors* (2×37) *and functionals* (19) *for static supra-segmental modeling*

independent of the length of the turn. This is done by using 19 different functionals. The list of the functionals can be found in Table 4.2.

Two optimization strategies can be also applied: First, speaker normalization (SN) by feature normalization taking into account speaker context. Second, feature-space optimization by removing highly correlated acoustic features (FS).

Within dynamic analysis, speech input is processed using a 25ms Hamming window, with a frame rate of 10ms. As in typical speech recognition, we employ a 39-dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. Specification of the MFCC features is discussed in detail in section 3.2.1.

To characterize vowels quality, first two resonant frequencies (formants) are used. The formants characterize the global shape of the immediate voice spectrum and are mostly defining the phonetic content and emotional prosody of the vowels [Benzeghiba et al., 2007]. For our evaluations, formant contours were extracted using PRAAT speech analysis software [Boersma and Weenink, 2008] and the Burg algorithm with the following parameters: the maximum number of formants tracked (five), the maximum frequency of the highest formant (set to 6,000 Hz), the time step between two consecutive analysis frames (0.01 seconds), the effective duration of the analysis window (0.025 seconds) and the amount of pre-emphasis (50 Hz).

4.4.1.1 Normalization and standardization

To help cope with channel characteristics, the cepstral mean substraction (CMS) can be applied. In our publication [Vlasenko et al., 2007a] we investigate the benefits of speaker normalization (SN), as we proposed to analyze emotion independent of the speaker, herein. SN is realized by a normalization

of each acoustic feature by its mean and standard deviation for each speaker individually. Thereby the whole speaker context is used. This has to be seen as an upper benchmark for ideal cases, where a speaker could be observed with a variety of emotional states. Yet, it is not essential to know the actual emotional state of observed utterances at the current moment.

4.4.1.2 Feature set optimization

It is common to use a high number of features for static modeling. A feature space optimization (FSO) is an important issue for increasing performance and real-time-capability. In order to optimize a set of acoustic features rather than combining the attributes of a single high relevance, we use a correlation-based analysis, herein [Vlasenko et al., 2007a]. Thereby acoustic features of high-class correlation and low inter-feature correlation are kept [Witten and Frank, 2005]. This does not employ the target static classifier in the loop. Likewise, it mostly reduces correlation within the acoustic feature space rather than an evaluation of influences on an improvement of single attributes. Still, this conducts to a very compact representation of the acoustic feature space which usually improves accuracy of the emotion classification while reducing feature extraction effort at the same time.

4.4.2 Static analysis

As pointed out earlier in section 4.2 mapping of the LLD contours by descriptive statistic functionals is justified by the supra-segmental nature of the emotional content occurring in spontaneous speech [Schuller et al., 2009b], [Schuller et al., 2009c]. For suprasegmental modeling of the speaker's emotional state we use a static analysis in combination with systematic feature brute-forcing. In order to represent a typical state-of-the-art emotion-recognition engine operating on a turn level, we use a set of $1,406$ acoustic features basing on 37 low-level descriptors (LLD) as seen in Table 4.2 and their first-order delta coefficients [Shahin, 2006]. These 37×2 LLDs are next smoothed by low-pass filtering with an SMA filter. The static analysis derives statistics per utterance by a projection of each uni-variate time series, respectively the low-level descriptors, X onto a scalar feature x independent of the length of the utterance. This is realized by use of a functional F, as depicted in equation 4.1.

$$F : X \rightarrow x \in \mathbf{R}^1 \tag{4.1}$$

19 functionals presented in Table 4.2 are applied to each contour on the turn-level covering *extremes, ranges, positions, first four moments and quar-*

tiles, etc. support vector machines (SVM) with linear kernel and pairwise multi-class discrimination have been used for classification purposes. One could consider the use of GMM here, as well. Yet, SVM provides better modeling of static acoustic feature vectors [Schuller et al., 2007b].

4.4.2.1 OpenEAR

In this section, we describe configuration parameters of a Munich open Affect Recognition Toolkit (openEAR) [Eyben et al., 2009] which have been used for our evaluations.

Feature Group	**Features in Group**
Raw Signal	Zero-crossing-rate
Signal energy	logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF).
	Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0 - 250 Hz, 0 - 650 Hz, 250 - 650 Hz, 1 - 4 kHz
	25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. of spectrum max. and min.
Mel-spectrum	Band 1-26
Cepstral	MFCC 0-12

Table 4.3: 33 low-level descriptors (LLD) used in acoustic analysis with openEAR

The OpenEAR is a toolkit for acoustic emotion recognition, which is based on static analysis. It is publicly available to anybody under the terms of the GNU General Public License (http://sourceforge.net/projects/openear).

For our evaluations we use the openEAR toolkit with 6,552 acoustic features extracted as 39 functionals of 56 acoustic low-level descriptors (LLD) and corresponding first- and second-order delta regression coefficients.

Table 4.4 lists the statistical functionals, which were applied to the LLD as shown in Table 4.3 to map a time series of variable length onto a static feature vector. The classifier of choice is support vector machines with polynomial kernel and pairwise multi-class discrimination based on sequential minimal optimization.

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

Table 4.4: 39 functionals applied to LLD contours and regression coefficients of LLD contours

4.4.3 Dynamic analysis

In our research we also applied a low-level feature modeling on a frame-level for emotion recognition from speech. The hidden Markov models (HMM) with Gaussian mixture models (GMM) have been used for this purpose. Three different units of analysis can by used for dynamic analysis: *utterance, chunk,* and *phoneme*. In this section we describe utterance-, chunk-, and phoneme-level dynamic analysis models for the recognition of emotions within speech.

4.4.3.1 Utterance-level classification

We consider using a statistical analysis applied for ASR to recognize emotion from speech in the first place [Vlasenko and Wendemuth, 2009b]. Likewise, instead of the usual task to deduce the most likely word sequence hypothesis Ω_k from a given vector sequence $\mathbf{O}$ of M acoustic observations $\mathbf{o}$, we will recognize the current speaker's emotional state. This is solved by a stochastic approach similar to the approach presented in equation 3.1, with a different argument interpretation:

$$\Omega_k = \arg\max_{\Omega} \log P(\Omega|\mathbf{O}) = \arg\max_{\Omega} \frac{P(\mathbf{O}|\Omega)P(\Omega)}{P(\mathbf{O})} \tag{4.2}$$

where $P(\mathbf{O}|\Omega)$ is called the emotion acoustic model, $P(\Omega)$ is the prior user-behavior information and Ω is one of all system known emotions.

In a case of turn-level analysis, the emotion acoustic model is designed by s state HMMs. Each state is associated with an output probability distribution $b_k(\mathbf{o}_t) = p(\mathbf{o}_t|s_t = k)$. The model distribution $b_j(\mathbf{o}_t)$ is based on the multivariate *Gaussian mixture model* (GMM), see equation 3.13. One emotion is assigned for a complete utterance. In other words within the training and testing observation feature vectors sequence $\mathbf{O}$ contains all feature vectors extracted within one utterance.

In simple cases the priors in the user-behavior model $P(\Omega)$ are chosen as an equal distribution among emotion classes. It is possible to provide context and an emotional-state-history-dependent complex user-behavior model. Within our evaluations presented in Chapter 5 we used a simple user-behavior model. During the recognition phase the emotion that results in the highest GMM score is chosen.

The HMM/GMM parameters are estimated by the EM-algorithm using speaker-independent training, namely *leave-one-speaker-out* strategy (LOSO) (see section 2.10.2), and a number of 1 to 120 Gaussian mixture components to approximate the original probability density functions (PDFs) [Young et al., 2009]. However, we also consider multiple states HMM/GMM $s = 1, 2, \ldots, 5$

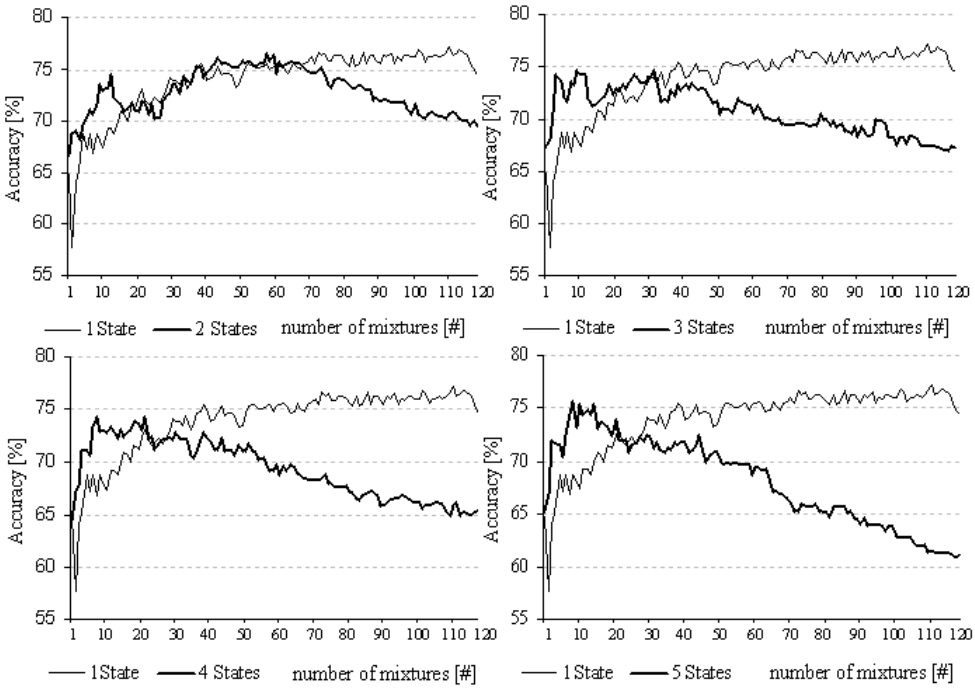


Figure 4.1: *Emotion-recognition accuracy (WA) depending on the number of Gaussian mixtures and number of HMM states, LOSO evaluation, database EMO-DB*

to better model dynamics. These are trained accordingly.

As can be seen in Figure 4.1 single-state HMM/GMM models show the most stable and robust results [Vlasenko et al., 2007b]. Within all emotion-classification evaluations presented in Chapter 5 based on utterance-level and chunk-level analysis we use single-state HMM/GMM models.

4.4.3.2 Chunk-level classification

This section describes another possible simple conceptual model of dynamic speaker's emotional state recognition. For classification purpose we can use HMM/GMM parameters estimated for utterance-level classification, see previous section. Instead of using turn-level classification, the time-synchronous one-pass Viterbi-beam search and the token passing algorithm with direct context-free grammar are used for decoding [Young et al., 2009]. This method is an integral component of continuous speech-recognition system based on HMM models, see section 3.2.8. To apply context-free grammar as constraints within the token passing scheme, these grammar rules are compiled into a set of linked syntax networks of the form illustrated in Figure 4.2. There are three types of the nodes of each syntax network: *links, terminals and non-terminals.* Link nodes are used to store tokens and are the points where recognition decisions are recorded. Terminal nodes correspond to emotion acoustic models and non-terminal nodes refer to separate sub-syntax networks representing the right-hand side (RHS) of the corresponding grammar rule. For our chunk-level emotion classification we did not use non-terminal nodes.

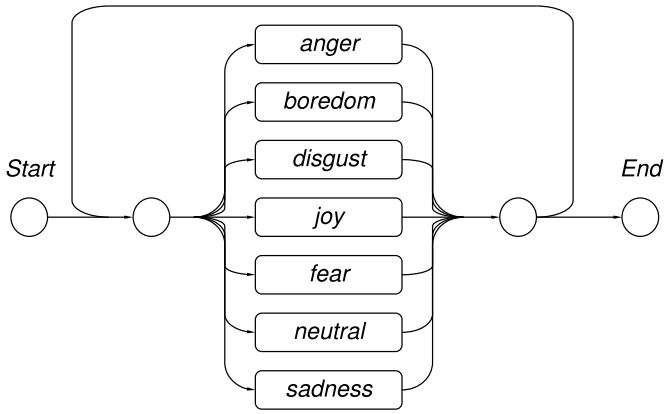


Figure 4.2: *Automatic chunking by acoustic properties and one-pass Viterbi beam search with token passing*

The three types of node are merged in such a way that every arc connects either a terminal or a non-terminal to a link node, or the other way around.

The syntax network presented in Figure 4.2 has exactly one entry, one exit and zero or more internal link nodes. Every terminal and non-terminal node could only have one arc leading into it, whereas each link node may have few arcs leading into it. Link nodes can thus be considered as filters, which remove all but lowest cost tokens passing through them [Young, S. J. et al., 1989]. More details about Viterbi-beam search with a token passing algorithm can be found in section 3.2.8.

The main idea is that tokens propagate through the networks just as in the finite state case: when a token node enters a terminal node, it is transferred to the entry node of the corresponding emotional state model.

This method can be used for detection of context-independent emotional chunks. Also this method can be modified for context-dependent emotional chunks detection. In this case the syntax network presented in Figure 4.2 should be combined with the user's emotional-state-driven language model. In section 4.4.4 we will present the two-stage emotion-classification technique which uses chunk-level classification as a first step of analysis.

4.4.3.3 Phoneme-level classification

Finally, the smallest possible units of analysis of emotional speech, namely *phonemes* have been chosen, as these should provide the most flexible basis for unit-specific models: if the emotion is feasible on a phoneme basis, then these sub-word units could be most easily re-used for any further content, and high numbers of training instances could be obtained [Vlasenko et al., 2008a], [Schuller et al., 2008]. Two different methods can be used for the phoneme-level emotion classification: *emotional phoneme classes* and *vowel-level formants tracking*.

Emotional phoneme classes: We use a simple conceptual model of dynamic emotional-state recognition on phoneme-level analysis: the full list of 36 phonemes (all phonemes which presented in EMO-DB dataset) is modeled for neutral and anger emotion speaking style, independently. As a first step of developing an emotion-classification module we decided that recognition of neutral and negative (anger) speaker's states is appropriate for an emotion adaptive dialog management. We integrated such speaker's emotion-recognition module in a prototype of the NIMITEK demonstrator [Wendemuth et al., 2008]. Within an interactive usability test we find that modeling only two speaker's emotional states, namely negative and neutral, is sufficient for development a user-friendly spoken dialog system. More details about an interactive usability test can be found in Chapter 6. Hence $2 \times 36 = 72$ phoneme emotion (PE) models are trained [Vlasenko et al., 2008a].

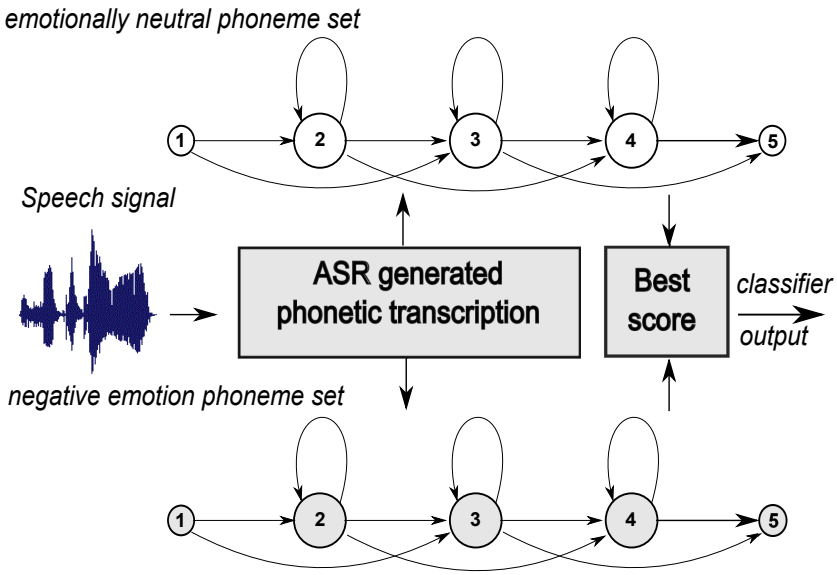


Figure 4.3: *Phoneme-level emotion recognition*

In the case of phoneme-level emotion analysis we can restate equation 4.2 in such a way:

Ω *is a possible emotional word (emotional phones sequence) from a defined vocabulary,*

$P(X|\Omega)$ *is an emotion acoustic model for word* Ω*,*

$P(\Omega)$ *is the affective speech language model.*

Emotional phonemes are modeled by training three emitting states HMM models with 16 Gaussian mixture components. There is not enough material in a selected part of EMO-DB database to train robust monophone models. Hence, in contrast to the previous models [Vlasenko et al., 2008a], [Vlasenko and Wendemuth, 2009a] we are using Kiel-trained monophones models as a background HMM/GMM model. The HTK toolkit was used for MLLR adaptation of the background model on two phoneme emotion subsets: neutral and anger. Neutral and anger samples from EMO-DB database were used for adaptation. In the case of phoneme-level emotion recognition we are using an ASR engine adapted for affective speech to recognize on word-level as a start point.

After this we are generating possible emotional phonetic transcriptions for sensible utterances by using an emotional phoneme set, see Figure 4.3. In our case, two transcriptions for neutral and anger speaking styles are generated. Emotional phoneme models which provide the highest recognition score are selected.

In the case of the Interspeech 2009 Emotion Challenge we used 72 phoneme emotion models for two emotional classes evaluation, and 180 phoneme emo-

tion models for five emotional classes. Results of the Interspeech 2009 Emotion Challenge will be presented in section 5.3.4.

Vowel-level formants tracking: It is also possible to classify emotions with an average formants value extracted from vowel segments [Vlasenko et al., 2011a], [Vlasenko et al., 2011b]. The phoneme boundaries estimation was based on a *forced alignment*, see section 3.3.2.1, provided by the HTK [Young et al., 2009]. Within our evaluation we use a simplified version of a BAS SAM-PA [SAM, 1996] with a set of 39 phonemes (18 vowels and 21 consonants). Table 3.2 and Table 3.1 present lists of German vowels and consonants, with their corresponding IPA and BAS SAM-PA symbols [SAM, 1996]. A list of vowels with their corresponding instances number can be found in Table 3.3. To receive the most reliable phoneme boundaries alignment mono-phone HMMs have been trained on each corpora independently.

Taking into account automatically extracted phoneme borders, we estimate an average first formant ($F1$) and second formant ($F2$) value for each vowel instance. Formant contours were extracted by using PRAAT speech analysis software [Boersma and Weenink, 2008] and the Burg algorithm. As one can see from Figure 3.7, the vowel triangles form and their position are different for different emotional states of the speaker. Now we want to find out if there are any discriminative changes to the average vowel's formant values as a function of the level of arousal of the speaker's emotional state. One can see that all emotional vowel triangles expand along the $F1$ axis more than along the $F2$ axis. As a consequence, we decided to use only the average $F1$ values for our evaluations.

Taking into account the *central limit theorem*, the mean of a sufficiently large number of vowel-level discrete estimations of first formant values, which definitely have a finite mean and a finite variance, will be approximately normally distributed. We define a new variable X which corresponds to an average $F1$ value estimated on vowel-level. The value of X can be calculated by:

$$X = \frac{1}{t_k} \sum_{i=1}^{t_k} {f_i}^1 \tag{4.3}$$

where t_k is a number of discrete estimations of first formant values within a vowel segment, ${f_i}^1$ is an estimation of the first formant value at discrete time i. The random variable X can be represented as $\mathcal{N}(x|\mu, \sigma^2)$ with the following probability density function:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\,\sigma^2}} \tag{4.4}$$

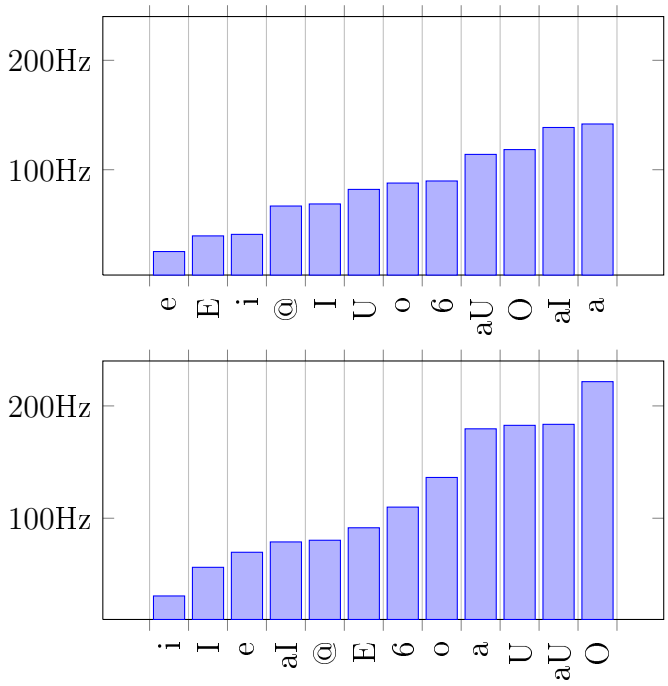


Figure 4.4: *Mean of the centralized $F1$ values for high-arousal emotions (fear, anger, joy). Speakers: male (top), female (bottom)*

To characterize the vowels quality changes under the influence of the different speaker's emotional state, we estimated the mean of the centralized $F1$ values for each vowel individually. For this evaluation, we use all vowels which contain a sufficient amount of instances for low and high-arousal emotions.

To specify the vowel quality variation, we use the mean of the centralized $F1$ value. The centralized $F1$ value shows the difference between the estimated average $F1$ value on an emotional vowel segment, and the mean of the average $F1$ value of the same vowel pronounced in a neutral way. Figures 4.4 and 4.5 display the mean of the centralized $F1$ values for the 12 vowels presented in the EMO-DB database. Due to the sparse amount of instances, we do not estimate the mean of the centralized $F1$ values for the following list of vowels [2,u,Y,9,OY] with corresponding IPA symbols [ʌ,u,ʏ,ɵ,ɔʏ].

As one can see from Figures 4.4 and 4.5, the most indicative vowels are [a, e, E, @, 6, aI, aU] with the corresponding IPA symbols [a, e ɛ, ə, ɒ, aɪ, aʊ]. Now we want to find out if it is possible to build a reliable simple emotion classifier based on the Neyman-Pearson criterion which will use the average $F1$ value as a parameter. This criterion is quite often used for speaker classification, identification and authorization tasks [Roberts et al., 2005]. The average $F1$ value will be extracted within the alignment boundaries of the most indicative vowels.

