Institute of Electronic Music and Acoustics (IEM)
University of Music and Performing Arts, Graz

Investigations on
A Robust Feature Set for Classification of Speech under Stress

Diploma Thesis

Johannes Luig

Supervisor: Dr. Alois Sontacchi
Assessor: Prof. Robert Höldrich

June, 2009

Abstract

The objective of this diploma thesis is the selection and evaluation of appropriate low-level features and derived feature characteristics for automated recognition and classification of speech under varying emotions and mental stress levels. Importance is attached to obtaining results which are applicable to speech under a broad spectrum of stress types and independent of the language spoken. For this purpose, speech data from an English database of speech under stress (SUSAS) is analyzed as well as a German database of emotional speech (Emo-DB) and an English corpus of non-prompted air traffic control speech (ATCOSIM).

Basic features are extracted using the speech analysis software Praat; including pitch, intensity, F1/F2 frequency and bandwidth, harmonicity, MFCCs, and properties of the glottal source spectrum. Further processing steps, implemented in MATLAB, comprise a phoneme boundary and class detection with subsequent feature extraction utilizing the phoneme grid as a new time base. These additional features include phoneme durations and a feature based on the nonlinear Teager Energy Operator (TEO).

The discriminative power of single features is estimated by means of appropriate statistical tests on the derived characteristics. This results in a feature ranking list for a selected combination of two emotional classes, from which the best performing set of features is then determined iteratively. Using this feature set, a supervised classification method (k-nearest neighbours) is employed in a cross-validation process. Its outcome is the percentage of correctly assigned emotional classes, which is taken as a measure of performance. Finally, a "shared" feature set is found by intersecting optimum feature sets of individual experiments.

For acted emotional speech, results of up to 98% correct classification rate (CCR) are achieved using individual feature sets, which are degraded by not more than 12% when taking the shared feature set for classification. Workload level classification performance reaches up to 70% CCR for individual feature sets and likewise degrades by 12% maximum when using the shared set, what ends up in rather moderate classification rates around 60% CCR though.

Zusammenfassung

Ziel dieser Diplomarbeit ist die Auswahl und Beurteilung geeigneter Sprachmerkmale (Features) und daraus abgeleiteter *Feature Characteristics* zur automatischen Erkennung und Einteilung von gesprochener Sprache in verschiedenen emotionalen Zuständen und bei unterschiedlicher psychischer Belastung. Die Ergebnisse sollen für unterschiedliche Arten von Stress anwendbar und unabhängig von der gesprochenen Sprache sein. Zu diesem Zweck werden Sprachdaten einer englischen Datenbank mit Sprache unter Stressbedingungen (SUSAS), einer deutschen mit emotionaler Sprache (Emo-DB) sowie eines englischen Sprachkorpus mit Fluglotsen-Funkverkehr analysiert.

Grundlegende Features wie Tonhöhe, Intensität, Frequenzen und Bandbreiten der ersten beiden Formanten, Harmonizität, MFCCs und Eigenschaften der glottalen Anregung werden mit Hilfe der Sprachanalyse-Software Praat extrahiert. Anschließend wird eine Phonemgrenzenerkennung und -klassifizierung durchgeführt, was Voraussetzung für die Berechnung der Phonemdauer sowie eines auf dem Teager Energy Operator (TEO) basierenden Features ist. Diese Berechnungen werden – wie auch die weiteren Schritte – in MATLAB implementiert.

Das Differenzierungspotential der einzelnen Merkmale wird mit Hilfe geeigneter statistischer Tests bestimmt, woraus sich eine Rangliste der Features für eine Auswahl zweier emotionaler Klassen ergibt. Aus dieser wird iterativ diejenige Kombination von Features ermittelt, die die besten Ergebnisse bei der Klassifikation mit einer überwachten Methode (k-nearest neighbours) liefert. In einem Vergleichsprüfungsverfahren wird so der Prozentsatz der korrekt zugeordneten emotionalen Klassen berechnet, der das Ergebnis darstellt. Ein “allgemeines” Set von Merkmalen wird schließlich durch Bildung der Schnittmenge aus den Einzelergebnissen gewonnen.