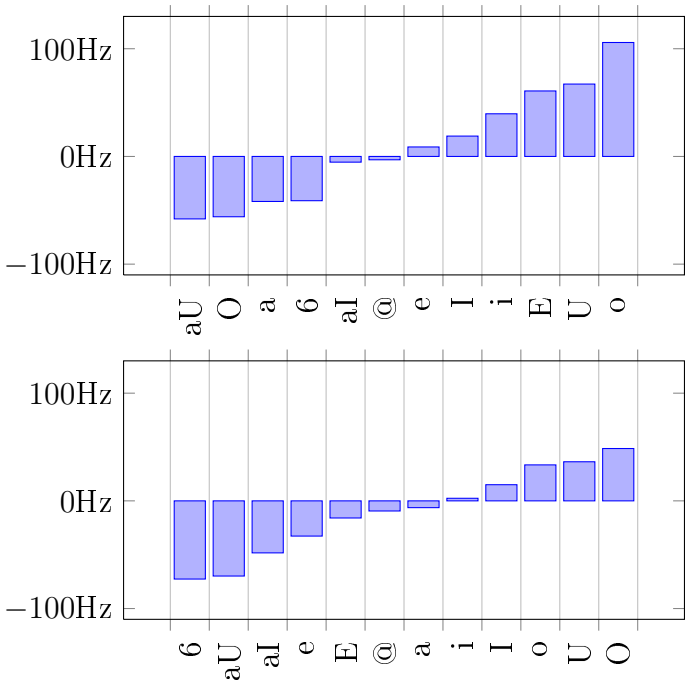


Figure 4.5: *Mean of the centralized $F1$ values for low-arousal emotions (boredom, sadness) in comparison with neutral speech. Speakers: male (top), female (bottom)*

We pointed out earlier that the random variable X defined in equation 4.3 is approximately normally distributed. As a result, it can be represented by $\mathcal{N}(x|\mu, \sigma^2)$. Now we shall compute the normal distribution parameters for each indicative vowel pronounced in a neutral speaking style. Due to the high variability of speaker vocal tract lengths for male and female voices we decided to calculate the pair of estimations (μ, σ) for each gender individually. For calculating the mean value estimations μ, we use two neutral speech sentences per speaker for the EMO-DB dataset and one utterance per speaker with the smallest absolute arousal value for the VAM dataset. These sentences have been removed from the test sets. For gender-dependent σ estimations of seven of the most indicative vowels we use the Kiel corpus. It is clear that there is not enough material within two sentences to calculate a reliable standard deviation estimation. To solve this problem, instead of using speaker-dependent σ estimations we use gender-dependent (male,female) estimations calculated on the Kiel corpus [Vlasenko et al., 2011a]. The list of normal distribution parameters for indicative vowels can be found in Table 4.5.

For our evaluation we generate male and female (μ, σ) estimations pools. Mean and standard deviation values from these pools will be adopted for each utterance according to speaker's gender. This can be expressed as follows:

Vowel	EMO-DB		VAM		Kiel	
ID	**male** μ [Hz]	**female** μ [Hz]	**male** μ [Hz]	**female** μ [Hz]	**male** σ [Hz]	**female** σ [Hz]
a	644.1	749.4	658.8	769.8	62.0	119.4
e	443.7	439.1	488.9	607.5	67.6	88.1
E	440.8	439.2	579.2	623.0	66.1	111.1
@	509.1	475.0	555.2	584.5	123.6	124.6
6	547.8	584.1	594.6	690.7	89.5	127.1
aI	610.6	731.7	615.5	756.0	48.7	78.1
aU	514.7	594.1	684.9	694.4	48.1	77.6

Table 4.5: *Estimations of the normal distribution parameters calculated on Kiel, EMO-DB and VAM corpus material*

$\mu_{ik} = \mu_{ig(k)}$, $\sigma_{ik} = \sigma_{ig(k)}$, where i is an index of an indicative vowel, k is an utterance index, $g(k)$ is a function which specifies a speaker's gender of utterance k.

For classification purposes we use the Neyman-Pearson criterion:

$$\Lambda(U) = \frac{L(\Theta_0|U)}{L(\Theta_1|U)} \leq \eta \tag{4.5}$$

In our case, Θ_0 is a hypothesis that all indicative vowels included in utterance U are being pronounced with high-arousal emotion, and Θ_1 is a hypothesis that all vowels included in utterance U are being pronounced with neutral or low-arousal emotion.

Now we estimate $L(\Theta_0|U)$ and $L(\Theta_1|U)$. The cumulative distribution function (CDF) for the random variable X_i which corresponds to the average $F1$ value of an indicative vowel i is defined by:

$$P(X_i \leq x) = F_{X_i}(x) = \int_{-\infty}^{x} f(x_i)dx_i \tag{4.6}$$

Taking into account that the random variable X_i has a normal distribution, equation 4.6 can be expressed as:

$$F_{X_i}(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\,\sigma_i{}^2}} dx = \frac{1}{2}(1 + \operatorname{erf}\left\{\frac{x-\mu_i}{\sigma_i\sqrt{2}}\right\}) \tag{4.7}$$

where erf is a Gauss error function:

$$\operatorname{erf}\{x\} = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{4.8}$$

Taking into account equation 4.6, our conditional likelihoods $L(\Theta_0|U)$ and $L(\Theta_1|U)$ can be expressed as:

$$L(\Theta_0|U_k) = \sum_{\forall i\,:x_{ij(k)}\in U_k} F_X(x_{ig(k)})$$

$$= \sum_{\forall i\,:x_{ij(k)}\in U_k} \frac{1}{2}\left(1 + \mathrm{erf}\left\{\frac{x_{ij(k)} - \mu_{ig(k)}}{\sigma_{ig(k)}\sqrt{2}}\right\}\right) \tag{4.9a}$$

$$L(\Theta_1|U_k) = N_k - L(\Theta_0|U_k) \tag{4.9b}$$

where i is an index of an indicative vowel, $g(k)$ is a function which specifies a speaker's gender of the utterance k, k is an index of utterance, and N_k is the number of indicative vowels in the utterance U_k.

As a consequence, the Neyman-Pearson criterion can be estimated as:

$$\Lambda(U_k) = \frac{N_k}{P(\Theta_1|U_k)} - 1 \leq \eta \tag{4.10}$$

Equation 4.10 can be used for estimation of $\Lambda(U_k)$ during training and test stages. During training we should estimate the optimal η value. Also, the criterion threshold η can be estimated with leave-one-speaker-out (LOSO) strategy or by using some a-priory value $\eta = 1$ (it is a case when we simply select the hypothesis with higher likelihood). Within the test stage all utterances U_k with $\Lambda(U_k) \leq \eta$ will be classified as utterances pronounced in low-arousal emotional or neutral state. In other cases they will be classified as utterances articulated by the speaker with a high-arousal emotional state.

4.4.4 Combined analysis

Most parts of emotion-classification techniques usually employ static feature vectors extracted on a turn or linguistically completed sub-turn entities [Batliner et al., 2011]. Dynamic processing on the short-term frame-level is a less popular technique applied for the emotion recognition from speech [Polzin and Waibel, 1998], [Schuller et al., 2003]. In [Schuller et al., 2003], [Schuller et al., 2009] the latter has also been shown superior to dynamic modeling. This derives mostly from the fact, that by statistical functional application to the low-level descriptors (LLD) an important information reduction takes place, which avoids phonetic (respectively spoken-content) over-modeling. Yet, it is also considered that thereby important temporal information is lost due to a high degree of abstraction [Vlasenko et al., 2007a]. This led to the first successful attempts to integrate information on different processing levels [Murray and Arnott, 1993], [Li and Zhao, 1998], [Jiang and Cai, 2004], [Schuller and Rigoll, 2006].

In this section we describe two possible combined speech-based emotion-classification techniques: *two-stage processing and middle-level fusion.*

Two-stage processing: As the standard unit of emotional speech analysis a whole turn can be named [Polzin and Waibel, 1998], [Li and Zhao, 1998], [Schuller et al., 2003], [Jiang and Cai, 2004], [Batliner et al., 2006]. From an application point of view, this seems appropriate in most cases: a change of speaker's emotional state during a turn seems to occur seldom enough for many applications. However, from a classification point of view, it was often reported that sub-timing levels seem to be advantageous [Jiang and Cai, 2004], [Murray and Arnott, 1993], [Schuller and Rigoll, 2006]. Still, apart from a few attempts to recognize speaker's emotions within speech dynamically [Polzin and Waibel, 1998], [Schuller et al., 2003], current approaches usually employ static feature vectors derived on a utterance-, turn-, word-, or chunk-level [Schuller et al., 2007b]. In [Schuller et al., 2003] such static modeling has also been shown superior to dynamic modeling. In this section we therefore investigate a two-stage approach to acoustic modeling for the recognition of emotion from speech: a first stage segments utterances into chunks which are analyzed in detail in a second stage.

The two-stage approach is implemented to provide a higher temporal resolution by chunking of utterances according to their acoustic properties, and multi-instance learning for the turn mapping after an individual chunk analysis. For the chunking fast pre-segmentation into emotionally quasi-stationary segments the HMMs-/GMM-based one-pass Viterbi beam search with token passing is used. The chunk analysis is realized by brute-force large feature space construction with subsequent subset selection, support vector machines classification, and speaker normalization.

For the first stage we use the chunk-level analysis described in section 4.4.3.2. We train the chunk-level emotion-recognition models in a speaker-independent manner with LOSO strategy (see section 2.10.2) by using the Baum-Welch re-estimation algorithm presented in Chapter 3 and 50 Gaussian mixture components. Afterwards each original utterance is chunked by application of the one-pass Viterbi beam search as described. For the latter processing, only the obtained chunk boundaries are used from this stage. The motivation behind this processing is to find an acoustically motivated sub-turn splitting.

For the second stage we use the turn-level analysis described in section 4.4.2. In order to map the static analysis results of each chunk onto the turn-level, we consider three strategies known from multi-instance learning for each chunk:

- *an un-weighted majority vote (MV),*
- *a maximum length vote (MLV),*
- *a maximum classifier prediction score multiplied with the length vote (MSL)*

Likewise, we computed the majority label of each turn based on the chunk-level. In the case of a weighted vote, the length of the chunk in frames is used as a multiplicative weighting function. In the MSL case we also use the classifier prediction score for each class as additional weight. Note that in the case of an unweighted majority vote, turns may occur that cannot be uniquely assigned to an emotional class. This happens, if two or more emotional classes, which are the majority of classes, have the same number of chunks. This case will be separately denoted in the ongoing. In the case of time-based weighting this case can almost be ignored, as the majority of classes – if there are several – will rarely have an equal number of frames [Schuller et al., 2007]. This is even more likely, if length and prediction scores are used for weighting (MSL). As a disadvantage it has to be mentioned that temporal information is thereby lost. Alternatively, the duration of each chunk can be used as weight. Also, the time order of appearance of chunks is lost. However, we suppose that this information can be neglected under the precondition of constant emotion throughout an utterance. Employing majority voting (MV) we can observe two cases: utterances that are clearly assignable, and such that have two or more emotions assigned due to a draw. In the second case, a further discrimination can be considered: utterances that have the correct emotion among the majority classes, and such that are simply incorrectly assigned. Evaluation results of a two-stage speech-based emotion-classification technique will be presented and discussed in Chapter 5.

Middle-level fusion: To receive higher classification performance it is possible to use independent classification results for *middle-level fusion*. In most cases, with this method we can obtain a composite classification performance which is higher than that of the individual classifiers. As presented in [Batliner et al., 2006], with ROVER framework [Fiscus, 1997], authors showed an absolute improvement of up to 5.8 % of emotion-recognition accuracy on four class problems on AIBO [Batliner et al., 2008] dataset with respect to the best independent site result. Within early fusion, when combining acoustic features from all sites, authors achieved still a 2.1 % absolute improvement.

So far the two individual approaches to emotion recognition based on information processing directly on the frame level, or on a higher turn level, have been presented. In order to fuse these two approaches it seems beneficial to keep utmost amounts of information for the final decision process. However,

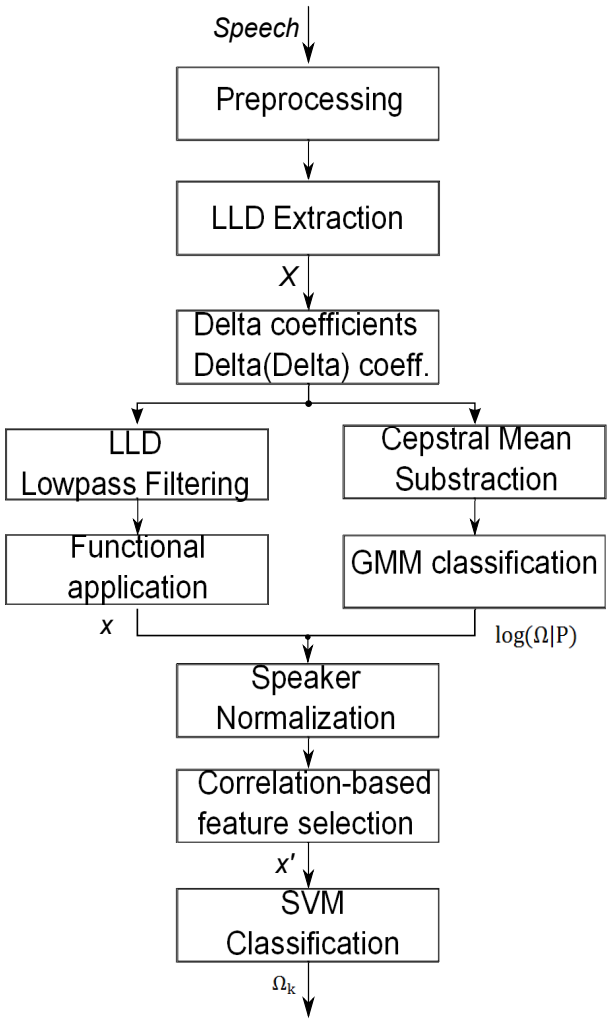


Figure 4.6: *Processing flow for the middle-level fusion of frame- and turn-level analysis*

an *early fusion* (acoustic features fusion) is not feasible, due to the different acoustic feature sets (frame-level vs. turn-level) [Vlasenko et al., 2007b]. We therefore decided to include the final HMM/GMM scores ($\log P(\Omega|X)$) within the static acoustic feature vector x, forming an argument vector x', and provide a *middle-level fusion*. The process of speaker normalization and feature space optimization is extended to the likewise obtained new feature vector x'. Overall feature selection having the HMM/GMM scores within the space reveals their high importance, as they are kept among high ranks. Figure 4.6 depicts the overall processing flow from an input speech signal via the two streams to the final classification result [Vlasenko et al., 2007a].

4.5 Context-dependent and context-independent models

Usually emotion recognition from speech uses spoken content independent acoustic models. One general model per speaker's emotional state is trained independent of the phonetic structure of affective speech samples. Given sufficient training samples, this approach provides acceptable emotion-recognition performance on test material which has similar phonetic content [Schuller et al., 2009]. This section tries to answer the question of whether emotion recognition from speech strongly depends on the content, and if models tailored for the spoken unit can lead to higher accuracies [Vlasenko et al., 2008a]. We therefore evaluate phoneme-, word-, utterance-models by use of a large prosodic, spectral, and voice quality feature space, HMM/GMM models and SVM.

Practically every approach to the emotion recognition from speech ignores the spoken content when it comes to acoustic modeling (see [Polzin and Waibel, 1998], [Li and Zhao, 1998], [Schuller et al., 2003], [Jiang and Cai, 2004], [Batliner et al., 2006]). A general model is trained for each speaker's emotional state, and applied on test-utterances which have a similar phonetic content. While this is a common practice, it seems surprising how well this works, especially considering that many acoustic features highly depend on phonetic structure, such as spectral and cepstral features which have become very popular recently [Batliner et al., 2006]. It is common to provide a high reduction of information: e.g., rather than using the original time-series, higher order statistics, such as means, deviations, extremes, etc., are used. Another possible solution is to use dynamic modeling, e.g., by the HMM, of low-level descriptors (MFCC, etc.) extracted on the frame-level [Schuller et al., 2003], [Vlasenko et al., 2007a].

We first investigate the influence of spoken content variation on the turn-level. We use dynamic analysis (see section 4.4.3) with utterance-level classification. Test runs on EMODB and SUSAS datasets for utterance models are carried out speaker independently by leave-one-speaker-out (LOSO) evaluation. Table 4.6 reports average among all speakers and all utterances accuracies for three cases to address context-independent evaluation. A total of 10 different utterances are found in EMODB and 35 in SUSAS databases, respectively. We included all utterances from training set for general model training. In other cases we left out all samples with target or non-target utterance from training set.

From 4.6 it is clear that removal of target utterance from training set fundamentally reduce accuracy of emotion recognition in comparison with re-

WA	EMO-DB	SUSAS
General model	77.1	46.0
Non-target utterance left out	75.9	45.4
Target utterance left out	72.7	44.2

Table 4.6: *Weighted average recalls (WA) [%] for turn-level modeling on EMODB and SUSAS. Dynamic analysis with utterance-level classification, LOSO evaluation*

moval non-target utterance. Random removal non-target utterances preserves the context, which results in the higher accuracy than removing the target utterance, which makes the training data context-independent.

Yet, the question is if phonetic content variance influences emotion-recognition performance negatively, and if models trained specifically on the phonetic unit at hand, can help. In this section, we aim to shed light on this question by training phoneme-, syllable-, and word-models for the emotion recognition in the following application. Unit-definite models require knowledge of the phonetic content, opposing "blind" sub-turn entities, as introduced in (see [Murray and Arnott, 1993], [Polzin and Waibel, 1998], [Li and Zhao, 1998], [Jiang and Cai, 2004], [Schuller and Rigoll, 2006]).

Likewise, recognition of the spoken content becomes essential, in order to choose the correct unit-definite model. Facing real-world cases, we do not report on transcribed content, as, e.g., in [Batliner et al., 2006], but do include the HMM-based state-of-the-art approach to ASR. The HMM of three emitting states and 16 Gaussian mixture components was built for each phoneme emotion (PE) and phoneme-level of interest (PLOI) models. The HTK toolkit was used to build these models, using standard techniques such as forward-backward and Baum-Welch re-estimation algorithms [Young et al., 2009]. We also use an automatic speech-recognition (ASR) engine adapted with MLLR and regression class tree on affective speech samples to recognize linguistic units (sentence, word) [Vlasenko et al., 2008a]. We report results considering superiority of unit-definite models over general models, and combine speech and emotion recognition in a real system.

Next, word-definite emotion models have to be selected for emotion recognition. This may lead to a downgrade, if word insertions, deletions or substitutions occur, provided the spoken content *does* influence emotion recognition [Vlasenko et al., 2008a]. Therefore, we test emotion recognition in matched and mismatched word condition (that is picking the correct or any other word model at a time) in comparison to a general model trained on all words. Note that for mismatched condition one vs. one training and testing of each word vs. each other is necessary.

Model description	Conditions	G1	G2	All
EMO-DB	matched	57.2	46.9	48.9
	mismatched	36.6	37.7	37.4
SUSAS	matched	64.6	60.3	60.7
	mismatched	52.4	54.4	55.2
AVIC	matched	79.7	57.8	60.9
	mismatched	49.2	51.3	50.1

Table 4.7: *Weighted average recalls (WA) [%] at word level for word emotion models in matched and mismatched condition. Static features, SVM, LOSO. Investigated are "worth-it" words (G1) and "non-worth-it" candidates (G2), as well as all (All) terms*

In total 73 different words are pronounced in EMO-DB database [Burkhardt et al., 2005]. From these we select only those that have a minimum frequency of occurrence of 3 within each emotion (likewise having 50 plus instances per word) comprising a total of 41 words with roughly 200 instances per word. Then, we employ static acoustic features and SVM classification for word emotion models after selection of according words by ASR. Table 4.7 visualizes the results received by two groups of frequency of occurrence in the corpus:

Group 1 (G1) are high frequency of occurrence words. For the EMO-DB dataset these words (10 out of 41) are *"abgeben (give away), am (on), auf (on top of), besucht (visits), gehen (walk), ich (I), sein (to be), sich (oneself), sie (her),* and *sieben (seven)"*. For the AVIC dataset these words (7 out of 50) are *"ah, but, is, it, mh, not,* and *you"*. For the SUSAS dataset this word (1 out of 11) is *"fifty"*.

In contrast, *group 2 (G2)* is "not worth it" due to low occurrence in the dataset. Likewise emotion unit-definite models for these words cannot be trained sufficiently. Besides, results of emotion recognition for all words are shown (All). Our evaluations are realized in a speaker-independent (SI) manner using LOSO strategy (see section 2.10.2). In the following, we stick to words as unit of analysis, which allow for incremental emotion recognition.

First, matched vs. mismatched conditions are examined: spoken content clearly does influence accuracy throughout word-model comparison in any case, as can be seen in Table 4.7. In fact, detailed analysis of complete results shows that the length of words and phonetic distance are the main influencing factors.

Considering results of word-level analysis for acted and spontaneous emotion and spontaneous level of interest, we discovered notable differences between matched and mismatched condition for words presented in group G1

Training size factor	1%	2%	5%	10%	100%
EMO-DB	43.1	44.7	49.1	51.7	55.5
SUSAS	50.6	56.1	60.7	61.5	64.7
AVIC	58.0	62.6	65.2	68.6	68.6

Table 4.8: *Weighted average recalls (WA) [%] at word level for word emotion models for general models at diverse relative sizes of training corpora. Static features, SVM, LOSO*

and G2. As can be seen from Table 4.7, in matched cases word-definite models for words from the group G1 provide better performance in comparison with general emotion models.

As mis-selection of word-definite emotion models would evidently significantly downgrade performance, we next address the question of how a general model trained on the whole dataset would perform.

We set this in relation to the amount of training data available for each word-definite emotion model by the relative training size factor by random down-sampling preserving emotion class-balance, see Table 4.8

Acoustic material for the each word correspond from 1.0% to 2.0% of complete acoustic material presented in EMO-DB, SUSAS, AVIC datasets. It can be seen that for all databases a general model with that training size factor will perform between matched and mismatched conditions for all words. With more training material available, the general model outperforms the matched case picking "All" and approaches the "G1" matched case. Without "G1" selection it seems preferable to use the general emotion model, simply as more training data is available. With "G1" matched cases accuracy of emotion recognition with word-definite emotion models outperform general models with a 100% training size factor.

However, the introduced unit-specific emotion-recognition models clearly outperformed common general models provided sufficient amount of training material per unit. Appearance of word-level-labeled corporas can improve current performance of phoneme- and word-level emotion and level of interest models. We found that emotional and level of interest activity is distributed irregularly among words within a sentence. For example in AVIC dataset, accuracy of level of interest recognition for the words *"ah, but, is, it, mh, not, you"* by word depended models exceeds accuracy of level of interest detection by general models. More details about this dataset can be found in section 2.6.2.7 on page 28. This is not the case for other evaluated datasets.

4.6 Summary

This chapter reviews existing speech-based emotion-recognition methods and provides a description of our developed emotion-recognition approaches. A variety of emotion descriptors is discussed first. Two different types of emotional speech analyses are applied for speech-based emotion recognition: *frame-level* and *turn-level*, are then presented. First of all we described the set of acoustic features which can be applied for different emotion-classification techniques. Two different optimization techniques applied on feature extraction level, namely *normalization and standardization* and *feature set optimization* have been presented afterwards. Static analysis applied for speech-based emotion-classification developed by our partners from TUM has been presented first. Then we introduced *utterance-*, *chunk-*, *phoneme-level* dynamic analysis models for the recognition of emotions within speech. During description of utterance-level dynamic analysis we determined the optimal HMM/GMM architecture. As a result within our evaluations of utterance-, chunk-level dynamic analysis we will use the single-state HMM/GMM architecture for emotion-classification models. Two different phoneme-level emotion-classification methods are described. The first is *emotional phoneme classes*. It provides context-dependent emotion classification and can be easily combined with automatic speech recognition for a user-behavior-adaptive spoken dialog system. Results of emotional phoneme classes evaluations can be found in Chapter 5. Also a prototype spoken dialog system with a user-behavior-adaptive spoken dialog system created within NIMITEK, which includes this technique will be discussed in Chapter 6. The second is *vowel-level formants tracking*. This method is our new technique, which showed applicable results of emotion recognition based on an extremely small acoustic feature set.

Within this chapter, we described two possible information integration techniques which use different processing levels. The first is a *two-stage processing* approach which is used to provide higher temporal resolution by chunking of utterances according to acoustic properties, and multi-instance learning for turn mapping after individual chunk analysis. The second is *middle-level fusion*. Within this method we integrate important information on temporal sub-layers as the frame-level within turn-level feature space. Finally, this chapter addresses the question on which phonetical level there is the onset of emotions and level of interest. We therefore compare phoneme-, word- and sentence-level analysis for emotional sentence classification by use of a large prosodic, spectral, and voice quality feature space for SVM and MFCC for HMM/GMM. Results of evaluations of our static and dynamic emotion classifiers will be presented in Chapter 5.

Chapter 5

Recognition experiments

Contents

5.1 Introduction

In this chapter we present results of experiments concerning our emotion-recognition and automatic speech-recognition methods. All experiments were conducted on The Kiel Corpus of Read Speech [KIE, 2002] and on the affective speech datasets presented in Table 2.3 on page 24. Building of the acoustic models and speech-recognition evaluation setup for neutral and affective speech samples, and adaptation on affective speech samples for acoustic models trained on neutral speech samples are presented in section 5.2. Section 5.3 discusses evaluation results of various emotion-classification methods presented earlier in Chapter 4. Then, we present our results within INTERSPEECH 2009 Emotion Challenge [Schuller et al., 2009c] and cross-corpus acoustic emotion recognition.

5.2 Evaluation of our ASR methods

This section presents the development of experiments on the German speech recognition with HMM/GMM models. All HMM/GMM models presented in this section are constructed as 18 Gaussian mixture components per state. ASR models presented in this section are evaluated with the bigram language model and a grammar scale factor $s = 5$. A larger number of the Gaussian mixture components and a higher grammar scale factor could improve performance of a defined thematic domain (system known fixed textual content of the evaluated database) – oriented automatic speech recognition.

The main issue of this section is to show that training ASR models on neutral speech, and subsequent adaptation on affective speech samples, does have an impact on the recognition performance within emotional speech recognition. Two different HMM/GMM models sets are presented and evaluated. First, we describe our non-adapted HMM/GMM models, trained independently on neutral speech samples and affective speech samples. Afterwards, we describe our affective-speech-adapted ASR models and present evaluation results on the EMO-DB database.

5.2.1 Corpora

As an emotionally neutral speech corpus we used part of The Kiel Corpus of Read Speech (PHONDAT90 and PHONDAT92: Kiel-CD #1, 1994) [KIE, 2002]. The Kiel Corpus is a growing collection of read and spontaneous German speech which has been collected and labeled segmentally since 1990. For our ASR engine evaluation we used speech samples from 12 female (1801 utterance in all) and 13 male (2000 utterance in all) speakers. The list of speakers is k01,...,k12, k61 (also defined as kko), k62 (also defined as rtd), k63,...,k70, dlm, hpt, uga. Within speech recording sessions a Neumann U87 condenser microphone (cardioid settings) was placed approximately 30 cm from the speaker's mouth. The microphone signals were amplified by a John Hardy M1 pre-amplifier and recorded on a SONY PCM 2500 DAT-recorder at a sampling rate of 44.1 kHz for PHONDAT90 and of 48 kHz for PHONDAT92, respectively, with 16 bit quantization. Afterwards, collected speech samples were then digitally transferred to a computer hard disk and downsampled to 16 kHz as well as high-pass filtered at 40 Hz.

For affective speech corpora we decided to use the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [Burkhardt et al., 2005]. Speech material recordings took place in the anechoic chamber of the Technical University Berlin, Technical Acoustics Department using a Sennheiser MKH 40 P 48 microphone and a Tascam DA-P1 portable DAT recorder. Recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz. The microphone was placed approximately 30 cm from the speaker's mouth. 10 professional actors (5 male and 5 female) spoke 10 German emotionally undefined sentences. One of these sentences is "**b03:** *An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. (At the weekends I have always gone home now and seen Agnes.)*". To provide reliable measures twenty evaluators took part in a perception-test. Each "rater" heard all of the utterances in a random order. They were allowed to listen to each utterance only once before the perception-test evaluator had to decide in which emotional state the speaker had been and how natural the

performance was. During perception test raters provided rates of naturalness and recognizability for each performance. An average rates of naturalness and recognizability have been included in dataset material. In total we used 494 utterances: 416 affective speech samples and 78 neutral speech samples. Each of these utterances has a level of naturalness not less than 60% and a level of recognizability not less than 80%.

5.2.2 Evaluation of non-adapted ASR models

In our ASR models, only diagonal covariance GMM matrix systems are considered where the features in each feature vector are assumed uncorrelated. The monophone set consists of 39 HMMs including *silence* and *short pause* (sp). Within our ASR evaluations we use a standard 39-dimensional feature vector which includes 12 MFCC coefficients, zero-order Cepstral coefficients, and their delta and acceleration coefficients.

The parameters of the models are re-estimated in 2 consecutive runs of the Baum-Welch algorithm (see section 3.2.6) using the monophone transcription of the training data. To handle impulsive noises in the training speech samples, additional transitions are added from state second to forth and from state forth to second in the silence HMM model. The backward transition provides a technique to assimilate impulsive noises without exiting the silence model. Besides, in order to deal with continuous speech, a one state short pause (sp) model was created whose emitting state is tied to the third state (central emitting state) of the silence model. This short pause model has a direct transition from entry to exit state. Then two more iterations of the Baum-Welch algorithm are run.

Finally, we convert the single-Gaussian component models to 18 mixtures Gaussian component models. After each mixture component increment, the resulting HMM models are re-estimated with 4 consecutive iterations of the Baum-Welch algorithm. During training of our HMM parameters we added one mixture per 4 consecutive runs of the Baum-Welch algorithm. For language modeling we used a bigram language model trained on transcriptions of the complete training set.

Test-runs on EMO-DB, Kiel for non-adapted ASR models are carried out in leave-one-speaker-out (LOSO) manner to address speaker independence (SI), as required by most applications. For each speaker presented in EMO-DB or Kiel we trained a speaker-independent ASR system based on speech samples from other speakers presented in the corresponding database. As a result we trained 10 ASR HMM/GMM models for the EMO-DB database and 25 ASR HMM/GMM models for the Kiel database. Within our evaluations on non-adopted models we also used cepstral mean subtraction (CMS), which is the

Speaker ID [%]	With CMS		Without CMS	
	Acc	*Corr*	*Acc*	*Corr*
k01	96.15	96.15	96.72	96.92
k02	92.37	93.32	92.75	93.32
...	...	...	...	...
k08	99.00	99.40	98.60	99.20
...	...	...	...	...
k61	87.93	90.13	87.78	89.82
k62	87.72	89.86	87.56	89.78
...	...	...	...	...
dlm	79.29	81.02	78.73	80.69
hpt	85.38	86.60	85.00	86.41
uga	91.85	93.02	91.61	92.83
Total	**90.20**	**91.58**	**90.04**	**91.47**

Table 5.1: *Recognition rates [%] for non-adapted ASR HMM/GMM models trained and evaluated on the Kiel database with LOSO*

simple method applied for the compensation of the long-term spectral effects such as those induced by different microphones and audio channels [Young et al., 2009].

Recognition rates of the HMM/GMM models trained on the Kiel dataset and evaluated with the bigram language model can be found in Table 5.1. In general, it can be seen that the performance of German affective speech recognition for speaker-independent models are substantially different. For example, we obtained the speech-recognition accuracy rate for speaker k08 up to $Acc = 98.6\%$ (acoustic features without CMS) at the same time the accuracy rate for speaker dlm was only $Acc = 78.73\%$ (acoustic features without CMS). Such a high performance variation can be explained by low-level textual content annotation in the Kiel dataset. Some speakers do not pronounce the corresponding prompted text within recordings, also paralinguistic events (like breathing and etc.) have not been transcribed. However, we will use Kiel datasets for training our basic ASR models for German emotionally neutral speech.

Recognition rates of the HMM/GMM models trained on speech samples from the EMO-DB dataset and evaluated with bigram language model can be found in Table 5.2. In general, it can be seen that the performance of German affective speech recognition for speaker-independent models are comparable. Only for speaker 10 we obtained a comparable low affective-speech-recognition accuracy rate $Acc = 83.55\%$ (acoustic features without CMS). Such comparable low performance can be explained by very specific vocal tract characteristics of speaker 10 and a high-level of intensity of the simulated emotions.

Recognition rates of the ASR models trained on the Kiel database and eval-

Speaker ID [%]	With CMS		Without CMS	
	Acc	*Corr*	*Acc*	*Corr*
03	98.09	98.09	98.33	98.33
08	98.14	98.52	98.70	99.07
09	97.12	97.12	96.07	96.60
10	84.84	86.77	83.55	86.45
11	97.43	97.82	97.03	97.62
12	97.23	97.63	96.84	97.23
13	98.11	98.11	98.11	98.11
14	98.62	99.23	98.31	98.92
15	97.44	97.44	97.44	97.44
16	95.49	95.80	95.33	95.80
Total	**96.70**	**97.06**	**96.49**	**96.99**

Table 5.2: *Recognition rates [%] for non-adapted ASR HMM/GMM models trained and evaluated on the EMO-DB database with LOSO*

Database [%]	With CMS		Without CMS	
	Acc	*Corr*	*Acc*	*Corr*
Kiel	85.99	86.97	87.37	88.27

Table 5.3: *Recognition rates [%] for non-adapted ASR HMM/GMM models trained on the Kiel database, evaluated on the EMO-DB database*

uated with bigram language model on the EMO-DB database can be found in Table 5.3. As one can see from Table 5.3, HMM/GMM models trained on the complete Kiel dataset without cepstral mean subtraction (CMS) provides the best German speech-recognition rates within cross-corpora ASR evaluation. As a result we decided to use HMM/GMM models trained on acoustic features extracted from the Kiel dataset without CMS. In the next section we will describe affective-speech-adaptation techniques which have been applied for these ASR models, referred to as *basic* ASR models for German emotionally neutral speech.

5.2.3 Evaluation of affective-speech-adapted ASR models

As one can see from Table 5.3, HMM/GMM models trained on neutral speech samples from the Kiel database could not provide sufficient recognition performance on affective speech material from the EMO-DB database. Therefore, in order to obtain robust acoustic models that can perform well with affective speech, we adapted the speaker-independent HMM/GMM models trained on natural speech data from the Kiel dataset. Various adaptation techniques have

Speaker	GBC		3 base classes		RCT	
ID [%]	*Acc*	*Corr*	*Acc*	*Corr*	*Acc*	*Corr*
03	94.02	94.50	94.02	94.26	95.22	95.22
08	83.67	84.23	82.93	83.30	85.71	86.09
09	81.15	83.25	83.77	85.86	81.94	84.29
10	82.26	82.90	82.26	83.87	83.23	83.87
11	90.50	90.89	89.70	90.10	90.89	91.29
12	90.12	90.51	89.72	90.12	93.68	93.68
13	92.28	92.45	91.94	92.28	92.97	92.97
14	90.00	90.92	89.69	90.77	90.62	91.69
15	93.29	93.49	93.89	94.08	94.08	94.28
16	73.87	74.34	73.56	74.34	73.72	74.34
Total	**86.95**	**87.56**	**86.91**	**87.62**	**87.87**	**88.43**

Table 5.4: *ASR recognition rates [%] for HMM/GMM models trained on the Kiel database, MLLR adapted on EMO-DB neutral speech samples, evaluated on the EMO-DB database with LOSO*

been used for this purpose: Maximum Likelihood Linear Regression (MLLR) (see section 3.2.9.2) with global base class (GBC) presented in listing 3.2 on page 65, 3 base classes (silence with short pause, vowels and consonants in three different base classes) presented in listing 3.3 on page 65 and regression class tree (RCT), Maximum a Posteriori (MAP) (see section 3.2.9.1) and combined MLLR(RCT)+MAP. For the MLLR, optimal performance was obtained with 39 regression classes where only means are transformed. For the MAP, optimal performance was obtained with $\tau = 10$ which is the MAP parameter which controls the impact of the MAP prior, see equation 3.47 on page 63.

First, we used for adaptation only neutral speech samples from the EMO-DB database for acoustic channel adaptation. We applied the MLLR adaptation technique with global base class (GBC), three base classes and the regression class tree (RCT).

Recognition rates of the basic ASR models adapted with MLLR on neutral speech samples and evaluated with bigram language model can be found in Table 5.4. As one can see from Table 5.4, adaptation on neutral speech samples from EMO-DB does have an insufficient impact on the recognition of the affective speech samples from the same database (recorded within the same acoustic channel). This has been found to yield a slight gain (about 0.5%) over the basic ASR models (*accuracy* 87.37%) trained on neutral speech samples from the Kiel database.

Secondly, we used for adaptation affective samples from the EMO-DB database. We applied LOSO strategy and the MLLR adaptation technique for basic ASR models trained on neutral speech samples from the Kiel database.

Recognition rates of the basic ASR models adapted with MLLR on affec-

Speaker ID [%]	GBC		3 base classes		RCT	
	Acc	*Corr*	*Acc*	*Corr*	*Acc*	*Corr*
03	94.50	94.74	94.74	94.74	96.41	96.41
08	87.38	88.13	87.57	87.94	94.25	94.25
09	87.43	88.74	87.96	89.01	93.19	93.98
10	85.16	85.81	85.16	85.81	87.10	87.74
11	90.89	91.29	91.09	91.49	94.85	95.25
12	94.07	94.07	94.07	94.07	96.44	96.44
13	95.20	95.37	94.17	94.17	98.63	98.63
14	90.31	90.92	90.00	90.92	95.69	96.00
15	94.67	94.67	94.28	94.28	95.86	95.86
16	82.12	82.43	82.58	83.83	91.29	91.91
Total	**90.00**	**90.44**	**89.96**	**90.46**	**94.57**	**94.84**

Table 5.5: *ASR recognition rates [%] for HMM/GMM models trained on the Kiel database, MLLR adapted on EMO-DB affective speech samples, evaluated on the EMO-DB database with LOSO*

tive speech samples and evaluated with the bigram language model can be found in Table 5.5. As one can see from Table 5.5 adaptation on the affective speech from EMO-DB does have a sufficient impact on the recognition of the affective speech samples. In contrast to the ASR models trained on the neutral speech samples from the Kiel database the accuracy of affective speech recognition with the MLLR (RCT) adapted HMM/GMM models was about 7.2% absolute better than that of the basic ASR models (*accuracy* 87.37%). It is well-known that MLLR and MAP can be effectively combined to im-

Speaker ID [%]	MAP				MLLR+MAP	
	neutral speech		affective speech		affective speech	
	Acc	*Corr*	*Acc*	*Corr*	*Acc*	*Corr*
03	94.02	94.26	95.93	95.93	96.89	96.89
08	84.97	85.71	91.28	91.47	96.85	96.85
09	82.98	85.08	91.36	91.88	96.60	96.86
10	83.23	83.55	83.87	84.52	85.16	86.77
11	92.08	92.08	94.46	94.65	95.64	95.84
12	92.89	93.28	96.05	96.05	96.84	96.84
13	94.00	94.17	96.74	96.74	98.63	98.63
14	88.62	89.69	93.54	94.46	98.00	98.62
15	94.48	94.67	95.27	95.27	96.25	96.25
16	76.36	77.14	87.87	88.65	96.73	96.89
Total	**88.10**	**88.71**	**92.73**	**93.09**	**96.24**	**96.49**

Table 5.6: *ASR recognition rates [%] for HMM/GMM models trained on the Kiel database, MAP or MLLR(RCT)+MAP adapted on EMO-DB affective speech samples, evaluated on the EMO-DB database with LOSO*

prove speech-recognition performance [Wong and Mak, 2000] by using MLLR transformed mean values as the priors for the MAP adaptation method. As a result we decided to use combined MLLR with regression class tree and the MAP method for adaptation on affective speech material.

Recognition rates of the basic ASR models adapted with MAP or combined MLLR(RCT)+MAP on affective speech samples from EMO-DB database and evaluated with bigram language model can be found in Table 5.6. As one can see from Table 5.6 the accuracy of affective speech recognition with the combined MLLR(RCT)+MAP adapted HMM/GMM models was about 8.9% absolute better than that of the basic ASR models (*accuracy* 87.37%).

For our ASR engine integrated into the NIMITEK demonstrator we used the HMM/GMM models trained on the Kiel database material and adapted with MLLR(RCT) on affective speech samples from the EMO-DB database. Also, for phoneme-level emotion recognition we use ASR models adapted with MLLR(RCT). Our first results of affective-speech-recognition evaluations with ASR models adapted on emotional speech data can be found in [Vlasenko and Wendemuth, 2009b].

5.3 Emotion-recognition methods evaluation

This section presents the development of experiments on emotion recognition from speech. We present experiments on all emotion-classification methods presented earlier in Chapter 4. Speech-based emotion recognition is a comparably new research field. In comparison with acoustic segments (words, phonemes) used as unit of analysis for automatic speech recognition, emotional classes used for speech-based emotion classification do not have so high discriminative characteristics. Providing "ground truth" measures for emotional content annotation (especially for spontaneous emotions) is a way more complex task in comparison to the reliable textual transcription of ASR speech corpora. Hence, in some cases of multi-class emotion classification we obtained results which are just slightly higher than classification by chance.

5.3.1 Phoneme-level classification

In this section we describe the evaluation results for two different methods which can be used for phoneme-level emotion classification: *emotional phoneme classes* and *vowel-level formants tracking*.

5.3.1.1 Emotional phoneme classes

First, we used ASR models adapted on affective speech samples with MLLR(RCT) to recognize a unit (sentence, word). Secondly, we generated all possible emotional or level of interest phonetic transcriptions for the recognized sentence or words by using the corresponding phoneme set (PE or PLOI), more details can be found in section 4.4.3.3. In the case of EMO-DB we considered 7 phoneme emotion models (PE) transcriptions, 5 phoneme emotion models (PE) transcription for SUSAS and 3 phoneme level of interest (PLOI) transcriptions for AVIC. Emotional phoneme or level of interest models which provide the highest recognition score are chosen.

Test-runs on EMO-DB, SUSAS and AVIC for phoneme-level models are carried out in leave-one-speaker-out (LOSO) manner to address speaker independence (SI), as required by most applications. Recognition rates of the emotional phoneme models evaluated with the bigram language model can be found in Table 5.7.

classification unit	EMO-DB	SUSAS	AVIC
word	51.0	49.5	45.8
sentence	66.2	49.5	54.1

Table 5.7: *Weighted average recalls (WA) [%] of emotion and level of interest recognition on sentence-, word-level applying phoneme-level analysis, MFCC, HMM/GMM, LOSO. Databases EMO-DB, SUSAS, AVIC*

In the case of the SUSAS dataset we have just one word per sentence. Detailed results from EMO-DB and AVIC evaluations show that some words within a sentence are classified wrong when the whole sentence is classified right. This means that emotional and level of interest activity is distributed irregularly among words inside a sentence. As a result phonemes which belong to the different words within a sentence have diverse emotions and levels of interest activity.

In Table 5.8 results are shown for emotion recognition on a word-, and phoneme-level in diverse constellations. Zero-gram for word-level analysis

Language model	WA
word-level zero-gram	32.1
phoneme-level bigram	38.8

Table 5.8: *Weighted average recalls (WA) [%] of emotion recognition on word-, and phoneme-level applying phoneme emotion models, dynamic features, HMM, LOSO. Evaluated on the EMO-DB database*

shows many insertions, hence low accuracy. Bi-gram LM can balance the insertions by grammar scale factor, hence higher accuracy. This is also the reason why phoneme-level accuracy is only reported with the bi-gram language model: zero-gram leads here to too-high insertion rates.

5.3.1.2 Vowel-level formants tracking

For affective speech we decided to use the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [Burkhardt et al., 2005] and The Vera am Mittag (VAM) corpus [Grimm et al., 2008]. The EMO-DB contains acted emotional speech samples. 10 professional actors (5 male and 5 female) spoke 10 German emotionally undefined sentences. Within our evaluation we used only 20 *neutral* utterances for training (2 utterances per speaker). The EMO-DB test set included *neutral (rest 58 utterances)*, low-arousal emotions (*boredom (79), sadness (53)*) and high-arousal emotions (*anger (127), fear (55)* and *joy (64))*. In total we used 456 utterances. Each of these utterances has a level of naturalness not less than 60% and a level of recognizability not less than 80%, as indicated by the raters.

The VAM database consists of 12 hours of audio-visual recordings taken from a German TV talk show. The corpus contains 947 utterances with spontaneous emotions from 47 guests of the talk show which were recorded from unscripted, authentic discussions. A large number of human labelers were used for annotation (17 labelers for one half of the data, six for the other). The labeling is based on a discrete five-point scale for three dimensions (valence, arousal, dominance) mapped onto the interval of [-1,1]. For our evaluations we use only arousal measures received with an *evaluator weighted estimator*. For training we selected one utterance per speaker with smallest absolute arousal value (19 *negative* and 28 *positive* arousal emotional utterances at all). The VAM test set included 483 *negative* and 417 *positive* arousal emotional utterances.

In order to execute a vowel-level analysis a phoneme-level ASR transcription is needed, which requires a corresponding lexicon containing phonetic transcription of words presented in a corpus. Unfortunately, the VAM corpus does not provide such a lexicon, so we created it by ourselves using a combined approach. The major part of the word transcriptions (1216 items) was taken from other German corpora, namely Verbmobil [Hess et al., 1995] and SmartKom [Schiel et al., 2002]. For the rest (688 words) we created transcriptions using grapheme-to-phoneme conversion with a Sequitur G2P converter [Bisani and Ney, 2008]. The converter was trained on a joined lexicon based on SmartKom and Verbmobil lexicons (12460 German words at all). Prior to applying the G2P software to the missing VAM lexicon, we tested it

on the constructed united lexicon, where 1% of randomly selected words were moved into the test set. The phoneme error rate was 5.33% (56 from 1050), the word error rate was 29.13% (37 from 127). In later experiments (force alignment for vowel boundaries extraction) the quality of the vowel boundaries specification cannot be expected to be absolutely reliable because of the word error rate (WER) about 29.13% which is intrinsically due to the transcription process. We decided to use this inaccurate method, because further transcription improvement required professional phonetician expert to reliable transcription and additional development expenses. In addition, we use the Sequitur G2P converter only for one-third of words presented in required lexicon, another two-third words transcription have been adopted from available lexicons.

As evaluation measures we employ the weighted (WA, i.e. accuracy) and unweighted (UA) average of class-wise recall rates. For estimation of the η values, which is only one parameter of our classifier, we applied a leave-one-speaker-out (LOSO) strategy. We used two different optimization criteria: maximum unweighted and maximum weighted average recall. For each speaker we estimated the optimal η values based on utterances from other speakers presented in the corresponding database.