Bei der Analyse gespielter Emotionen werden unter Verwendung der jeweils besten Feature-Sets Ergebnisse von bis zu 98% korrekter Erkennungsrate (CCR) erzielt; bei Verwendung des allgemeinen Sets verschlechtert sich die CCR um maximal 12%. Die Erkennung von Arbeitsbelastung (Workload) erreicht bis zu 70% CCR, eine vergleichbare Abnahme von 12% bei Verwendung des allgemeinen Sets bedeutet hier im Endeffekt jedoch eher mäßige Erkennungsraten um etwa 60% CCR.

Acknowledgements

I would like to take this opportunity to say "thank you" to a couple of people who were more or less involved in my diploma thesis or who I met on the way there ...

To my supervisor, Alois Sontacchi – for finding the right mixture of demand and assistance, for his nearly 24/7 availability, and for his good sense of humor.

To the whole IEM staff, especially to my colleagues in the first floor – for creating such a pleasant but motivating atmosphere at the institute. Christian, thanks for your help with the HTK toolkit.

To my fellow students and friends – for all the time spent together. I hope not all our paths go separate ways!

To my parents – for always supporting me in whatever I wanted to do. And to my brothers. As well as to my "family-in-law". You rock!

To my little family, Evelyne, Simon and David – just for being there and bringing joy into my life. And for being that patient during the last weeks and months, when I sometimes could have recorded my own voice as a perfect sample of "speech under stress"... I love you.

Contents

Abstract i
Zusammenfassung ii
Acknowledgements iii

1. Preface **1**
1.1. Motivation 1
1.2. Objective 2
1.3. Restrictions 2
1.4. Explanation of Terms 3
1.5. Outline 4

2. Analysis of Speech under Stress **5**
2.1. Stress and its Influence on the Speech Production Process 5
2.1.1. A Definition of “Speech under Stress” 5
2.1.2. An Extended Model of Speech Production 6
2.1.3. Stress and Emotions 8
2.2. Related Work 8

3. Feature Selection and Evaluation **10**
3.1. Methodology 10
3.1.1. General Analysis Framework 10
3.1.2. Guidelines 11
3.1.3. Verification of Transferability 11
3.2. Databases 12
3.2.1. SUSAS 12
3.2.2. Emo-DB 12
3.2.3. ATCOSIM 12
3.3. Selection of potentially meaningful Features 13
3.3.1. Pitch 13
3.3.2. Intensity 13
3.3.3. Duration 15
3.3.4. Glottal Source Characteristics 16

3.3.5. Vocal Tract Spectrum . . . 18
3.3.6. A TEO-based Feature: TEO-CB-AutoEnv . . . 20
3.3.7. Mel-Frequency Cepstral Coefficients (MFCCs) . . . 22
3.3.8. Harmonicity . . . 23
3.3.9. Zero-Crossing Rate . . . 25
3.3.10. Summary . . . 25
3.4. Phoneme Boundary Detection . . . 28
3.4.1. A Text-independent Approach . . . 28
3.4.2. Detection using Hidden Markov Models (HMMs) . . . 30
3.5. Feature Evaluation and Classification . . . 31
3.5.1. Numerical Feature Evaluation . . . 31
3.5.2. Graphical Feature Evaluation . . . 32
3.5.3. Dimension Reduction . . . 32
3.5.4. Classification . . . 34

4. Implementation 36
4.1. Feature Extraction . . . 36
4.1.1. Overview . . . 37
4.1.2. Basic Features (Praat) . . . 37
4.1.3. Further Feature Extraction (Matlab) . . . 37
4.2. Feature Evaluation . . . 41
4.2.1. A Note on Iterative Feature Evaluation . . . 42
4.2.2. Classification . . . 42
4.3. Other Issues . . . 43
4.3.1. Frame Rate . . . 43
4.3.2. ATCOSIM workload calibration . . . 43
4.4. Occured Problems . . . 44
4.4.1. Bad SUSAS Label Files . . . 44
4.4.2. Reduced Number of ATCOSIM Files . . . 44
4.4.3. Invalid Values in Feature Matrix . . . 45

5. Results 47
5.1. Experiment 1: Talking Styles . . . 47
5.1.1. Individual Results for SUSAS . . . 47
5.1.2. Individual Results for Emo-DB . . . 49
5.1.3. Combination of Results . . . 50
5.2. Experiment 2: Workload Tasks . . . 50
5.2.1. Individual Results for SUSAS . . . 50

5.2.2. Individual Results for ATCOSIM 52
5.2.3. Combination of Results 52
5.3. What about the Benchmark? 53

6. Discussion of Results and Perspective **54**

A. Code Examples **56**
A.1. Praat Feature Extraction Script 56

Colophon **vii**

Bibliography **xii**

List of Figures **xiv**

List of Tables **xv**

Chapter 1.