In Figure 5.1 and Figure 5.2 one can see the UA and WA rates of our two

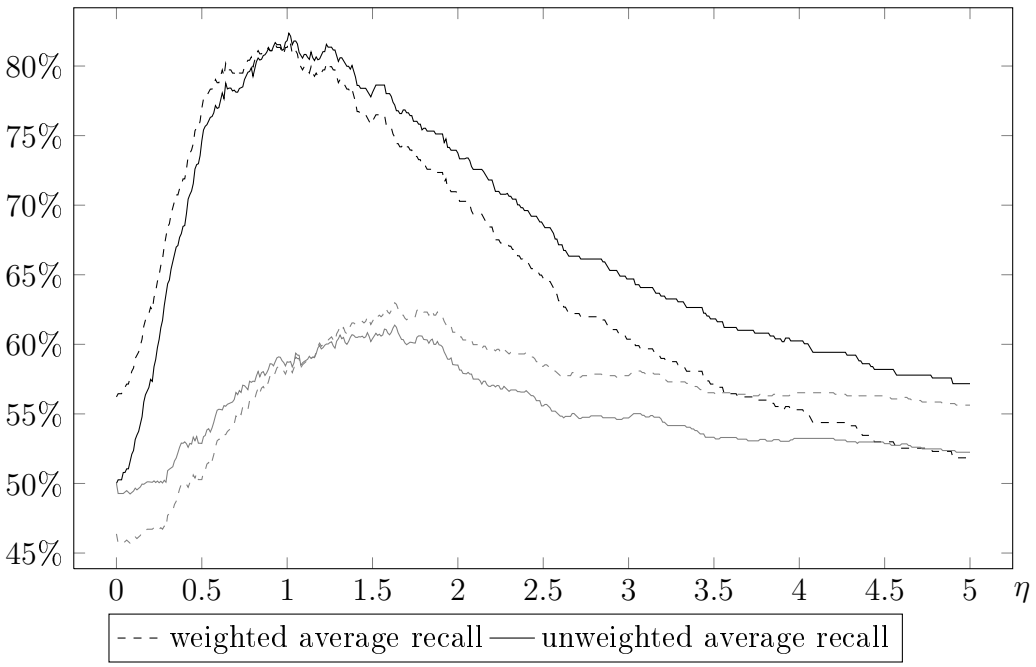


Figure 5.1: *Recognition rates of the two-class emotion classifier. Black - EMO-DB, gray - VAM*

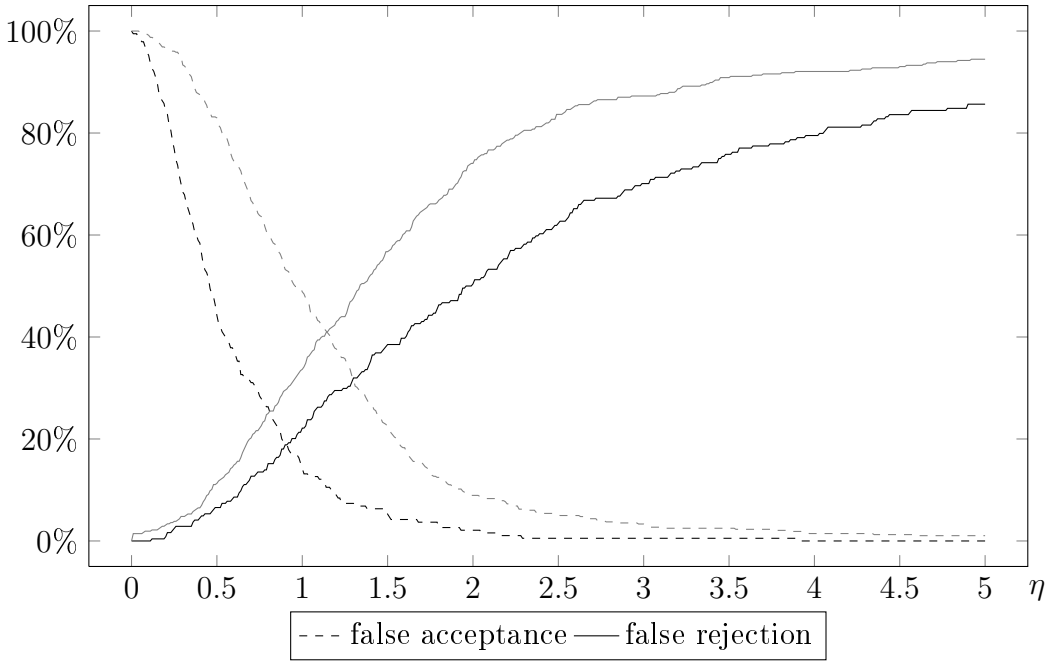


Figure 5.2: *Receiver operating characteristics curve, for high-arousal emotion detection task. Black - EMO-DB, gray - VAM*

class emotion classifier and receiver operating characteristics (ROC) which represent the false acceptance (FA) and false rejections (FR) rates for the high-level arousal emotions detection task as a function of η.

In Table 5.9 one can see performances of the two class emotion classifier for $\eta = 1$ and η values estimated within LOSO (with UA and WA as optimization criteria).

With LOSO strategy and UA optimization criteria we found the optimal η value for each speaker; these values are in range $1.01 \leq \eta \leq 1.23$ (EMO-DB) and $1.37 \leq \eta \leq 1.63$ (VAM). In the case of WA optimization criteria optimal η values are follows: $0.62 \leq \eta \leq 1.01$ (EMO-DB) and $\eta = 1.63$ (VAM).

By using gender-dependent models instead of speaker-dependent models

	EMO-DB				**VAM**			
	UA	WA	FA	FR	UA	WA	FA	FR
UA	81.3	80.6	13.1	24.2	60.2	61.8	18.7	60.8
WA	79.4	79.3	19.5	21.7	61.4	63.0	16.4	60.8
$\eta = 1$	81.8	81.3	14.2	22.1	58.7	58.2	48.9	33.7

Table 5.9: *Recognition rates [%] of vowel-level emotion classifier with different optimization strategies (UA,WA, $\eta = 1$) evaluated on EMO-DB and VAM corpora*

we can provide a statistically sufficient number of instances for the calculation of $\mu_{ig(k)}$ estimations. Due to more accurate mean values estimations we improve our results presented in [Vlasenko et al., 2011a], [Vlasenko et al., 2011b]. The presented results can be compared with the results presented in [Schuller et al., 2009]. In this article, we presented benchmark evaluation results for two-class emotion-recognition task (positive/negative arousal) with a HMM/GMM general model described in detail in section 4.4.3.1. We reached UA rates of up to 91.5% for EMO-DB and 76.5% for VAM. In our current research, instead of using 39 MFCC we used only one average $F1$ value. In contrast to HMM/GMM we used a straightforward Neyman-Pearson criterion. In the case of a priory defined η our classifier does not require any affective speech samples for training. Within practical application of our simple method the η value can be selected based on task-oriented balance between FA and FR rates, see Figure 5.2.

These results can be also compared with the results presented earlier in our paper [Schuller et al., 2008]. In this paper, we reached an emotion-recognition accuracy rate on EMO-DB database with phoneme-level analysis (see section 4.4.3.3) of up to 66.2%. Instead of using 41 phonemes for emotion recognition, we used only 7 indicative vowels. In the current approach we used only one Gaussian for each phoneme model instead of $3 \times 32 = 96$ Gaussians used in [Schuller et al., 2008]. Also our results can be improved by using more than two neutral utterances for the estimation of the mean values. Starting from our simple classifier, we can develop a more complex classification technique and provide better results.

We showed that the average F1 values extracted on a vowel-level are strongly correlated with the speaker's level of arousal. We estimated the optimal criteria thresholds for acted and spontaneous emotions. It has been shown that spontaneous emotions required higher η values. Most of the state-of-the-art emotion recognizers required sufficient amount of affective speech samples within the training phase. In the case of a priory defined η (for example $\eta = 1$) value within the training phase we require only one or two neutral (or close to "neutral" for VAM dataset) speaking style samples. As a result our method can be easily implemented for speaker-independent emotion classification.

5.3.2 Utterance-level emotion classification with dynamic and static analysis

In this section we provide results of the benchmark comparison [Schuller et al., 2009] under equal conditions on nine standard emotional speech corpora in the field using the two pre-dominant paradigms: dynamic analysis on a frame-

Corpus	All		Arousal		Valence	
	UA	WA	UA	WA	UA	WA
ABC	48.8	57.7	71.5	74.7	81.1	81.2
AVIC	65.5	66.0	74.5	77.5	74.5	77.5
DES	45.3	45.3	82.0	84.2	55.6	58.0
EMO-DB	73.2	77.1	91.5	91.5	78.0	80.4
eNTERFACE	67.1	67.0	74.9	76.8	78.7	80.5
SAL	34.0	32.7	61.2	61.6	57.2	57.0
SmartKom	28.6	47.9	58.2	64.6	57.1	68.4
SUSAS	55.0	47.9	56.0	68.0	67.3	67.8
VAM	38.4	70.2	76.5	76.5	49.2	89.9
Mean	50.7	56.9	71.8	75.0	66.5	73.4

Table 5.10: *Recognition rates [%] for benchmark evaluation of the dynamic-analysis-based emotion-recognition engine*

level by means of hidden Markov models and static analysis (supra-segmental) by systematic feature brute-forcing. The corpora investigated were the ABC, AVIC, DES, EMO-DB, eNTERFACE, SAL, SmartKom, SUSAS, and VAM databases. To provide better comparability among sets, we additionally cluster each of the database's emotions into binary valence and arousal discrimination tasks, see section 2.7.

For all databases, test-runs are carried out in the leave-one-speaker-out (LOSO) or leave-one-speakers-group-out (LOSGO) manner to face speaker independence, as required by most applications. In the case of 10 or fewer speakers in one dataset we applied the LOSO strategy; otherwise, namely for the AVIC, eNTERFACE, SmartKom, and VAM databases, we selected 5 speaker groups with almost equal amount of male and female speakers and samples per group for LOSGO evaluation. For evaluation measures we employed weighted (WA, i. e. accuracy) and unweighted (UA, thus better reflecting unbalance among classes) average recall.

The results for frame-level (Table 5.10) and supra-segmental modeling (Table 5.11) with openEAR toolkit described in section 4.4.2.1 are found for all emotion classes contained per database and for the clustered two-class tasks of binary arousal and valence discrimination as described in section 2.7.

Note that for supra-segmental modeling SVM with speaker standardization in constant parameterization are used for the given results. The delta of the mean in Table 5.11 to the mean of the best-performing individual configurations is 1.7 % (UA) and 0.7 % (WA) for class-wise results, 0.2 % (UA) and 1.8 % (WA) for arousal and 9.4 % (UA) and 9.5 % (WA) for valence (mostly due to variations on SAL).

Among the two result tables, very similar trends can be observed: the

Corpus	All		Arousal		Valence	
	UA	WA	UA	WA	UA	WA
ABC	55.5	61.4	61.1	70.2	70.0	70.0
AVIC	56.5	68.6	66.4	76.2	66.4	76.2
DES	59.9	60.1	87.0	87.4	70.6	72.6
EMO-DB	84.6	85.6	96.8	96.8	87.0	88.1
eNTERFACE	72.5	72.4	78.1	79.3	78.6	80.2
SAL	29.9	30.6	55.0	55.0	50.0	49.9
SmartKom	23.5	39.0	59.1	64.1	53.1	75.6
SUSAS	61.4	56.5	63.7	77.3	67.7	68.3
VAM	37.6	65.0	72.4	72.4	48.1	85.4
Mean	53.5	59.9	71.1	75.4	64.5	68.3

Table 5.11: *Recognition rates [%] for benchmark evaluation of the static-analysis-based emotion-recognition engine*

best performance is achieved on the datasets containing acted, prototypical emotions, where only emotions with high inter-labeler agreement were selected (EMO-DB, eNTERFACE datasets). A little exception here is the DES database, where performance is well behind EMO-DB database, even though the DES dataset also contains acted, prototypical emotions. This difference is not so obvious for the arousal task as it is for the full classification task. One reason for this might be that no selection of high inter-labeler agreements were done on the DES dataset and labelers may agree more upon arousal than on the emotion categories. The remaining six corpora are more challenging since they contain non-acted or induced emotions. On the lower end of recognition performance the SAL, SmartKom, and VAM corpora can be found, which contain the most spontaneous and naturalistic emotions, which in turn are also the most challenging to label. However, the SmartKom database contains long pauses with a high noise level, and it includes system output cross-talk segments and annotations that are multi-modal, i.e. mimic and audio based, thus the target emotion might not always be detectable from speech. The results for the SAL corpus are only marginally above chance level, which is due to speaker-independent evaluation on highly naturalistic data with only four speakers in total.

When comparing the dynamic analysis with static analysis an interesting conclusion can be drawn: dynamic analysis seems to be slightly superior for corpora containing variable content (AVIC, SAL, SmartKom, VAM), i.e. the subjects were not restricted to a predefined script, while static analysis outperforms frame-level modeling on corpora where the topic/script is fixed (ABC, DES, EMO-DB, eNTERFACE, SUSAS), i.e. where there is an overlap in verbal content between test and training set. This can be explained by the

nature of supra-segmental modeling: in corpora with non-scripted content, turn lengths may strongly vary. While frame-level modeling is mostly independent of highly varying turn length, in supra-segmental modeling each turn gets mapped onto one feature vector, which might not always be appropriate.

5.3.3 Combined analysis

In this section we describe evaluation results for two possible combined speech-based emotion-classification techniques: *two-stage processing and middle-level fusion.*

5.3.3.1 Two-stage processing

Within this section we present a number of results for the two-stage processing method presented in section 4.4.4. Evaluation test-runs are realized in leave-one-speaker-out (LOSO) manner for speaker-independent tests. For evaluation we used the EMO-DB database.

WA [%]	SN	FS	EMO-DB
Turn	-	-	74.9
Turn	✓	-	79.6
Turn	✓	✓	83.2

Table 5.12: *Baseline results by turn-level analysis. Weighted average recalls [%] for EMO-DB, turn-wise feature extraction, considering speaker-normalization (SN), and feature selection (FS) for optimization, speaker-independent (SI) LOSO evaluation with SVM*

In Table 5.12 we present the baseline results for speaker-independent classification on the turn-level described in section 4.4.2 employing standard turn-wise derived acoustic features presented in Table 4.2 on page 83.

[#]	Turns	Chunks	Syllables
anger	127	269	1,843
boredom	79	225	1,151
disgust	38	173	516
fear	55	160	794
joy	64	179	927
neutral	78	213	1,093
sadness	53	143	823
sum	494	1,362	7,147

Table 5.13: *Distribution among emotions, database EMO-DB. Considered are turns, automatically extracted chunks and syllables*

[#]	Chunks	Syllables
1	167	-
2	86	-
3	95	-
4	65	-
5-9	78	94
10-14	3	135
15-19	-	156
20-29	-	109

Table 5.14: *Number of automatically extracted chunks and syllables per utterance. Database EMO-DB*

Table 5.13 presents a detailed number of automatically extracted chunks and syllables per emotion obtained by HMMs-/GMM-based one-pass Viterbi beam search with token passing within the first stage of processing. As one can see, automatically extracted chunks comparably longer than syllables. Note that an almost constant factor of chunks per emotion resembling 3 is obtained [Schuller et al., 2007]. Disgust, however, shows a slightly different behavior. Apart from the mean number of chunks and syllables per emotion, Table 5.14 depicts their frequencies of appearance in more detail.

Table 5.15 below presents the emotion-recognition results for chunks and syllables, aimed at sub-turn entities. As for the base-line turn-level features, speaker normalization and feature space optimization are applied for optimization. Finally, we present results for the mapping of chunks or syllables onto turns by the diverse strategies: an un-weighted majority vote (MV), a maximum length vote (MLV), a maximum classifier prediction score multiplied with the length vote (MSL) introduced in section 4.4.4. The second stage of processing, based on the chunk analysis, is realized by brute-force large feature space construction with subsequent subset selection, support vector

WA [%]	SN	FS	EMO-DB
Chunk	-	-	42.6
Chunk	✓	-	46.7
Chunk	✓	✓	51.4
Syllable	-	-	42.1
Syllable	✓	-	44.6
Syllable	✓	✓	47.6

Table 5.15: *Results by chunk-level analysis. Weighted average recalls [%] for EMO-DB, chunk-wise feature extraction, considering speaker-normalization (SN), and feature selection (FS) for optimization, speaker-independent LOSO evaluation with second-stage static analysis*

	Strategy	Correct	Correct*
Chunk	MV	45.3	64.2
Chunk	MLV	60.1	64.2
Chunk	MLS	70.6	70.6
Syllable	MV	42.8	60.1
Syllable	MLV	56.9	60.1
Syllable	MLS	67.8	67.8

Table 5.16: *Results by turn-level mapping. Weighted average recalls [%] for EMO-DB, chunk-wise features with speaker-normalization and feature selection, considering Correct and Correct* cases, by addition of non-unique winning-classes, speaker-independent LOSO evaluation with second-stage static analysis*

machines (SVM) classification, and speaker normalization.

Thereby only the optimal cases with speaker normalization and feature space optimization are considered, as chunk-level accuracy is crucial for the overall success. First, we describe the speaker-independent evaluation results presented in Table 5.16. Thereby the three strategies: majority vote (MV), maximum length (MLV) and maximum length times prediction score (MLS) are considered.

As can be seen form Table 5.16, we discriminate between correct assignment (column Correct) and cases, where the correct class has been the winning class among one or more other emotional classes (column Correct). The main outcomes of these results are that the proposed chunking seems superior to annotation-based syllable chunking. However, recognition results with turn-level acoustic features cannot be reached. This holds even after mapping on the turn-level by the investigated three different strategies.

The introduced two-stage processing approach was superior to syllables speaker-independent analysis. This may be due to the fact that it produces roughly 5 times longer segments, though at the same time 5 times fewer instances are obtained for robust training. Still, results for both of these sub-turn entities clearly fall behind those for turn-level analysis. We secondly investigated mapping of these context-independent chunks on the turn level by multi-instance learning. Yet, as a result for the evaluated database no advantage over direct turn-level acoustic feature extraction can be reported. However, no turn-level feature information was integrated, which may lead to an advantage as reported in [Schuller and Rigoll, 2006], where chunk- and turn-level features were integrated in one super-vector.

5.3.3.2 Middle-level fusion

With this combined method we integrated frame-level information within a state-of-the-art large feature space static analysis for speaker's emotion recognition [Vlasenko et al., 2007a]. In order to fuse this information with turn-based modeling, output scores are added to a super-vector combined with static acoustic features. Thereby a variety of low-level descriptors and functionals to cover prosodic, speech quality, and articulatory aspects are considered. Starting from 1,406 acoustic features presented in Table 4.2 we selected optimal configurations including and excluding emotion-recognition scores from HMM-/GMM-based classifier. The final decision task is realized by use of SVM. Extensive test-runs are carried out on two popular public databases, namely EMO-DB and SUSAS, to investigate acted and spontaneous data.

Emotion-recognition results are presented for each modeling technique individually (turn-level (TL) and frame-level (FL)), and for the combination of these two. Thereby the effects of speaker normalization (SN) and feature space optimization (FS) as described in section 4.4.1 are shown, too. For the EMODB database, we provide results of a leave-one-speaker-out (LOSO) evaluation to face the challenge of speaker independence. For the SUSAS database we used 10-fold stratified cross-validation (SCV), as only 7 speakers are contained in the chosen spontaneous emotional speech subset. On the other hand, this is possible, as roughly 500 phrases are available per speaker.

During feature selection the original 1,406 features have been reduced to 76 for the EMODB dataset. For the SUSAS 71 features have been selected on the whole dataset, and 33–107 features were observed as optimum for the individual speakers. This underlines the brute-force nature of the creation of feature space with more than 1,000 acoustic features in order to find a very

WA [%]	SN	FS	EMO-DB	SUSAS
TL	-	-	74.9	80.8
TL	✓	-	79.6	80.8
TL	✓	✓	83.2	80.8
FL	-	-	77.1	67.1
TL+FL	✓	-	81.6	81.3
TL+FL	✓	✓	89.9	83.8

Table 5.17: *Combination of turn-level and frame-level analysis, databases EMODB with LOSO evaluation and speaker-dependent 10-fold SCV for SUSAS. TL and FL abbreviate turn and frame levels. SN and FS represent speaker normalization and feature space optimization. (✓) indicates that the technique has been applied*

compact robust final set. Table 5.17 shows the summarized results.

As one can see from Table 5.17, speaker normalization and feature space optimization both clearly help to improve overall results. Thereby it has to be noted that less than 10 % of the original feature space suffices to get an optimum performance. The highest accuracy is however obtained by the suggested fusion of both approaches. This is particularly true for the EMODB dataset. For the SUSAS dataset it is not too clear whether the extra effort is justified or not.

5.3.4 Interspeech 2009 Emotion Challenge

A CEICES initiative [Batliner et al., 2006] was the first cooperative emotion-recognition experiment, where seven sites compared their classification results under exactly the same conditions and fused their acoustic features together for a combined emotion indicative acoustic features selection process. This challenge was not public, which motivates the INTERSPEECH 2009 Emotion Challenge [Schuller et al., 2009c] to be organized with strict comparability, using the same emotional speech database. Three sub-challenges are addressed using non-prototypical five or two emotion classes (including a garbage model): *Open Performance Sub-Challenge, Classifier Sub-Challenge, and the Feature Sub-Challenge*. We participated in the Open Performance Sub-Challenge, where we evaluated our developed acoustic features and classification algorithm.

Due to the unbalanced number of the emotional class instances included in training and test sets, the primary emotion-recognition measure to optimize is unweighted average (UA) recall, and secondly the weighted average (WA) recall (i.e. accuracy). For tuning our classifier we used affective speech samples from the training set and the LOSO strategy. Afterwards we used an optimal

Level of analysis	**Classes** [#]	**UA**	**WA**
Utterance	2	69.21	70.36
Phonemes	2	68.09	73.26
Combined	2	68.45	70.35
Baseline	2	67.7	65.5
Utterance	5	41.40	47.44
Phonemes	5	35.21	52.78
Combined	5	40.62	49.38
Baseline	5	38.2	39.2

Table 5.18: *Recognition rates [%] on test set of FAU AIBO database within INTERSPEECH 2009 Emotion Challenge. Baseline results are taken from [Schuller et al., 2009c]*

[#/%]	**NEG**	**IDL**
NEGative	**1,635**	830
IDLe	1,617	**4,175**
NEGative	**66.3%**	33.7%
IDLe	27.9%	**72.1%**
All [#]	**2,465**	**5,792**

Table 5.19: *Confusion matrix for the two-classes emotion-recognition task and accuracies for each class individually and complete test set*

emotion classifier configuration with corresponding acoustic features set for challenge trials on test set material. The best results obtained on the challenge test set and baseline results provided by organizers are presented in Table 5.18. Baseline results were adopted from [Schuller et al., 2009c], they represent the baseline of emotion-recognition performance for static modeling. Baseline results for dynamic modeling presented by organizers were comparably lower.

The best results for two classes (**NEG***ative* and **IDL***e*) were achieved with utterance-level analysis with the feature set which included 12 MFCC coefficients normalized with gender-dependent vocal tract length normalization, energy and their deltas and acceleration. For the five classes (**A***nger*, **E***mphatic*, **N***eutral*, **P***ositive* and **R***est*) emotion-recognition task the best results were received with 13 MFCC coefficients normalized by gender-dependent vocal tract length normalization after CMS included zero coefficient instead of energy and their delta and acceleration. Confusion matrices for the best results for two-class and five-class task are presented in Tables 5.19 and 5.20.

As one can see from Table 5.19 for *NEG* class false acceptance error is quite high. This confusion can be explained by low discriminative acoustic diversity of some *NEG* and *IDL* subclasses (i.e. emphatic vs. motherese). List of all subclasses covered by emotional categories (**NEG***ative* and **IDL***e*) can be found in section 2.6.2.1.

In the case of the five emotion classes evaluation, classes are unbalanced in the training set, see Table 2.4 on page 26. As a result, we have to be very careful with over tuning of sparse emotional classes like *Positive, Rest.* As one can see from Table 5.20, there is quite high confusion among the leaders of the emotion classification: *Anger, Emphatic* and *Neutral.* At the same time *Positive* and *Rest* have a high level of confusion with all other emotional classes. We suppose that the main reason of so a high level of confusion among all five emotional classes lies in unreliable emotional annotation. We think that students, like any other adult who is not a natural relative to the child, could not provide reliable emotional annotation of the child's emotional speech even thought they are advanced students of linguistics.

[#/%]	**A**	**E**	**N**	**P**	**R**
Anger	**315**	189	67	9	31
Emphatic	202	**944**	276	10	76
Neutral	592	1,551	**2,485**	217	532
Positive	17	17	90	**53**	38
Rest	95	108	176	47	**120**
Anger	**51.6%**	30.9%	11.0%	1.5%	5.0%
Emphatic	13.4%	**62.6%**	18.3%	0.7%	5.0%
Neutral	11.0%	28.8%	**46.3%**	4.0%	9.9%
Positive	8.0%	8.0%	41.8%	**24.6%**	17.6%
Rest	17.4%	19.8%	32.2%	8.6%	**22.0%**
All [#]	**611**	**1,508**	**5,377**	**215**	**546**

Table 5.20: *Confusion matrix for the five-classes emotion-recognition task and accuracies for each class individually and complete test set*

The results of the challenge were presented at a special session of the conference Interspeech 2009. A ranking list of the best results can be found in Table 5.21.

As one can see from Table 5.21, we got second place for the two emotion classes task and forth place for the five emotion classes task over 33 research

Rank	**UA**[%]	**WA**[%]	**Authors**
two emotion classes task			
1	70.29	68.68	[Dumouchel et al., 2009]
2	69.21	70.36	[Vlasenko and Wendemuth, 2009b]
3	68.33	65.84	[Kockmann et al., 2009]
4	67.90	63.03	[Bozkurt et al., 2009]
5	67.19	63.26	[Luengo et al., 2009]
6	67.55	72.67	[Polzehl et al., 2009]
7	67.06	62.29	[Barra-Chicote et al., 2009]
8	66.40	66.56	[Vogt and André, 2009]
five emotion classes task			
1	41.65	44.17	[Kockmann et al., 2009]
2	41.59	44.17	[Bozkurt et al., 2009]
3	41.57	39.87	[Lee et al., 2009b]
4	41.40	47.44	[Vlasenko and Wendemuth, 2009b]
5	41.38	43.35	[Luengo et al., 2009]
6	39.40	52.08	[Dumouchel et al., 2009]
7	39.40	41.12	[Vogt and André, 2009]
8	38.24	36.68	[Barra-Chicote et al., 2009]

Table 5.21: *Results and ranking list for two emotion classes and five emotion classes INTERSPEECH 2009 Emotion Challenge. Data for ranking list are taken from [Schuller et al., 2011]*

groups registered to get access to the data [Schuller et al., 2011]. In total our classification results (sum of unweighted average recalls for two tasks) are the best. With our emotion-classification technique we prove that only by using spectral features (Mel-frequency Cepstral coefficients (MFCC)) with dynamic analysis we can reach one of the best emotion-recognition performances for spontaneous emotional speech samples [Vlasenko and Wendemuth, 2009b].

5.3.5 Cross-corpus acoustic emotion recognition

A great advantage of cross-corpora evaluations is the well definedness of test and training datasets and thus the easy reproducibility of the results. Since most emotion corpora, in contrast to speech corpora for automatic speech recognition or speaker identification, do not provide fixed training, development, and test sets, individual splitting and cross-validation are mostly found, which makes it hard to reproduce the results under equal conditions. In contrast to this, cross-corpus experiments are well defined and thus easy to reproduce and compare.

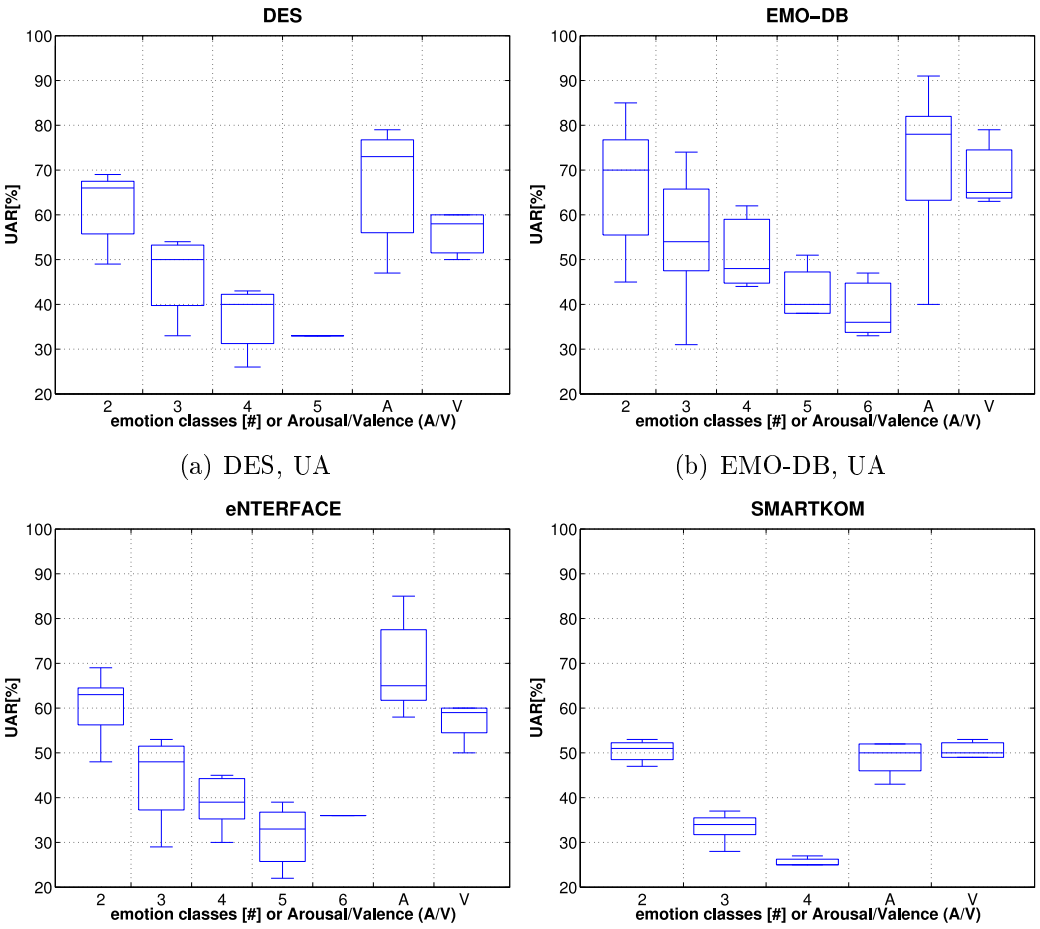
In Table 5.22 one can find a list of all 23 different training and test set combinations which have been used for evaluation in our cross-corpus experiments. Affective speech samples from the SUSAS and AVIC databases are only used for training, since they do not cover the sufficient overlapping "basic" emotions for the testing. Furthermore, we omitted combinations for which the number of emotion classes occurring were lower than tree in both the training and the test dataset (e.g. we did not evaluate training on AVIC database material and testing on DES database affective speech samples, since only *neutral* and *joyful* occur in both corpora – see also Table 2.3 on page 24). In order to obtain combinations for which up to six emotion classes occur in the training and test set, we included evaluations in which more than one dataset was used for training (e.g. we combined eNTERFACE and SUSAS databases for training in order to be able to model six classes when testing on the EMO-DB database). Depending on the maximum number of different emotion classes that can be modeled in a certain experiment, and depending on the number of classes we actually use (two to six) for evaluation, we got a certain number of possible emotion class permutations according to Table 5.22. For example, if we aimed to model two emotion classes when testing on the EMO-DB database and training on the DES dataset, we obtained six possible permutations. Evaluating all permutations for all of the 23 different training-test combinations leads to 409 different evaluations (sum of the last line in Table 5.22). Additionally, we evaluated the discrimination between positive and negative arousal as well as the discrimination between high and low valence for all 23 combinations, leading to 46 additional evaluations.

Test set	Training set	number of classes				
		2	3	4	5	6
EMO-DB	AVIC	3	1	0	0	0
	DES	6	4	1	0	0
	eNTERFACE	10	10	5	1	0
	SmartKom	3	1	0	0	0
	eNTERF.+SUSAS	15	20	15	6	1
	eNTERF.+SUSAS+DES	15	20	15	6	1
DES	EMO-DB	6	4	1	0	0
	eNTERFACE	6	4	1	0	0
	SmartKom	6	4	1	0	0
	EMO-DB+SUSAS	6	4	1	0	0
	EMO-DB+eNTERFACE	10	10	5	1	0
eNTERFACE	DES	6	4	1	0	0
	EMO-DB	10	10	5	1	0
	SmartKom	3	1	0	0	0
	EMO-DB+SUSAS	10	10	5	1	0
	EMO-DB+SUSAS+DES	15	20	15	6	1
SmartKom	DES	6	4	1	0	0
	EMO-DB	3	1	0	0	0
	eNTERF.	3	1	0	0	0
	EMO-DB+SUSAS	3	1	0	0	0
	EMO-DB+SUSAS+DES	6	4	1	0	0
	eNTERF.+SUSAS	6	4	1	0	0
	eNTERF.+SUSAS+DES	6	4	1	0	0
Total		**163**	**146**	**75**	**22**	**3**

Table 5.22: *Number of emotion class permutations dependent on the used training and test set combination and the total number of classes used in the respective experiment*

To summarize the results of permutations over cross-training datasets and emotion classes groupings, box-plots indicating the unweighted average recall (UA) are shown (see Figures 5.3(a) to 5.3(d)). All recognition rates are averaged over all constellations of cross-corpus training to provide a raw general impression of performances to be expected. The plots show the median, the lower and upper quartile, and the extremes for a varying number (from two to six) of emotion classes and the binary valence and arousal tasks. In a case of DES dataset (5 classes evaluation) and eNTERFACE dataset (6 classes evaluation) we have only one permutation, as a result in the corresponding box plot's columns one can see only medians.

First, the DES dataset is chosen for testing, as depicted in Figure 5.3(a). For training, five different combinations of the remaining datasets are used (see Table 5.22). As expected the weighted (i. e., accuracy – not shown) and unweighted recall monotonously drop on average with an increased number of



(a) DES, UA (b) EMO-DB, UA

(c) eNTERFACE, UA (d) SMARTKOM, UA

Figure 5.3: *Box-plots for unweighted average recall (UA) in % for cross-corpora testing on four test corpora. Results obtained for varying number of classes (2–6) and for classes mapped to high/low arousal (A) and positive/negative valence (V)*

classes. For the DES experience holds: arousal discrimination tasks are 'easier' on average. While the average results are constantly found considerably above chance level, it also becomes clear that only selected groups are ready for real-life application – of course allowing for some error tolerance. These are two-class tasks with an approximate error of 20 %. An interpretation of the results in multi-class recognition is given below.

A very similar overall behavior is observed for the EMO-DB dataset in Figure 5.3(b). This seems no surprise, as the two databases have very similar characteristics. For the EMO-DB a more or less additive offset in terms of recall is obtained, which is owed to the known lower 'difficulty' of this dataset.

Switching from acted to mood-induced, we provide results on the eNTERFACE dataset in Figure 5.3(c). However, the picture remains the same, apart from lower overall results: again a known fact from experience, as eNTERFACE database is not a 'gentle' dataset, partially for being more natural than the DES corpus or the EMO-DB database.

Finally, considering testing on spontaneous affective speech with non-restricted varying spoken content and natural emotion, we note the challenge arising from the SmartKom dataset in Figure 5.3(d): as this set is – due to its nature of being recorded in a user-study – highly unbalanced, the mean unweighted recall is again mostly of interest. Here, rates are found only slightly above chance level. Even the optimal groups of emotions are not recognized in a sufficiently satisfying manner for a real-life usage. Though one has to bear in mind that SmartKom was annotated multimodally, i. e., the emotion is not necessarily reflected in the speech signal, and overlaid environment noise is often present due to the setting of the recording, this shows in general that the reach of our results is so far restricted to acted data or data in well-defined scenarios: the SmartKom results clearly demonstrate that there is a long way ahead for emotion recognition in user studies (cf. also [Schuller et al., 2009c]) and real-life scenarios. At the same time, this raises the ever-present and in comparison to other speech analysis tasks unique question on ground truth reliability: while the labels provided for acted data can be assumed to be double-verified, as the actors usually wanted to portray the target emotion which is often additionally verified in perception studies, the level of emotionally valid material found in real-life data is mostly unclear due to the reliance on few labelers with often high disagreement among them [Schuller et al., 2010].

5.4 Summary

This chapter reviews results of experiments concerning our developed emotion-recognition and automatic speech-recognition methods. Afterwards, we present results of evaluations on non-adapted and adapted ASR models. In section 5.2, we showed that the combined MLLR(RCT)+MAP adapted HMM/GMM models was about 8.9% absolute better than that of the basic ASR models (*accuracy* 87.37%) trained on emotionally neutral speech samples.

In section 5.3 we present evaluation results for various speech emotion-classification techniques. As a starting point for our experiments we chose phonemes, as these should provide the most flexible basis for unit-specific models: if emotion recognition is feasible on phoneme basis, these units could most easily be integrated into a user-behavior-adaptive spoken dialog sys-

tem [Vlasenko et al., 2008a]. However, the introduced unit-specific (phoneme-, word-level) emotion models clearly outperformed context-independent general models provided enough training material per unit. Appearance of high-standard word-level-labeled emotional speech corpora can improve the current performance of phoneme and word-level emotion models. A prototypical spoken dialog system with a user-behavior-adaptive spoken dialog system created within NIMITEK collaboration, which includes phoneme-level emotion recognition, will be discussed in Chapter 6. With a vowel-level formants tracing technique we showed that the average F1 values extracted on a vowel-level are strongly correlated with the level of arousal of the speaker's emotional state. We estimated the optimal criteria thresholds for acted and spontaneous emotions. It was shown that spontaneous emotions required higher η values in comparison with optimal η values for acted emotions. We showed that the list of the most indicative German vowels [Vlasenko et al., 2011a], [Vlasenko et al., 2011b] within the task of measuring the level of arousal of the speaker's emotional state can be used for spontaneous emotion classification.

When comparing the dynamic analysis with the static analysis an interesting conclusion can be drawn: frame-level modeling seems to be slightly superior for corpora containing variable content (AVIC, SAL, SmartKom, VAM), i.e. the subjects were not restricted to a predefined script, while supra-segmental modeling (turn-level analysis) slightly outperforms frame-level modeling on corpora where the topic is fixed (ABC, DES, EMO-DB, eNTERFACE, SUSAS), i.e. where there is an overlap in textual content between training and test dataset [Schuller et al., 2009]. This can be explained by the nature of static analysis: in corpora with non-fixed content, turn lengths may strongly vary. While dynamic analysis is mostly independent of highly varying turn length, in supra-segmental modeling each turn gets mapped onto one feature vector, which might not always be appropriate. In section 5.3.4, we present our results within the INTERSPEECH 2009 Emotion Challenge [Schuller et al., 2009c]. With our emotion-classification technique based on dynamic analysis we prove that only by using spectral features (Mel-frequency Cepstral coefficients (MFCC)) we can reach one of the best emotion-recognition performances for spontaneous emotional speech samples [Vlasenko and Wendemuth, 2009b].

Finally, in section 5.3.5 we present evaluation results for cross-corpus acoustic emotion recognition. To sum up, we have shown results for intra- and inter-corpus speech-based emotion recognition. By that we have learnt that the recognition rates highly depend on the specific sub-group of emotions considered. In any case, emotion-recognition performance decreases dramatically when operating cross-corpora-wise. As long as conditions remain similar, cross-corpus training and testing seems to work to a certain degree: the DES,

EMO-DB, and eNTERFACE datasets led to partly useful results [Schuller et al., 2010]. These are all rather prototypical, mood-induced or acted with pre-defined spoken content. The fact that three different languages – Danish, English, and German – are contained, seems not to generally disallow inter-corpus testing: these are all Germanic languages, and a highly similar cultural background may be assumed. However, the cross-corpus testing on a spontaneous dataset (SmartKom) clearly showed limitations of the current systems. Here only a few groups of emotions stood out in comparison to chance level. To better cope with the emotional corpora's differences, we evaluated different normalization approaches, whereas speaker normalization led to the best results. For all experiments we had used static analysis based on a broad variety of prosodic, voice quality, and articulatory features (see Table 4.3 on page 85) and SVM classification.

Chapter 6

User-behavior-adaptive dialog management

Contents

6.1 Introduction

This chapter is not dealing with the full specification of techniques for developing a spoken dialog system (SDS). For this topic, the reader is referred to the excellent survey material [Gnjatović, 2009, Gnjatović and Rösner, 2008a, Gnjatović and Rösner, 2008c]. The focus is on the incorporation of the findings described earlier in this thesis into a prototype dialog system especially developed by the author and colleagues to demonstrate the adaptation of the system to the user's emotional state. In this chapter we present a prototype of the user-friendly spoken dialog system integrated into the NIMITEK demonstrator. The NIMITEK (Neurobiologically inspired, multimodal intention recognition for technical communication systems) demonstrator is a spoken dialog system prototype which provides an "intelligent" support for users while they solve tasks in a graphics system interface (e.g., Towers-of-Hanoi puzzle). The "intelligent" feature of the system is a user-behavior-adaptive

dialog management. The system dynamically selects a dialog strategy according to the current user's emotional state. In this chapter we describe the data collection strategy within the NIMITEK Wizard of Oz experiment, and the structure of the conventional and user's behavior adaptive dialog systems. Finally we discuss the results of an interactive usability test.

6.2 Framework: NIMITEK demonstrator

This chapter presents a part of the work in the framework of the NIMITEK project [Wendemuth et al., 2008] in the period from 2005 to 2010 that includes an interdisciplinary research on human-machine interaction. Various cognitive aspects of user-friendly interfaces were investigated within the current project. Also, this interdisciplinary research combines the fields of electrical engineering, computer science and neuro-biology to carry out the study into processing of an audio-visual user's interaction interfaces, the development of a task-oriented knowledge representation and modeling different dialog situations.

The NIMITEK project has various research goals: multimodal emotion recognition from the user's speech (i.e., prosodic cues and spectral features analysis), mimic and text-based analysis; developing robust affective-speech-

Figure 6.1: *Prototype of a multimodal spoken dialog system, NIMITEK Demonstrator*

recognition models; analysis of the task-oriented interaction experiments; modeling of the adaptive dialog management; developing neuro-biological perception, cognitive and behavior models. The NIMITEK spoken dialog system prototype presented in Figure 6.1 was developed to demonstrate research achievements in emotion recognition and user's emotion adaptive dialog management.

6.3 Interface, chosen tasks and WOZ experiments

In this section we specify the main issues in developing the NIMITEK spoken dialog system prototype: *flexibility and adaptivity, interface design* and *task selection for evoking user's emotions.*

6.3.1 Flexibility and adaptivity

The importance of the user-behavior-driven dialog strategies in human-machine interaction (HMI) lies in the existing limitations of automatic speech-recognition technologies. Current state-of-the-art automatic speech-recognition (ASR) methods still cannot deal with flexible, unrestricted user's language and emotionally colored speech [Lee, 2007]. Therefore, problems caused by misunderstandings of a user during interaction with SDS with a predefined, and usually restricted set of interaction rules seems to be inevitable. In our spoken dialog system we want to provide a flexible interaction speech-based interface. In such a way the user will be able to find out suitable commands by himself.

In the domain of human-machine interaction [Gnjatović and Rösner, 2008a], we witness the rapid increase of research interest in affective user behavior. However, some aspects of the affective user behavior during HMI still turns out to be a challenge for SDS developers. Detecting and utilizing non-lexical or paralinguistic cues as part of the user-behavior state descriptors is one of the major challenges in the development of reliable human-machine interfaces. Knowing the current user's emotional state can help to adjust system responses so that the user of such a system can be more engaged and have a more effective interaction with the system [Schuller et al., 2007b], [Busso et al., 2007]. To make our system user-centered we implemented an intention recognition module, which is dealing with *motivational intention.* Psychologist also distinguish a *functional intention* [Anscombe, 2000]. But for our practical implementation we decided to concentrate on *motivational* aspects of intention. Examples will be given in section 6.5 below.

In this section we present the implementation of adaptive dialog management in the NIMITEK prototype spoken dialog system for supporting users while they solve the Towers-of-Hanoi puzzle which is displayed in Figure 6.1.

Within the human-machine interaction users are able to follow the ASR recognition results. When the garbage model was not able to encapsulate out-of-vocabulary words, the users were able to see misrecognized system perceptible commands. We expect that users will try to adapt their commands vocabulary to contribute to the right system reaction.

6.3.2 Interface design and task selection for evoking user's emotions

It is quite difficult to motivate naïve users to experience, express and utilize emotions while using any graphical application. We decided to use a graphical system with a verbal interaction interface to simulate an intelligence test. In such a way, we expected to achieve a strong user's motivation and emotional involvement. For modeling user behavior during human-machine interaction we decided to develop a spoken dialog system for simple logical games (i.e. Towers-of-Hanoi, Tangram) with system users support while they use a graphical tool. This graphical tool has been developed using an existing software package[1] that implements visual reflection, alteration and movement of different graphical objects.

In the NIMITEK demonstration system, users are allowed only to use a verbal interaction interface (i.e., mouse or keyboard interaction interfaces are not supported by system).

Two different prototypical graphical tasks were implemented in the NIMITEK demonstrator prototype: *Towers-of-Hanoi* and *Tangram*. The Towers-of-Hanoi puzzle (3-disks version) was introduced by Édouard Lucas in 1883. The puzzle consists of three pegs and three disks (small, middle and large). At the beginning of the game, the disks are stacked in order of size on the left peg, as one can see in Figure 6.2 [Gnjatović and Rösner, 2008c]. The aim of the game is to move the complete stack to the right peg shifting disks according to the following rules: only one disk can be shifted at a time, all three pegs can be used, and no disk can be located on the top of a smaller disk.

Another prototypical graphical task is the Tangram puzzle. It is a famous Chinese puzzle. Its origins are lost in time. It was introduced to the western world by a Captain M. Donaldson in 1815. The goal of this graphical task is

[1] This graphical engine was developed at the Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany.

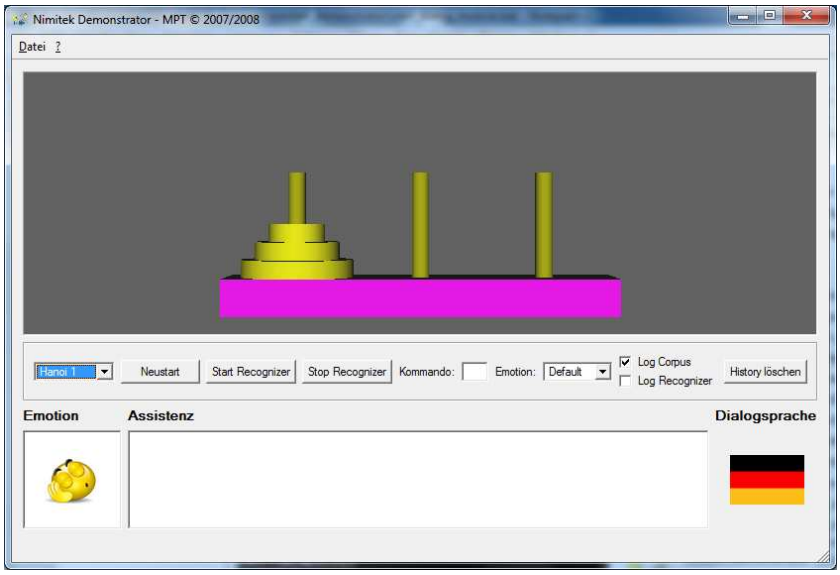


Figure 6.2: *Towers-of-Hanoi Puzzle: Screen shot of the NIMITEK demonstrator*

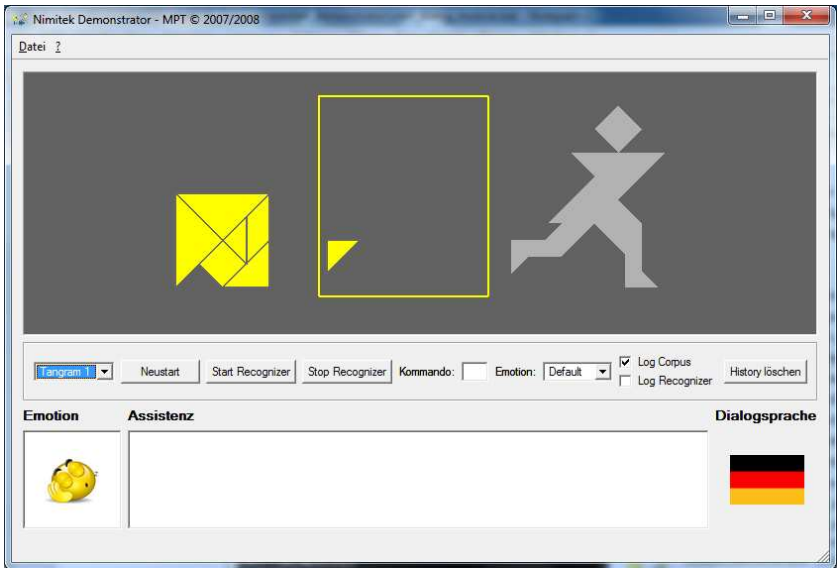


Figure 6.3: *Tangram: Screen shot of the NIMITEK demonstrator*

to seamlessly form a specific construction by using seven Tangram two dimensional objects (e.g., triangles, quadrant, rhombus). Two kinds of action over corresponding objects were possible: *relocation* and *rotation*. In Figure 6.3 one can see a screen shot of the desktop representing the NIMITEK demonstrator with an active Tangram puzzle. These two game applications were used for the experiments described below.

6.3.3 NIMITEK Wizard of Oz experiments

Affective speech corpora provide an important empirical foundation for investigation when researchers aim at implementing emotion-aware spoken dialog systems [Gnjatović and Rösner, 2010]. In this section we describe the applied Wizard of Oz (WOZ) technique in order that a scenario designed to extract emotional speech within human-machine interaction could result in useful and natural data. This data can be used for the development of a user-friendly dialog strategy. Corresponding Wizard of Oz experiments were conducted in the framework of the NIMITEK project. The schema of the laboratory settings used for the NIMITEK dataset collection is presented in Figure 6.4.

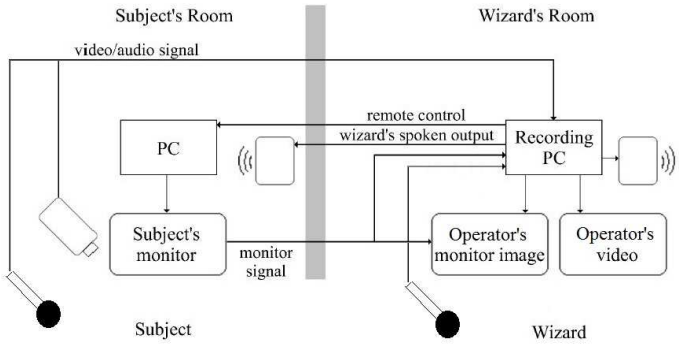


Figure 6.4: *Schema of the NIMITEK WOZ laboratory settings*

As usual for WOZ studies [Fraser and Gilbert, 1991], subjects believe they are interacting with a real spoken dialog system driven by the computer, while the assumed instructions and system's support is actually provided by a human "wizard". We used two different rooms for our experiment to hide the "wizard". A simulated spoken dialog system was installed on the subject's computer. The "wizard" pretends to have automatic speech recognition, remotely controls the interaction interface of the system, and declaims speech output of the dialog system. The video screen shots from the subject's computer desktop and the video recordings of subject (facial expressions, gestures and body movements) are displayed on two different monitors in the wizard's room.

Ten native German subjects (7 female, 3 male) aged 18 to 27 (mean 21.7) participated in the WOZ experiments. None of them had user experience or engineering knowledge related to state-of-the-art spoken dialog systems. The NIMITEK corpus contains 15 hours of speech and video recordings collected during the Wizard-of-Oz experiments specially designed to provoke user's emotional reactions. More technical details about affective data collection strategy can be found in [Gnjatović and Rösner, 2008c, Gnjatović, 2009, Gnja-

tović and Rösner, 2010]. The used NIMITEK dataset contains approximately 3 hours recordings which are related to the Towers-of-Hanoi game.

Gnjatović et al. [Gnjatović, 2009] analyzed all 6798 commands presented in the NIMITEK dataset. They found that users do not follow a predefined grammar during interaction with the system. Still, by using the grammar-based language model presented in listing 3.1 on page 60 we developed the system which can recognize and process users' commands of different syntactic forms: *elliptical commands, verbose commands (i.e., the commands that were only partially recognized by the speech-recognition module), and context-dependent commands.*

6.4 Architecture I: Conventional spoken dialog system

In this section we present the possible architecture of a spoken dialog system, later referred to as the *conventional spoken dialog system* (CSDS). In Figure 6.5 one can see the interaction of the submodules of the CSDS.

The interaction within CSDS submodules can be presented as follows. The possible textual meaning of the user's utterances is delivered to the natural language understanding module. This module detects the command and for-

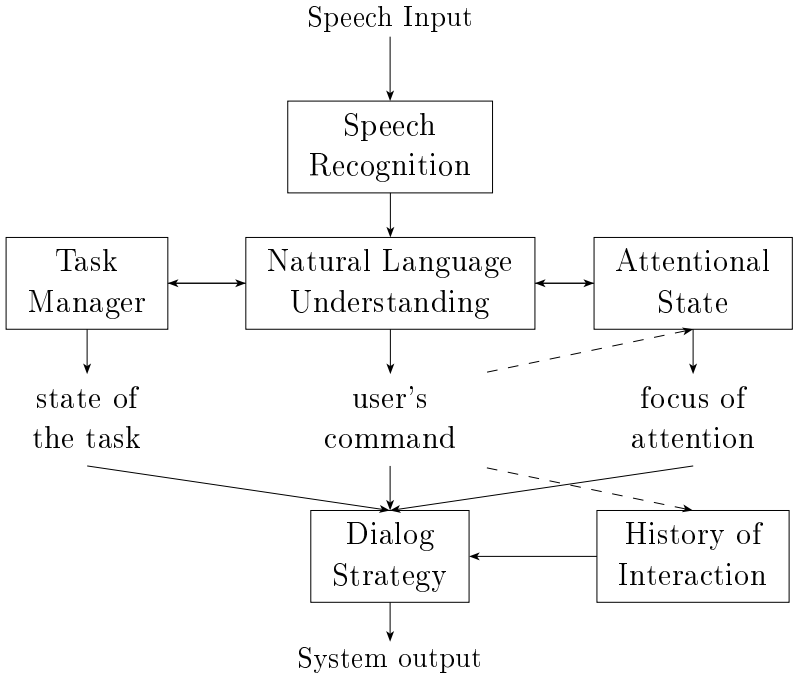


Figure 6.5: *Schema of the conventional spoken dialog system (CSDS)*

wards it:

- *to the attentional state module for updating the focus of attention,*
- *to the history of the interaction module to save the current values of other interaction features and process the context-dependent user's commands,*
- *to the task manager module (including the graphical platform) for executing the detected command, update of the state of the task, and appropriate graphical display,*

A new entry is added to the history of the interaction, containing: updated state of the task, the detected command, and the current focus of attention.

For real-time automatic speech recognition (ASR) within the conventional spoken dialog system, we used the ATK and HTK [Young et al., 2009]. Monophones ASR models are designed by training three emitting state hidden Markov models (HMM) with 16 Gaussian mixture components for each phoneme model. We use a short version of German SAMPA which includes the 39 phonemes presented in section 3.3.2. ASR models have been trained on the emotionally neutral speech samples from the Kiel dataset.

6.5 Architecture II: User-behavior-adaptive spoken dialog system

During the WOZ experiments we have seen that users employ several output modalities (mimics, speech, prosody) to communicate with a computer. In the NIMITEK demonstrator prototype [Wendemuth et al., 2008], we include recognition of the user's emotional state. The emotion classifier integrated in the NIMITEK demonstrator prototype uses three modalities: emotional prosody within spoken communication, literal meaning of user's utterances and user mimics. For the current usability test we evaluate the NIMITEK demonstrator prototype with speech-based emotion classification [Vlasenko et al., 2010]. We provide two different dialog strategies for two concerned user's emotional states (neutral and negative).

In Figure 6.6 one can see a spoken dialog system which is adaptive to the user's behavior, later referred to as *user-behavior-adaptive spoken dialog system* (UASDS). Figure 6.7 presents an interaction of submodules of the UASDS.

Phonetic transcriptions and the hypothesis word sequence generated by the speech-recognition module is transferred to the natural language understanding (NLU) and emotion-recognition module. Later, based on phonetic transcriptions and the speech signal, the emotion classifier recognizes the cur-

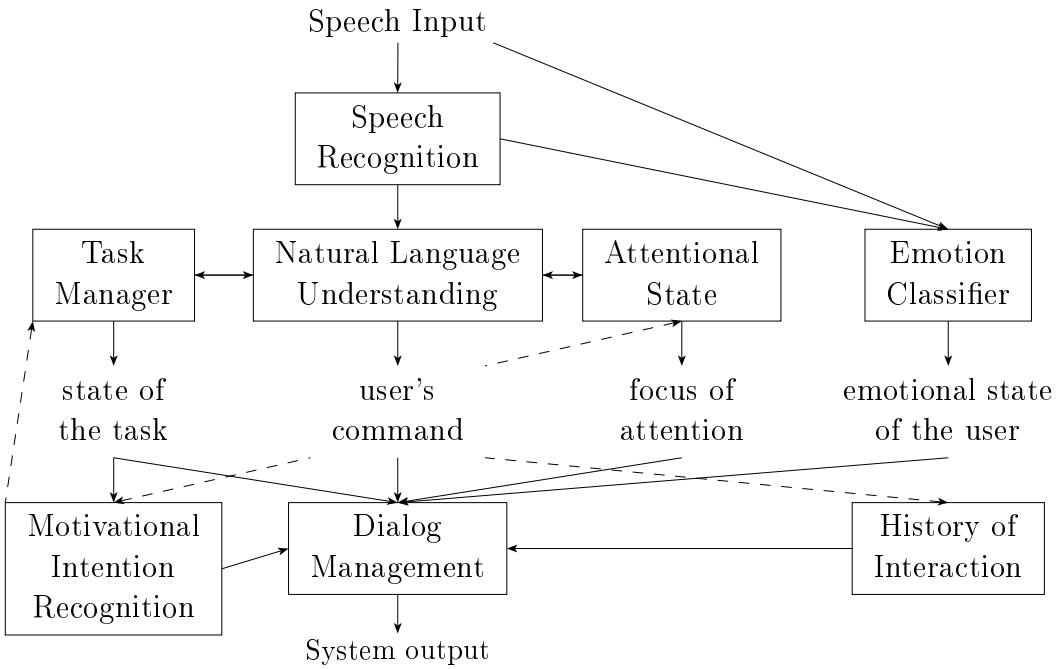


Figure 6.6: *Schema of the user-behavior-adaptive spoken dialog system (UASDS)*

rent speaker's emotional state. The NLU module interprets the command and forwards it:

- *to the attentional state module for updating the focus of attention,*
- *to the history of the interaction module to save the current values of other interaction features and process the context-dependent user's commands,*
- *to the motivational intention recognition module for defining the user's motivational intention based on his last command and current state of the task,*
- *from motivational intention recognition to the task manager module (including the graphical platform) for executing the detected command, update of the state of the task, and appropriate graphical display,*

Then, a new entry is added to the history of the interaction, containing: the updated state of the task, detected command, current focus of attention, and the detected user's emotional state. For delimitation of type of frustration (communication incomprehension or task related) we take into account the current state of the focus and history of interaction. When the user's game manipulations are far away from solving the Towers-of-Hanoi task the system indicates a task related frustration. Then, the system provides user support according to the current state of the task, and the emotional and motivational

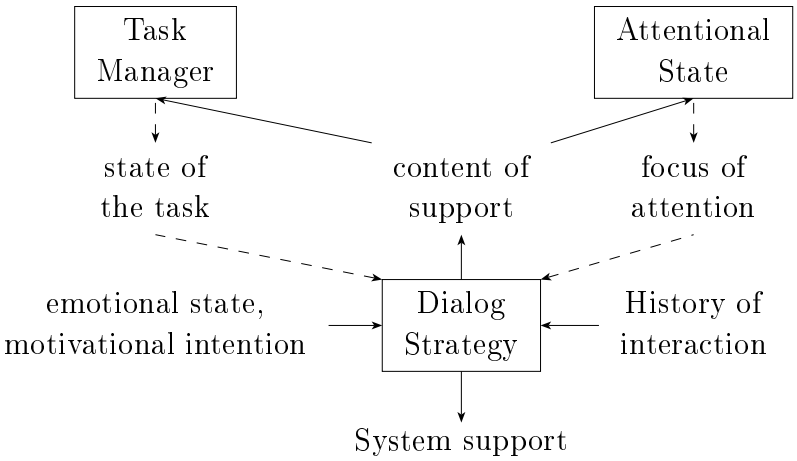


Figure 6.7: *System support processing within UASDS*

intentional state of the user. The processing of a user's command in the NIMITEK prototype UASDS is presented in Figure 6.7.

The adaptive dialog management designed to support the user addresses the negative user state on two tracks: (i) to help a frustrated user to overcome problems that occur within the interaction, and (ii) to motivate a discouraged or apathetic user. The recognized user's motivational intention determines the direction of system support: for a *cooperative* user, the next logical step is explained; for an *explorative* user, comprehensive coverage of possible steps is given; for a *destructive* user, the limitations of the next steps are explained. Generally, the support information may contain a proposed move, an audio system support and various animations. In the case when system support contains only the audio system message or the animation, this information is delivered to the task manager module for appropriate display. If support contains also a proposed move, this information is sent:

- *to the task manager module for a performance of the proposed command and an update of the state of the task,*
- *to the attentional state module for an update of the focus of attention.*

More technical details of the dialog management model can be found in [Gnjatović and Rösner, 2008a, Gnjatović, 2009, Gnjatović and Rösner, 2008c] and other publications of Gnjatović.

Like in CSDS, for real-time automatic speech recognition (ASR) within the user adaptive spoken dialog systems, we used the ATK and HTK [Young et al., 2009]. Monophones ASR models are designed by training three emitting state hidden Markov models (HMM) with 16 Gaussian mixture components for each phoneme model. We use a short version of German SAMPA which includes the 39 phonemes presented in section 3.3.2. The HMM/GMM models

have been trained on the Kiel database material and, in addition to CSDS, adapted with MLLR(RCT) on affective speech samples from the EMO-DB database. The emotion classifier integrated into UASDS based on emotional phoneme classes method, the full list of 36 phonemes (all phonemes which presented in EMO-DB dataset) is modeled for neutral and negative speaker's states.

6.6 Experiment

For our experiments we established two different SDS systems: conventional (CSDS) and user-behavior-adaptive (UASDS) with emotion adaptive dialog strategy and affective-speech-adapted ASR models. Other systems' technical characteristics are identical: vocabulary, language model, and a garbage model for OOV words.

For the usability test we hired 8 students (4 female and 4 male). Half of the test persons played the Towers-of-Hanoi game with UASDS including a behavior-based dialog management strategy and the remaining testers used the CSDS system with standard support, i.e. repeating the rules of the game or asking for the command to be repeated. The UASDS varies the answers depending on the behavior of the user like asking for a specific peg or disk, repeating the rules, or giving general hints.

All together, we collected audio material which in total lasts 16:21 minutes for the UASDS system and 27:40 minutes for the CSDS system. These recordings also include the time the system support recommendations or provides help to the user and the silences caused by the user. This data is not related to the NIMITEK corpus discussed earlier and described in detail in [Gnjatović and Rösner, 2008c].

The main point of interest are interaction time and required number of the dialog turns to solve the task. Also interesting values which were collected are measures related to user adaptation (number of the dialog turns required for adaptation and their total duration) to the systems "command list" as a response of ASR's textual output. When the user starts using commands from the system vocabulary at the beginning of the HMI, we set duration of the adaptation time to 00:00. As a start point we did not provide any information to subjects about ASR active vocabulary and grammar structure, other than the rules of the game. In the case of support requirements, users are able to ask the SDS system for "help".

6.7 Results

The experimental results of the spoken dialog systems evaluation are presented in Table 6.1. Comparing the numbers of dialog turns which are necessary to solve the puzzle, the UASDS performs better [Vlasenko et al., 2010]. On average, using the CSDS the user needs ca. 18 dialog turns more (47.4% more) to finish the game.

Trial	UASDS				CSDS			
	Complete task		Adaptation		Complete task		Adaptation	
	Turns	Time	Turns	Time	Turns	Time	Turns	Time
1.	34	05:43	1	00:00	44	05:40	1	00:00
2.	31	03:37	10	01:36	61	06:05	30	03:43
3.	34	02:44	10	01:04	81	11:48	10	01:51
4.	55	04:17	1	00:00	41	04:07	7	00:52
Mean	38.5	04:05	5.5	00:40	56.75	06:55	12	01:37

Table 6.1: *Number of turns [#], interaction time [mm:ss] for the complete task, and number of turns [#] with time intervals [mm:ss] required for user vocabulary adaptation for CSDS and UASDS*

Considering the overall time which includes pauses and the system support, the UASDS shows the better average results (04:05 vs. 06:55 minutes (40.9% less) absolute talk time). In the case of CSDS, independently of the user's behavior a standard output is given. This provides evidence that behavior dependent dialog strategies may provide better user support. Also, within interaction with the UASDS, users are more considerate to the ASR output. As a result they are adapting their commands vocabulary faster (00:40 vs. 1:37 minutes (58.7% less)).

Finally, we analyzed the dialog turns structure and commands vocabulary. The adaptation values given in Table 6.1 were counted until the first word, which is in the system's vocabulary, occurred. A total adaptation of the user could not be observed, but we would not expect this. In most cases, system specific and additional words are combined, e.g., "the smallest disk up" where "up" is not part of the (hidden) command set. Moreover, almost all users varied in words, but the longer the experiment lasted, the vocabulary used became more stable. Due to the behavior-based dialog management the user could get the right commands faster, because the strategy is directed to provide adequate information at any time.

In both versions, the user switches between two command forms: complete statements (e.g., "the smallest disk from one to the right peg") and context-dependent commands (e.g., "smallest disk" - pause - "to three"). In the recording we found a significant relation between the system version, dialog

management type, and the command form. In the UASDS almost all users uttered complete statements whereas in the CSDS the most common form is context-dependent separate commands. Moreover, due to the neutral behavior of the system the testers were not stimulated to change their strategy, because they mentioned that they thought they were interacting with an artificial system. In the other case, the users said that they think the system interacts more intuitively.

6.8 Conclusions and transition to Companion technology

Within the usability experiment we found out that during human-machine communication frustration situations, the UASDS provides comprehensive help and exhaustive recommendations in context of the current state of the task. The user-behavior-adaptive spoken dialog system built upon acoustic emotion recognition in combination with affective-speech-adapted ASR models decreases interaction time by 40.9%. During usability tests we found out that the affective-speech-adapted ASR models provide better spontaneous speech-recognition performance in real applications. At the same time user-behavior-based dialog management stimulates the user for a more cooperative interaction with the computer. As a result the user's commands vocabulary adaptation time is decreased by 58.7%. Methods developed and investigated in the NIMITEK project will lay the foundations for a technology which helps to provide a close to natural way of human-machine interaction.

In Figure 6.8 one can see the main research goals within the ongoing research project, the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems", started at 01.01.2009 (http://www.informatik.uni-ulm.de/ki/sfb-trr-62/).

The SFB/TRR 62 is an interdisciplinary (Computer Science, Electrical and Information Engineering, Psychology, and Neurosciences) research activity to investigate and optimize the interaction between human users and technical systems. It is particularly specialized on the consideration of so-called *Companion-features* - properties like adaptivity, accessibility, individuality, cooperativity, trustworthiness, and the ability to react to the user's emotional state appropriately and individually. The research program comprises of the fundamental and experimental investigation as well as the practical implementation of advanced cognitive processes in order to achieve Companion - like behavior of technical systems with an integrated human-centered multimodal (speech, mimics, gestures, biological signals) interaction interface. Within this interdisciplinary research activity we will integrate our methods (*user-*

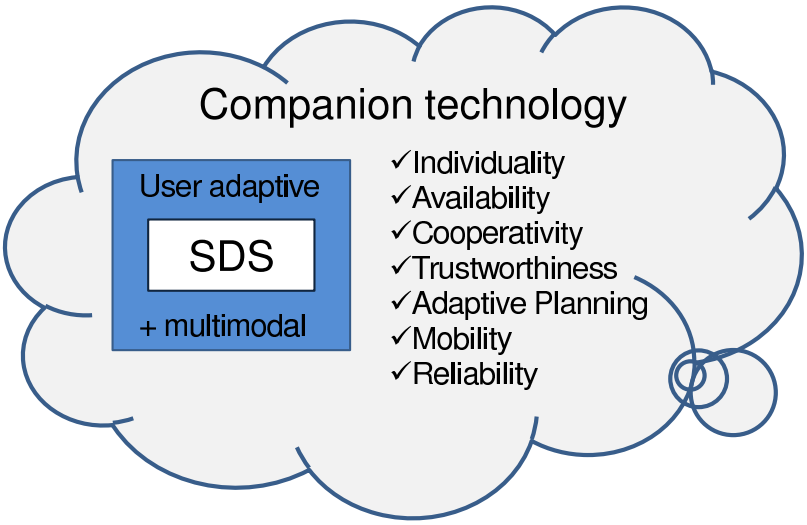


Figure 6.8: *Main research activities in the Transregional Collaborative Research Centre SFB/TRR 62 Companion-Technology for Cognitive Technical Systems*

behavior-adaptive dialog management, multimodal user's emotion processing) initiated within the NIMITEK project into a new Companion technology system. With that, it will lay the foundations for a technology which opens a completely new dimension of human-machine interaction.

Chapter 7

Conclusion and future work

Emotional speech analysis is a powerful instrument applied for development of a user-centered spoken dialog system. The fundamentals of the user-centered human-machine interaction, characteristics of the natural human speech, namely boundary and emotional prosody and emotion theory have been reviewed in Chapter 2. The main contributions of this work have been described in Chapter 3 and Chapter 4. The first contribution, described in Chapter 3, is to use an adaptation technique to increase the affective-speech-recognition rate. A concept of the adaptation on emotional speech samples of the ASR models trained on the emotionally neutral speech with MLLR(RCT)+MAP methods is proposed. This contribution will be summarized in section 7.1. The second contribution, described in Chapter 4, provides a detailed description of our various emotion-classification techniques. The summarized description of our developed emotion-classification techniques is presented in section 7.2. Phoneme-level user's emotion recognition has been integrated into a prototype dialog system especially developed by the author and colleagues to demonstrate adaptation of the system to the user's emotional state. Practical application of the previously described contribution is summarized in section 7.3. Finally, possible future research directions are discussed in section 7.4.

7.1 ASR model adaptation on affective speech data

Since we want to develop a spoken dialog system which will be able to process flexible, unrestricted user's language, spontaneous and emotionally colored speech, the acoustic model that is trained on emotionally neutral speech data is tailored to the vocal variability of the affective speech. In Chapter 3, we investigate the potency of adapting emotional speech acoustic models for German language. By the comparison of the vowel triangles for affective and neutral speech, we showed the vowel's pronunciation pattern similarity of non emotional read speech and affective speech samples. Within evaluations presented in Chapter 5, we proved that due to the pronunciation pattern similarity of affective and neutral speech, emotion-specific characteristics can

be captured from existing emotional speech corpora within adaptive transformation of model parameters of the initial neutral speech model to obtain an emotional speech acoustic model. The application of the maximum a posteriori (MAP) adaptation for the maximum likelihood linear regression (MLLR) transformed models gives a tremendous boost in emotional speech-recognition performance. The accuracy of affective speech recognition with the combined MLLR(RCT)+MAP adapted HMM/GMM models was about 8.9% absolute better than that of the ASR models trained on emotionally neutral speech samples (*baseline accuracy* 87.37%). This resulted in remarkable performance gain.

By using emotional speech adapted ASR methods we can provide better spontaneous-speech-recognition performance. This assumption has been confirmed by the usability experiment. Detailed results of this experiment can be found in section 6.7.

7.2 Recognition of the user's emotional state

To be able to design a user-centered spoken dialog system, we set up in Chapter 4 a speech-based emotion-recognition framework that should be robust enough to detect emotional events within human-machine interaction. A variety of emotion descriptors is discussed first. Two different types of emotional speech analyses are applied for speech-based emotion recognition: *frame-level* (dynamic analysis) and *turn-level* (static analysis) are presented. First of all we described the set of acoustic features which can be applied for different emotion-classification techniques. Two different optimization techniques applied on feature extraction level, namely *normalization and standardization* and *feature set optimization* have been presented afterwards. Then we introduced *utterance-*, *chunk-*, *phoneme-level* dynamic analysis models for the recognition of emotions within speech. Within experimental evaluations of the utterance-level dynamic analysis we determined the single-state HMM/GMM as an optimal architecture. In this framework we try to answer the question if phonetic content variance influences emotion-recognition performance negatively, and if models trained specifically on the phonetic unit at hand can help. During evaluation experiments we found out that the introduced unit-specific emotion-recognition models clearly outperformed common context-independent general models provided sufficient amount of training material per unit. Appearance of word-level labeled emotional corpora can improve current performance of phoneme and word-level emotion-recognition models.

In section 5.3.2 we provide results of the benchmark comparison under equal conditions on nine standard emotional speech corpora presented in Ta-

ble 2.3 in the field using the two pre-dominant paradigms: dynamic analysis on a frame-level by means of hidden Markov models and static analysis (supra-segmental) by systematic feature brute-forcing. To provide better comparability among sets, we additionally cluster each of the database's emotions into binary valence and arousal discrimination tasks (*positive, negative*), see section 2.7. When comparing the dynamic analysis with static analysis an interesting conclusion can be drawn: dynamic analysis seems to be slightly superior for spontaneous speech corpora containing variable textual content (AVIC, SAL, SmartKom, VAM), i.e. the subjects were not restricted to a predefined script, while static analysis outperforms frame-level modeling on corpora where the textual content is fixed (ABC, DES, EMO-DB, eNTERFACE, SUSAS), i.e. where there is an overlap in verbal content between test and training set. This can be explained by the nature of supra-segmental modeling: in corpora with non-scripted content, turn lengths may strongly vary. While frame-level modeling is mostly independent of highly varying turn length, in supra-segmental modeling each turn gets mapped onto one feature vector, which might not always be appropriate.

To show the robustness of our emotion-classification techniques, we presented in section 5.3.4 results of the INTERSPEECH 2009 Emotion Challenge [Schuller et al., 2009c]. With our emotion-classification technique based on dynamic analysis we proved that only by using spectral features (Mel-frequency Cepstral coefficients (MFCC)) and utterance-level analysis we can reach one of the best emotion-recognition performances for spontaneous emotional speech. We got second place for the two emotion classes task and forth place for the five emotion classes task over 33 research groups registered to get access to the data.

Finally, in section 5.3.5 we present evaluation results for cross-corpus evaluation for *intra-* and *inter*-corpus speech-based emotion recognition. We showed that the recognition rates highly depend on the specific sub-group of emotions considered. Emotion-recognition performance decreases dramatically when operating cross-corpora-wise. As long as conditions remain similar, cross-corpus training and testing seems to work to a certain degree: the DES, EMO-DB, and eNTERFACE datasets led to partly useful results. However, the cross-corpus testing on a spontaneous emotions dataset (SmartKom) clearly showed limitations of the current context-independent emotion-recognition systems. As a result, in section 4.4.3.3 we proposed to use a new context-dependent emotion-classification technique which is based on vowel-level formants tracking. By evaluating this technique we showed that the average F1 values extracted on a vowel-level are strongly correlated with the level of arousal of the speaker's emotional state. We defined the list of the most indicative German vowels within the task of measuring the level

of arousal of the speaker's emotional state. Also, we estimated the optimal Neyman-Pearson's criteria thresholds for acted and spontaneous emotions. It has been shown that spontaneous emotions required higher η values in comparison with optimal η values for acted emotions. We showed that the list of the most indicative German vowels within the task of measuring the level of arousal of the speaker's emotional state can be used for spontaneous emotion classification.

To summarize the overall results, the best emotion-recognition performance is achieved on the databases containing acted, prototypical emotions, where only emotions with high inter-labeler agreement were selected (EMO-DB, eNTERFACE, DES). The remaining emotional corpora are more challenging since they contain non-acted or induced emotions. On the lower end of recognition performance the SAL, SmartKom, and VAM corpora can be found, which contain the most spontaneous and naturalistic emotions, which in turn are also the most challenging to label. In this thesis we presented a variety of task-oriented suitable emotion-recognition methods. For example the context-independent utterance-level emotion-recognition method can be easily implemented for HMI systems which do not require textual interpretation of the user's speech. In contrast to the "brute-force" emotion-classification techniques we develop methodologically simple methods, which are universally usable for professional applications.

7.3 Application of the previously described contributions

A prototypical spoken dialog system with a user-behavior-adaptive spoken dialog system was created within the NIMITEK[2] collaboration. This system includes phoneme-level emotion recognition and ASR models adapted with MLLR(RCT) technique on emotional speech data. To prove an appropriateness of application of the previously described contributions we organized interactive usability experiments for our prototype spoken dialog system. Within the usability experiment we could show that during human-machine communication frustration situations, the user-behavior-adaptive spoken dialog system (UASDS) provides comprehensive help and exhaustive recommendations in context of the current state of the task. The UASDS built upon acoustic emotion recognition in combination with affective-speech-adapted ASR models decreases interaction time by 40.9%. During usability tests

[2]Neurobiologically Inspired, Multimodal Intention Recognition for Technical Communication Systems, 2005-2010, [Wendemuth et al., 2008]

we found out that the affective-speech-adapted ASR models provide better spontaneous-speech-recognition performance in real applications. At the same time user-behavior-based dialog management stimulates the user for a more cooperative interaction with the computer. As a result the user's commands vocabulary adaptation time is reduced by 58.7%.

7.4 Future work

The research on affective-speech-adapted ASR models and emotion recognition from speech may be further carried out in a number of directions:

- **Collection of emotional speech material with reliable textual and emotional annotation:**
  Creation of new well-annotated emotional corpora can help us to make a more detailed emotional speech analysis. Within the annotation process we should take into account two main issues: Firstly, transcription needs to acknowledge the full range of features involved in the acoustic expression of emotion, including voice quality, boundary prosody and non-linguistic features such as *laughter, crying, clatter, breath, etc.*. Secondly, it needs to describe the attributes (e.g., linguistic, dialog acts specification) that are relevant to emotion. Within the Transregional Collaborative Research Center SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG) we are collecting a new speech corpus with spontaneous emotions. Well transcribed data with reliable emotion annotation will be an important dataset for detailed context-dependent spontaneous-emotion-recognition experiments.
- **Improvement of ASR performance by creation of more reliable lexica:**
  In this work, the gain of emotion speech adapted ASR and context-dependent emotion recognition is limited due to the various errors included in existing German lexicons. To improve recognition performances, the lexica should be modified. All wrong phonetic transcriptions should be corrected; in a case of various phonetic transcriptions which are representative for the same word, all transcriptions should be included in the new lexicon.
- **More detailed fundamental context-dependent analysis of emotion indicative acoustic features:**
  Within our research we proved that the vowel can be used as the smallest emotional unit of analysis. We find out that using vowel-level analysis, namely formants tracking, can be an important issue during developing

a robust emotion classifier. We are pretty sure, that there exists some other qualitative and temporal characteristics of the smallest phonetic units which can be used for robust context-dependent emotion recognition. Future research may be carried out to specify this qualitative and temporal measures.

- **A Companion-Technology for Cognitive Technical Systems:**
  Within this interdisciplinary research activity of the Transregional Collaborative Research Centre SFB/TRR 62 we will integrate our methods (*user-behavior-adaptive dialog management, multimodal user's emotion processing*) initiated within the NIMITEK project into a new Companion technology system.
- **Dialog-state-dependent emotion recognition:**
  Combination of the speech-based emotion classification and dialog act features analysis could improve performance of miscommunication detection during HMI. For example, finding repetitions of the same dialog might contribute in addition to the acoustic-based emotion classification to the detection of trouble in communication.
- **Multimodal emotion recognition:**
  In future we want to combine audio, video emotion analysis with processing of some physiological responses (blood pressure, blood volume pulse, respiration rate, heart rate, galvanic skin response, ECG, EMG, etc.). In such a way we want to develop our own multimodal emotion-classification technique within ongoing Transregional Collaborative Research Center SFB/TRR 62. For fusion of these various processing streams we should take into account corresponding emotion indicative responses delays. For example, some physiological responses could indicate an emotional user's state slightly later than mimic expression.
- **Selection of suitable emotion categorization technique:**
  In future we would like to work on essential problems for the analysis of spontaneous emotional speech. For instance, we want to determine what an emotional episode is, where it starts and where it ends (*emotional events localization*) and which emotional annotation approach (multi-dimensional representation or classical emotion categories) to choose for annotation purposes.

Declaration

No part of this thesis has been submitted elsewhere for any other degree or qualification and it is my own work unless referenced to the contrary in the text. Some of the materials have been presented at international conferences and workshops [Schuller et al., 2007, Vlasenko et al., 2007a, Vlasenko et al., 2007b, Vlasenko and Wendemuth, 2007, Schuller et al., 2008, Vlasenko et al., 2008a, Vlasenko et al., 2008b, Vlasenko and Wendemuth, 2009b, Schuller et al., 2009, Vlasenko and Wendemuth, 2009a, Hübner et al., 2010, Vlasenko et al., 2010, Vlasenko et al., 2011b, Vlasenko et al., 2011a, Siegert et al., 2011], published as article in an academic journal [Schuller et al., 2010].

Author's Publications

[Hübner et al., 2010] Hübner, D., Vlasenko, B., Grosser, T., and Wendemuth, A. (2010). Determining Optimal Features for Emotion Recognition from Speech by applying an Evolutionary Algorithm. In *Proceedings of the Interspeech 2010*, pages 2358–2361, Makuhari, Japan. International Speech Communication Association.

[Schuller et al., 2007] Schuller, B., Vlasenko, B., Minguez, R., Rigoll, G., and Wendemuth, A. (2007). Comparing One and Two-Stage Acoustic Modeling in the Recognition of Emotion in Speech. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2007*, pages 596–600, Kyoto, Japan.

[Schuller et al., 2008] Schuller, B., Vlasenko, B., Arsic, D., Rigoll, G., and Wendemuth, A. (2008). Combining Speech Recognition and Acoustic Word Emotion Models for Robust Text-Independent Emotion Recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2008*, pages 1333–1336, Hannover, Germany.

[Schuller et al., 2009] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. (2009). Acoustic Emotion Recognition: A Benchmark Comparison of Performances. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2009*, pages 552–557, Merano, Italy.

[Schuller et al., 2010] Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. (2010). Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies. *IEEE Transactions on Affective Computing*, I:119–131.

[Siegert et al., 2011] Siegert, I., Böck, R., Philippou-Hübner, D., Vlasenko, B., and Wendemuth, A. (2011). Appropriate Emotional Labeling of Non-acted Speech Using Basic Emotions, Geneva Emotion Wheel and Self Assessment Manikins. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2011*, Barcelona, Spain.

[Vlasenko et al., 2007a] Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007a). Combining Frame and Turn-Level Information for Robust Recognition of Emotions within Speech. In *Proceedings of the Interspeech 2007*, pages 2249–2252, Antwerp, Belgium. International Speech Communication Association.

[Vlasenko et al., 2007b] Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2007b). Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. In *Proceedings of the Affective Computing and Intelligent Interaction, ACII 2007*, volume LNCS 4738, pages 139–147, Lisbon, Portugal. Springer Berlin, Heidelberg.

[Vlasenko et al., 2008a] Vlasenko, B., Schuller, B., Tadesse Mengistu, K., Rigoll, G., and Wendemuth, A. (2008a). Balancing Spoken Content Adaptation and Unit Length in the Recognition of Emotion and Interest. In *Proceedings of the Interspeech 2008*, pages 805–808, Brisbane, Australia. International Speech Communication Association.

[Vlasenko et al., 2008b] Vlasenko, B., Schuller, B., Wendemuth, A., and Rigoll, G. (2008b). On the Influence of Phonetic Content Variation for Acoustic Emotion Recognition. In *Proceedings of the 4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems, PIT 2008*, volume LNCS 5078, pages 217–220. Springer Berlin, Heidelberg, Kloster Irsee, Germany.

[Vlasenko et al., 2010] Vlasenko, B., Böck, R., and Wendemuth, A. (2010). Modeling affected user behavior during human-machine interaction. In *Proceedings of the 5th International Conference Speech Prosody 2010*, Chicago, Illinois, USA. http://speechprosody2010.illinois.edu/papers/100044.pdf.

[Vlasenko et al., 2011a] Vlasenko, B., Philippou-Hübner, D., Prylipko, D., Böck, R., Siegert, I., and Wendemuth, A. (2011a). Vowels Formants Analysis Allows Straightforward Detection of High Arousal Emotions. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2011, Top 15% paper*, Barcelona, Spain.

[Vlasenko et al., 2011b] Vlasenko, B., Prylipko, D., Philippou-Hübner, D., and Wendemuth, A. (2011b). Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In *Proceedings of the Interspeech 2011*, pages 1577–1580, Florence, Italy. International Speech Communication Association.

[Vlasenko and Wendemuth, 2007] Vlasenko, B. and Wendemuth, A. (2007). Tuning Hidden Markov Model for Speech Emotion Recognition. In *Proceedings of the "Tagungsband Fortschritte der Akustik" - DAGA 2007*, Stuttgart, Germany.

[Vlasenko and Wendemuth, 2009a] Vlasenko, B. and Wendemuth, A. (2009a). Heading Toward to the Natural Way of Human-Machine Interaction: The NIMITEK Project. In *Proceedings of the IEEE International*

Conference on Multimedia and Expo, ICME 2009, pages 950–953, New York, USA.

[Vlasenko and Wendemuth, 2009b] Vlasenko, B. and Wendemuth, A. (2009b). Processing Affected Speech Within Human Machine Interaction. In *Proceedings of the Interspeech 2009*, pages 2039–2042, Brighton, England. International Speech Communication Association.

References

[SAM, 1996] (1996). Bavarian Archive for Speech Signals. Extended SAM-PA(PhonDat-Verbmobil).

[KIE, 2002] (2002). The Kiel Corpus of Read Speech, Vol. I.

[GEW, 2008] (2008). Adapted Geneva emotion wheel used as digital questionnaire.

[Anscombe, 2000] Anscombe, A. (2000). *Intention*. Harvard University Press, Cambridge.

[Arnold, 1960] Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.

[Averill, 1980] Averill, J. R. (1980). A constructivist view of emotion. In R. Plutchik and H. Kellerman (Ed.), Emotion: Theory, research and experience. *New York: Academic Press*, pages 305–339.

[Baker, 1975] Baker, J. (1975). The Dragon System: An Overview. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23:24–29.

[Banse and Scherer, 1996] Banse, R. and Scherer, K. R. (1996). Acoustic Profiles in Vocal Emotion Expression. *Journal of Personality and Social Psychology*, 70(3):614–636.

[Barkhuysen et al., 2005] Barkhuysen, P., Krahmer, E., and Swerts, M. (2005). Problem detection in human-machine interactions based on facial expressions of users. *Speech Communication*, 45(3):343–359.

[Barra et al., 2006] Barra, R., Montero, J. M., Macias-Guarasa, J., D'Haro, L. F., San-Segundo, R., and Cordoba, R. (2006). Prosodic and Segmental Rubrics in Emotion Identification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2006*, volume I, pages 1085–1088. IEEE Computer Society.

[Barra-Chicote et al., 2009] Barra-Chicote, R., Fernández-Martínez, F. F., Lutfi, S. L., Lucas-Cuesta, J. M., Guarasa, J. M., Montero, J. M., Segundo, R. S., and Pardo, J. M. (2009). Acoustic emotion recognition using dynamic Bayesian networks and multi-space distributions. In *Proceedings of the Interspeech 2009*, pages 336–339. International Speech Communication Association.

[Batliner et al., 2000a] Batliner, A., Buckow, J., Niemann, H., Nöth, E., and Warnke, V. (2000a). *The Prosody Module*, pages 106–121. New York, Berlin.

[Batliner et al., 2000b] Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., and Fischer, K. (2000b). *Verbmobil: Foundations of Speech-to-Speech Translations The Recognition of Emotion*, pages 122–130. New York, Berlin.

[Batliner et al., 2003] Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2003). How to Find Trouble in Communication. *Speech Communication*, 40:117–143.

[Batliner et al., 2006] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2006). Combining Efforts for Improving Automatic Classification of Emotional User States. In Erjavec, T. and Gros, J. Z., editors, *Proceedings 5th Slovenian and 1st International Language Technologies Conference (IS LTC 2006), Ljubljana, Slovenia*, pages 240–245.

[Batliner et al., 2008] Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In Devillers, L., Martin, J.-C., Cowie, R., Douglas-Cowie, E., and Batliner, A., editors, *Proceedings of a Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect*, pages 28–31, Marrakesh.

[Batliner et al., 2010] Batliner, A., Seppi, D., Steidl, S., and Schuller, B. (2010). Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human-Computer Interaction*.

[Batliner et al., 2011] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., and Amir, N. (2011). Whodunnit - Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech. *Computer Speech and Language*, pages 4–28.

[Batliner and Nöth, 2003] Batliner, A. and Nöth, E. (2003). Prosody and Automatic Speech Recognition - Why not yet a Success Story and where to go from here. In *Proceedings of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, pages 357–364, Tokyo.

[Baum et al., 1970] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

[Bellman, 1957] Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

[Benzeghiba et al., 2007] Benzeghiba, M., Demori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., and Ris, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786.

[Bernstein et al., 1997] Bernstein, D. A., Clarke-Stewart, A., Roy, E. J., and Wickens, C. D. (1997). *Psychology: Fourth Edition*. Boston: Haughton Mifflin Company.

[Bertsekas, 2000] Bertsekas, D. P. (2000). *Dynamic Programming and Optimal Control*. Athena Scientific.

[Bianchi-Berthouze and Lisetti, 2002] Bianchi-Berthouze, N. and Lisetti, C. (2002). Modeling Multimodal Expression of User's Affective Subjective Experience. *User Modeling and User-Adapted Interaction*, 12(1):49–84.

[Bilmes, 1998] Bilmes, J. A. (1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.

[Bisani and Ney, 2008] Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451.

[Boersma and Weenink, 2008] Boersma, P. and Weenink, D. (2008). Praat: doing phonetics by computer (Version 5.0.0.07) [Computer program]. http://www.praat.org/.

[Bosch, 2003] Bosch, L. t. (2003). Emotions, speech and the ASR framework. *Speech Communication*, 40(1-2):213–225.

[Bozkurt et al., 2009] Bozkurt, E., Erzin, E., Çigdem Eroglu Erdem, and Erdem, A. T. (2009). Improving automatic emotion recognition from speech signals. In *Proceedings of the Interspeech 2009*, pages 324–327. International Speech Communication Association.

[Breazeal and Aryananda, 2002] Breazeal, C. and Aryananda, L. (2002). Recognition of Affective Communicative Intent in Robot-Directed Speech. *Autonomous Robots*, 12:83–104.

[Burkhardt et al., 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A Database of German Emotional Speech. In *Proceedings of the Interspeech 2005*, pages 1517–1520. International Speech Communication Association.

[Busso et al., 2007] Busso, C., Lee, S., and Narayanan, S. (2007). Using Neutral Speech Models for Emotional Speech Analysis. In *Proceedings of the Interspeech 2007*, pages 2225–2228, Antwerp, Belgium. International Speech Communication Association.

[Cairns and Hansen, 1994] Cairns, D. and Hansen, J. H. L. (1994). Nonlinear Analysis and Detection of Speech Under Stressed Conditions. *Journal of the Acoustical Society of America*, 96(6):3392–3400.

[Carletta, 1996] Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

[Clifton et al., 2002] Clifton, C. J., Carlson, K., and Frazier, L. (2002). Informative prosodic boundaries. *Language and Speech*, 45:87–114.

[Cohen et al., 2003] Cohen, I., Sebe, N., Gozman, F. G., Cirelo, M. C., and Huang, T. S. (2003). Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.*, volume I, pages 595–601.

[Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., N., T., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion Recognition in Human-Computer Interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.

[Cowie et al., 2000] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schröder, M. (2000). The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24, Newcastle, Northern Ireland.

[Damasio, 1994] Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Hitman Brain.* London, U.K.

[Darwin, 1872] Darwin, C. (1872). *The Expression of the Emotions in Man and Animals.* University of Chicago Press, Chicago.

[Dellaert et al., 1996] Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing Emotions in Speech. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 1996*, volume III, pages 1970–1973, Philadelphia, PA, USA.

[Devillers et al., 2005] Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

[Douglas-Cowie et al., 2003] Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 40(1-2):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., Abrilan, S., Batliner, A., Amir, N., and Karpousis, K. (2007). The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Proceedings of the Affective Computing and Intelligent Interaction, ACII 2007*, volume LNCS 4738, pages 488–500, Berlin-Heidelberg. Springer.

[Dumouchel et al., 2009] Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., and Boufaden, N. (2009). Cepstral and Long-Term Features for Emotion Recognition. In *Proceedings of the Interspeech 2009*, pages 344–347. International Speech Communication Association.

[Engbert and Hansen, 1996] Engbert, I. S. and Hansen, A. (1996). Documentation of the Danish Emotional Speech Database DES. Technical report, Center for Person Kommunikation, Aalborg University, Denmark.

[Eyben et al., 2009] Eyben, F., Wöllmer, M., and Schuller, B. (2009). openEAR - Introducing the Munich open-source Emotion and Affect Recognition toolkit. In *Proceedings of the Affective Computing and Intelligent Interaction, ACII 2009*, volume I, pages 576 – 581.

[Fairbanks and Pronovost, 1939] Fairbanks, G. and Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, 6:87–104.

[Fernandez and Picard, 2003] Fernandez, R. and Picard, R. W. (2003). Modeling drivers' speech under stress. *Speech Communication*, 40(1-2):145–159.

[Fernandez and Picard, 2005] Fernandez, R. and Picard, R. W. (2005). Classical and Novel Discriminant Features for Affect Recognition from Speech.

In *Proceedings of the Interspeech 2005*. International Speech Communication Association.

[Fiscus, 1997] Fiscus, J. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 1997*, Santa Barbara, USA.

[Fragopanagos and Taylor, 2005] Fragopanagos, N. and Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405.

[Fraser and Gilbert, 1991] Fraser, N. M. and Gilbert, G. N. (1991). Simulating Speech Systems. *Computer Speech and Language*, 5:81–99.

[Gales, 1996] Gales, M. F. (1996). The Generation And Use Of Regression Class Trees For MLLR Adaptation. Technical report, Cambridge University Engineering Department.

[Gauvain and Lee, 1994] Gauvain, J. and Lee, C. (1994). Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2:291–298.

[Gnjatović, 2009] Gnjatović, M. (2009). *Adaptive Dialogue Management in Human-Machine Interaction*. PhD thesis, Universität Magdeburg.

[Gnjatović and Rösner, 2008a] Gnjatović, M. and Rösner, D. (2008a). Adaptive Dialogue Management in the NIMITEK Prototype System. In *Proceedings of the 4th IEEE Tutorial and Research Workshop "Perception and Interactive Technologies for Speech-Based Systems" PIT'08*, pages 14–25, Kloster Irsee, Germany.

[Gnjatović and Rösner, 2008b] Gnjatović, M. and Rösner, D. (2008b). Emotion Adaptive Dialogue Management in Human-Machine Interaction. In *Proceedings of the 19th European Meetings on Cybernetics and Systems Research (EMCSR 2008)*, pages 567–572, Vienna, Austria. Austrian Society for Cybernetic Studies.

[Gnjatović and Rösner, 2008c] Gnjatović, M. and Rösner, D. (2008c). On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[Gnjatović and Rösner, 2010] Gnjatović, M. and Rösner, D. (2010). Inducing Genuine Emotions in Simulated Speech-Based Human-Machine Interaction: The NIMITEK Corpus. *IEEE Transactions on Affective Computing*, I:132–144.

[Goudbeek et al., 2009] Goudbeek, M., Goldman, J., and Scherer, K. R. (2009). Emotion dimensions and formant position. In *Proceedings of the Interspeech 2009*, pages 1575–1578, Brighton, UK. International Speech Communication Association.

[Grimm et al., 2007] Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800.

[Grimm et al., 2008] Grimm, M., Kroschel, K., and Narayanan, S. (2008). The Vera am Mittag German Audio-Visual Emotional Speech Database. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2008*, pages 865–868.

[Hansen and Bou-Ghazale, 1997] Hansen, J. and Bou-Ghazale, S. (1997). Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. In *Proceedings of the 5th European Conference on Speech Communication and Technology, Eurospeech 1997*, volume IV, pages 1743–1746, Rhodes, Greece. International Speech Communication Association.

[Harré, 1986] Harré, R. (1986). *The social construction of emotions*. Oxford: Basil Blackwell.

[Herzog et al., 2004] Herzog, G., Ndiaye, A., Merten, S., Kirchmann, H., Becker, T., and Poller, P. (2004). Large-scale Software Integration for Spoken Language and Multimodal Dialog Systems. *Natural Language Engineering*, 10(3/4):283–305.

[Hess et al., 1995] Hess, W., Kohler, K., and Tillman, H.-G. (1995). The Phondat-Verbmobil speech corpus. In *Proceedings of the 4th European Conference on Speech Communication and Technology, Eurospeech 1995*, pages 863–866, Madrid, Spain. International Speech Communication Association.

[Hohmann, 1996] Hohmann, G. W. (1996). Some effects of spinal cord lesions on experienced emotional feelings. *Psychophysiology*, III(2):143–156.

[Hönig et al., 2009] Hönig, F., Wagner, J., Batliner, A., and Nöth, E. (2009). Classification of User States with Physiological Signals: On-line Generic Features vs. Specialized. In *Proceedings of the 17th European Signal Processing Conference, EUSIPCO 2009*, pages 2357–2361, Glasgow, Scotland.

[Ijima et al., 2009] Ijima, Y., Tachibana, M., Nose, T., and Kobayashi, T. (2009). Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009*, pages 4157–4160, Taipei. IEEE Computer Society.

[Iliou and Anagnostopoulos, 2009] Iliou, T. and Anagnostopoulos, C.-N. (2009). Comparison of Different Classifiers for Emotion Recognition. *Proceedings of the Panhellenic Conference on Informatics*, pages 102–106.

[Inanoglu and Young, 2009] Inanoglu, Z. and Young, S. (2009). Data-driven emotion conversion in spoken English. *Speech Communication*, 51(3):268–283.

[Jaimes and Sebe, 2007] Jaimes, A. and Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1-2):116–134.

[Jaimes et al., 2006] Jaimes, A., Sebe, N., and Gatica-Perez, D. (2006). Human-Centered Computing: A Multimedia Perspective. In *Proceedings of the 14th annual ACM international conference on Multimedia, MULTIMEDIA 2006*, pages 855–864, New York, NY, USA. ACM.

[James, 1884] James, W. (1884). *What is an emotion?* Mind, Vol. 9(34).

[Jiang and Cai, 2004] Jiang, D.-N. and Cai, L.-H. (2004). Speech emotion classification with the combination of statistic features and temporal features. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2004*, volume III, pages 1967–1970.

[Kehrein, 2002] Kehrein, R. (2002). The Prosody of Authentic Emotions. In *Proceedings of the International Conference Speech Prosody 2002*, Aix-en-Provence, France.

[Kießling, 1996] Kießling, A. (1996). *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*. Shaker, Aachen.

[Kim et al., 2010] Kim, J., Lee, S., and Narayanan, S. S. (2010). An exploratory study of manifolds of emotional speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010*, pages 5142–5145, Dallas, TX, USA. IEEE Computer Society.

[Kockmann et al., 2009] Kockmann, M., Burget, L., and Cernocký, J. (2009). Brno University of Technology System for Interspeech 2009 Emotion Challenge. In *Proceedings of the Interspeech 2009*, pages 348–351. International Speech Communication Association.

[Lee et al., 2009a] Lee, C., Busso, C., Lee, S., and Narayanan, S. (2009a). Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Proceedings of the Interspeech 2009*, pages 1983–1986. International Speech Communication Association.

[Lee et al., 2009b] Lee, C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2009b). Emotion recognition using a hierarchical binary decision tree approach. In *Proceedings of the Interspeech 2009*, pages 320–323. International Speech Communication Association.

[Lee, 2007] Lee, C.-H. (2007). Fundamentals and Technical Challenges in Automatic Speech Recognition. In *Proceedings of the 12th International Conference Speech and Computer SPECOM 2007*, pages 25–44, Moscow, Russia.

[Lee and Narayanan, 2005] Lee, C. M. and Narayanan, S. S. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303.

[Lee et al., 2004] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., and Narayanan, S. (2004). Emotion recognition based on phoneme classes. In *Proceedings of the 8th International Conference on Spoken Language Processing, Interspeech 2004*. International Speech Communication Association.

[Li and Zhao, 1998] Li, Y. and Zhao, Y. (1998). Recognizing Emotions in Speech Using Short-term and Long-term Features. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 1998*, pages 2255–2258, Sydney, Australia.

[Litman and Forbes, 2003] Litman, D. and Forbes, K. (2003). Recognizing emotions from student speech in tutoring dialogues. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2003*, pages 25–30.

[Luengo et al., 2009] Luengo, I., Navas, E., and Hernáez, I. (2009). Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge. In *Proceedings of the Interspeech 2009*, pages 332–335. International Speech Communication Association.

[Lugger and Yang, 2007] Lugger, M. and Yang, B. (2007). An incremental analysis of different feature groups in speaker independent emotion recognition. In *Proceedings of the 16th International Congress of Phonetic Sciences, ICPhS 2007*, pages 2149–2152.

[Mangold, 1990] Mangold, M. (1990). *DUDEN: Das Aussprachewörterbuch 6*. Mannheim.

[Martin et al., 2006] Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The Enterface'05 Audio-Visual Emotion Database. In *Proceedings of the IEEE Workshop on Multimedia Database Management*, Atlanta.

[McGilloway et al., 2000] McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, C., Westerdijk, M., and Stroeve, S. (2000). Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark. In *Proceedings of the ISCA workshop on Speech and Emotion*, pages 207–212.

[Metallinou et al., 2010] Metallinou, A., Lee, S., and Narayanan, S. S. (2010). Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010*, Dallas, Texas, USA. IEEE Computer Society.

[Metze et al., 2010] Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., and Steidl, S. (2010). Emotion Recognition Using Imperfect Speech Recognition. In *Proceedings of the Interspeech 2010*. International Speech Communication Association.

[Morrison et al., 2007] Morrison, D., Wang, R., and Silva, L. C. D. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49(2):98–112.

[Murray and Arnott, 1993] Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. *Journal of the Acoustical Society of America*, 93(2):1097–1108.

[Nicholas et al., 2006] Nicholas, G., Rotaru, M., and Litman, D. (2006). Exploiting Word-level Features for Emotion Prediction. In *Proceeding of the IEEE/ACM workshop Spoken Language Technology, SLT 2006*, pages 110–113.

[Niemann et al., 1998] Niemann, H., Nöth, E., Batliner, A., Buckow, J., Gallwitz, F., Huber, R., Kießling, A., Kompe, R., and Warnke, V. (1998). Using

Prosodic Cues In Spoken Dialog Systems. In *Proceedings of the 2nd International Workshop Speech and computer SPECOM 1998*.

[Nogueiras et al., 2001] Nogueiras, A., Moreno, A., Bonafonte, A., and Mariño, J. (2001). Speech Emotion Recognition Using Hidden Markov Models. In *Proceedings of the Interspeech 2001*, pages 2679–2682, Aalborg, Denmark. International Speech Communication Association.

[Nöth et al., 2002] Nöth, E., Batliner, A., Warnke, V., Haas, J.-P., Boros, M., Buckow, J., Huber, R., Gallwitz, F., Nutt, M., and Niemann, H. (2002). On the Use of Prosody in Automatic Dialogue Understanding. *Speech Communication*, 36(1-2):45–62.

[Nöth et al., 2004] Nöth, E., Horndasch, A., Gallwitz, F., and Haas, J. (2004). Experiences with Commercial Telephone-based Dialogue Systems. *Information Technology*, 46(6):315–321.

[Nwe et al., 2003] Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Classification of stress in speech using linear and nonlinear features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003*, volume II, pages 9–12. IEEE Computer Society.

[Ortony and Turner, 1990] Ortony, A. and Turner, T. (1990). What's basic about basic emotions? *Psychological Review*, 97(3):315–331.

[Oudeyer, 2003] Oudeyer, P. (2003). The Production and Recognition of Emotions in Speech: Features and Algorithms. *International Journal of Human-Computer Studies*, 59:157–183.

[Paterson and Barney, 1952] Paterson, G. and Barney, H. (1952). Control Methods Used in a Study of the Vowels. *Journal of the Acoustical Society of America*, 24(2):175–194.

[Picard, 1997] Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA, USA.

[Plutchik, 2001] Plutchik, R. (2001). The nature of emotions. *American Scientist*, 89(4):344–350.

[Polzehl et al., 2009] Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., and Metze, F. (2009). Emotion classification in children's speech using fusion of acoustic and linguistic features. In *Proceedings of the Interspeech 2009*, pages 340–343. International Speech Communication Association.

[Polzin and Waibel, 1998] Polzin, T. S. and Waibel, A. (1998). Detecting Emotions in Speech. In *Proceedings of the 2nd International Conference on Cooperative Multimodal Communication CMC 1998*.

[Pompino-Marschall, 1992] Pompino-Marschall, B. (1992). PHONDAT. Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch. *Forschungsbericht des IPSK (FIPKM)*, 30:99–128.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Rabiner, 1989] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286.

[Reithinger et al., 2003] Reithinger, N., Alexandersson, J., Becker, T., Blocher, A., Engel, R., Löckelt, M., Müller, J., Pfleger, N., Poller, P., Streit, M., and Tschernomas, V. (2003). SmartKom : Adaptive and Flexible Multimodal Access to Multiple Applications. In *Proceedings of the 5th International Conference on Multimodal Interfaces, ICMI 2003*, pages 101–108, Vancouver, British Columbia, Canada. ACM.

[Roberts et al., 2005] Roberts, W. J. J., Ephraim, Y., and Sabrin, H. W. (2005). Speaker classification using composite hypothesis testing and list decoding. *IEEE Transactions on Speech and Audio Processing*, 13(2):211–219.

[Schachter and Signer, 1962] Schachter, S. and Signer, J. E. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69:379–399.

[Scherer, 1986] Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, 99:143–165.

[Scherer, 1999] Scherer, K. R. (1999). *Appraisal theory*, pages 637–663. Wiley, Chichester.

[Scherer, 2004] Scherer, K. R. (2004). *Feelings integrate the central representation of appraisal-driven response organization in emotion*, pages 136–157. Cambridge University Press, Cambridge.

[Scherer, 2005] Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44:695–729.

[Schiel et al., 2002] Schiel, F., Steininger, S., and Turk, U. (2002). The Smartkom multimodal corpus at BAS. In *Proceedings of the Language Resources and Evaluation, LREC 2002*.

[Schröder et al.,] Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., and Schuller, B. Towards responsive Sensitive Artificial Listeners. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, Bellagio, Italy.

[Schröder et al., 2007] Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., and Wilson, I. (2007). What should a generic emotion markup language be able to represent? In *Proceedings of the Affective Computing and Intelligent Interaction, ACII 2007*, volume LNCS 4738, pages 440–451, Lisbon, Portugal. Springer Berlin, Heidelberg.

[Schuller, 2002] Schuller, B. (2002). Towards intuitive speech interaction by the integration of emotional aspects. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, SMC 2002*, Yasmine Hammamet, Tunesia.

[Schuller et al., 2003] Schuller, B., Rigoll, G., and Lang, M. (2003). Hidden Markov Model-Based Speech Emotion Recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003*, volume II, pages 1–4. IEEE Computer Society.

[Schuller et al., 2004] Schuller, B., Rigoll, G., and Lang, M. (2004). Speech Emotion Recognition Combining Acoustic Features and Linguistic Information in a Hybrid Support Vector Machine - Belief Network Architecture. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004*, volume I, pages 577–580. IEEE Computer Society.

[Schuller et al., 2005a] Schuller, B., Lang, M., and Rigoll, G. (2005a). Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers. In Fastl, H. and Fruhmann, M., editors, *Proceedings of the "Tagungsband Fortschritte der Akustik" - DAGA 2005*, volume I, pages 329–330. DEGA, Berlin.

[Schuller et al., 2005b] Schuller, B., Müller, R., Lang, M., and Rigoll, G. (2005b). Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles. In *Proceedings of the Interspeech 2005 - Eurospeech*, pages 805–809, Lisbon, Portugal. International Speech Communication Association.

[Schuller et al., 2007a] Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2007a). The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals. In *Proceedings of the Interspeech 2007*, pages 2253–2256, Antwerp, Belgium. International Speech Communication Association.

[Schuller et al., 2007b] Schuller, B., Seppi, D., Batliner, A., Maier, A., and Steidl, S. (2007b). Towards More Reality in the Recognition of Emotional Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010*, volume II, pages 941–944, Honolulu, Hawaii, USA. IEEE Computer Society.

[Schuller et al., 2007c] Schuller, B., Wimmer, M., Arsic, D., Rigoll, G., and Radig, B. (2007c). Audiovisual Behavior Modeling by Combined Feature Spaces. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007*, volume II, pages 733–736, Honolulu, Hawaii, USA. IEEE Computer Society.

[Schuller et al., 2008] Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., and Rigoll, G. (2008). Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space? In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008*, Las Vegas, Nevada, USA. IEEE Computer Society.

[Schuller et al., 2009a] Schuller, B., Batliner, A., Seppi, D., and Steidl, S. (2009a). Emotion Recognition from Speech: Putting ASR in the Loop. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009*, pages 4585–4588, Taipei. IEEE Computer Society.

[Schuller et al., 2009b] Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., and Konosu, H. (2009b). Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 27(12):1760–1774.

[Schuller et al., 2009c] Schuller, B., Steidl, S., and Batliner, A. (2009c). The INTERSPEECH 2009 Emotion Challenge. In *Proceedings of the Interspeech 2009*, pages 312–315, Brighton, UK. International Speech Communication Association.

[Schuller et al., 2009d] Schuller, B., Wöllmer, M., Eyben, F., and Rigoll, G. (2009d). *The Role of Prosody in Affective Speech*, volume 97 of *Linguistic Insights, Studies in Language and Communication*, chapter Spectral or Voice Quality? Feature Type Relevance for the Discrimination of Emotion Pairs, pages 285–307. Peter Lang Publishing Group.

[Schuller et al., 2010] Schuller, B., Metze, F., Steidl, S., Batliner, A., Eyben, F., and Polzehl, T. (2010). Late Fusion of Individual Engines for Improved Recognition of Negative Emotion in Speech - Learning vs. Democratic Vote. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010*, pages 5230–5233. IEEE Computer Society.

[Schuller et al., 2011] Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*.

[Schuller and Rigoll, 2006] Schuller, B. and Rigoll, G. (2006). Timing Levels in Segment-Based Speech Emotion Recognition. In *Proceedings of the Interspeech 2006*, pages 1818–1821, Pittsburgh, USA. International Speech Communication Association.

[Scripture, 1921] Scripture, E. (1921). A study of emotions by speech transcription. *Vox*, 31:179–183.

[Seppi et al., 2010] Seppi, D., Batliner, A., Steidl, S., Schuller, B., and Nöth, E. (2010). Word Accent and Emotion. In *Proceedings of the International Conference Speech Prosody 2010*, Chicago, USA.

[Shahin, 2006] Shahin, I. (2006). Enhancing speaker identification performance under the shouted talking condition using the second order circular Hidden Markov Models. *Speech Communication*, 48(8):1047–1055.

[Shami and Verhelst, 2006] Shami, M. and Verhelst, W. (2006). Automatic Classification of Emotions in Speech Using Multi-Corpora Approaches. In *Proceedings of the second annual IEEE BENELUX/DSP Valley Signal Processing Symposium (SPS-DARTS 2006)*, pages 3–6, Antwerp, Belgium.

[Shriberg, 2005] Shriberg, E. (2005). Spontaneous Speech: How People Really Talk and Why Engineers Should Care. In *Proceedings of the Interspeech 2005 - Eurospeech*, pages 1781–1784. International Speech Communication Association.

[Skinner, 1935] Skinner, E. (1935). A calibrated recording and analysis of the pitch, force, and quality of vocal tones expressing happiness and sadness. *Speech Monographs*, 2:81–137.

[Slaney and McRoberts, 1998] Slaney, M. and McRoberts, G. (1998). Baby Ears: A recognition system for affective vocalizations. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1998*, volume II, pages 985–988. IEEE Computer Society.

[Steininger et al., 2002] Steininger, S., Schiel, F., Dioubina, O., and Raubold, S. (2002). Development of User-State Conventions for the Multimodal Corpus in SmartKom. In *Proceedings of the Workshop on Multimodal Resources and Multimodal Systems Evaluation*, pages 33–37, Las Palmas, Gran Canaria, Spain.

[Ververidis and Kotropoulos, 2003] Ververidis, D. and Kotropoulos, C. (2003). A Review of Emotional Speech Databases. In *Proceedings of the Panhellenic Conference on Informatics*, pages 560–574, Thessaloniki, Greece.

[Ververidis and Kotropoulos, 2004] Ververidis, D. and Kotropoulos, C. (2004). Automatic speech classification to five emotional states based on gender information. In *Proceedings of the 12th European Signal Processing Conference EUSIPCO 2004*, pages 341–344, Austria.

[Ververidis and Kotropoulos, 2006] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, 13:260–269.

[Vogt and Andre, 2005] Vogt, T. and Andre, E. (2005). Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME 2005*, pages 474–477, Amsterdam, The Netherlands.

[Vogt and André, 2009] Vogt, T. and André, E. (2009). Exploring the benefits of discretization of acoustic features for speech emotion recognition. In *Proceedings of the Interspeech 2009*, pages 328–331. International Speech Communication Association.

[Vogt et al., 2008] Vogt, T., André, E., and Wagner, J. (2008). Affect and Emotion in Human-Computer Interaction. chapter Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation, pages 75–91. Springer Berlin, Heidelberg.

[Weintraub et al., 1996] Weintraub, M., Taussig, K., Hunicke-smith, K., and Snodgrass, A. (1996). Effect of speaking style on LVCSR performance. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 1996*, pages 1457–1460, Philadelphia, PA, USA.

[Wendemuth et al., 2008] Wendemuth, A., Braun, J., Michaelis, B., Ohl, F., Rösner, D., Scheich, H., and Warnemünde, R. (2008). Neurobiologically inspired, multimodal Intention Recognition for technical communication Systems (NIMITEK). In *Proceedings of the 4th IEEE Tutorial and Research Workshop Perception and Interactive Technologies for Speech-Based Systems, PIT 2008*, volume LNCS 5078, pages 141–144. Springer Berlin, Heidelberg.

[Whissell, 1989] Whissell, C. (1989). The Dictionary of Affect in Language. In Plutchik, R. and Kellerman, H., editors, *Emotion: theory, research and experience*, volume 4, The measurement of emotions, pages 113–131. Academic Press, New York.

[Williams and Stevens, 1972] Williams, C. and Stevens, K. (1972). Emotions and speech: some acoustic correlates. *Journal of the Acoustical Society of America*, 52:1238–1250.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco.

[Wöllmer et al., 2008] Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of the Interspeech 2008*, pages 597–600, Australia, Brisbane. International Speech Communication Association.

[Wong and Mak, 2000] Wong, K.-M. and Mak, B. (2000). Map adaptation with subspace regression classes and tying. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume III, pages 1551–1554, Los Alamitos, CA, USA. IEEE Computer Society.

[Wundt, 1897] Wundt, W. (1897). *Outlines of Psychology*. Wilhelm Engelmann, Leipzig.

[Yacoub et al., 2003] Yacoub, S., Simske, S., Lin, X., and Burns, J. (2003). Recognition of Emotions in Interactive Voice Response Systems. In *Proceedings of the Interspeech 2003 - Eurospeech*, pages 729–732, Geneva, Switzerland. International Speech Communication Association.

[Yildirim et al., 2011] Yildirim, S., Narayanan, S. S., and Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Computer Speech and Language*, 25:29–44.

[Young et al., 2009] Young, S., Evermann, G., Gales, M., T., H., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2009). *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department.

[Young, 1995] Young, S. J. (1995). Large Vocabulary Continuous Speech Recognition: A Review. In *Proceedings of IEEE workshop on Automatic Speech Recognition*, pages 3–28, Snowbird, Utah, USA.

[Young, S. J. et al., 1989] Young, S. J., Russell, N. H., and Thornton, J. H. S. (1989). Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems.

[Yu et al., 2001] Yu, F., Chang, E., Xu, Y., and Shum, H.-Y. (2001). Emotion Detection from Speech to Enrich Multimedia Content. In *Proceedings of the 2nd IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing*, pages 550–557. Springer Berlin, Heidelberg.

[Yu, 2006] Yu, K. (2006). *Adaptive Training for Large Vocabulary Continuous Speech Recognition*. PhD thesis, Cambridge University, Cambridge, UK.

[Zeißler et al., 2006] Zeißler, V., Adelhardt, J., Batliner, A., Frank, C., Nöth, E., Shi, R. P., and Niemann, H. (2006). *The Prosody Module*, volume 1, pages 139–152. Berlin.

[Zeng and Martinez, 2000] Zeng, X. and Martinez, T. R. (2000). Distribution-Balanced Stratified Cross-Validation for Accuracy Estimation. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(1):1–12.

[Zeng et al., 2009] Zeng, Z., Pantic, M., Rosiman, G. I., and Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.

[Zhou et al., 1998] Zhou, G., Hansen, J. H. L., and Kaiser, J. F. (1998). Linear and Nonlinear Speech Feature Analysis for Stress Classification. In *Proceedings of the International Conference on Spoken Language Processing, ICSLP 1998*, Sydney, Australia.