Preface

1.1. Motivation

Stress is the response to physical or mental challenges and is accompanied by specific emotions. The emotional state can – consciously or unconsciously – affect the speech behavior.

With this knowledge, it must be possible to draw conclusions on the emotional state from *identifying speech markers of stress* within the speech signal, which can be obtained by extracting appropriate (i.e., descriptive) features.

The speech signal is an interesting source for stress analysis, since it can be measured in a non-invasive, contact-free, and non-intrusive way (see Fig. 1.1).

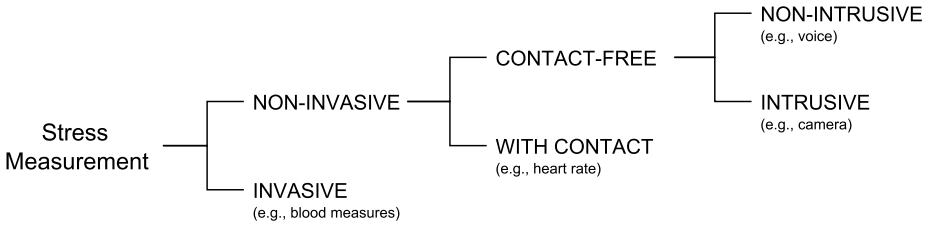


Figure 1.1.: Taxonomy of stress measurement methods.

A speech monitoring system which is able to quantify a speaker's *degree of stress* could serve as a measure of human performance wherever an individual is responsible for other people's safety. In the area of air traffic management (ATM), an air traffic control operator is in charge of a certain *sector* of the airspace. Sector sizes are regularly conformed to the respective traffic loads, subject to a supervisor's subjective decision. These decisions could be objectified by means of indicators of human stress provided by such a monitoring system [19].

Likewise, the physical and mental state of a pilot could be observed in order to activate emergency procedures in case of impending loss of control over the aircraft. Knowledge of a speaker's emotional state can also be beneficial in improving the performance of speech detection algorithms by taking into account its impact on the physiological properties of speech

production; other interesting fields include neutral-to-emotional speech synthesis or "emotion equalization" in speech.

1.2. Objective

Many approaches in the area of emotional speech and speech under stress have been presented over the last years, most of them concentrating on improving classification rates for different talking styles. Prevalently, the genus of *talking styles* implies classes of different qualities; while some are apparently related to certain emotions or kinds of stress (e.g. angry, loud or fast), others like question depend on the situation rather than on the present emotional state.

The majority of these studies employ the SUSAS database (see 3.2.1), which is a discrete set of single-word utterances with very limited vocabulary. This fact implies that up to now, a significant amount of research results have been obtained using speech fragments without any contextual relation.

By contrast, the aim of this thesis is to construct a "transferable" analysis framework which performs equally well on any kind of emotional or stressful speech material; without requesting additional metadata such as labelled phoneme boundaries or a-priori knowledge on the speaker (as it is the case with, e.g. model-based approaches). A set of features is to be found that matches the above-mentioned prerequisites best possible.

An existing approach operating on the SUSAS database is expanded by introducing additional features, before the analysis framework is applied to a German database of emotional speech and an English corpus of non-prompted air traffic control speech to examine the feature set's transferability with respect to foreign languages, continuous speech samples, and workload-related tasks.

1.3. Restrictions

The speech material representing different emotional states consists of acted speech solely; i.e., it has not been recorded in "real-life" situations. While this ensures full control over recording quality and content on the one hand, it prevents control over the quality of the expressed emotion on the other hand, since the speaker may tend to accentuate some features while suppressing others [17].

Concerning workload-related stress, it is important to point out the difference between the terms of *taskload* and *workload*. While the former is an objective measure for the demand of the work (and the parameter that can be specified via the experimental setup), the latter stands for the subjective capacity utilization (and the parameter that impacts the speech production process). Since human beings can hardly be approximated as linear, time-invariant systems, there will always be a nonlinear relationship between taskload and workload.

An interesting issue in this context is the relation between taskload and *performance*, as a speech monitoring system's intended field of application could be called "performance prediction". The effect that a very low taskload can also lead to high workload and thus to lower performance has been described by de Waard [16] and is sketched in figure 1.2.

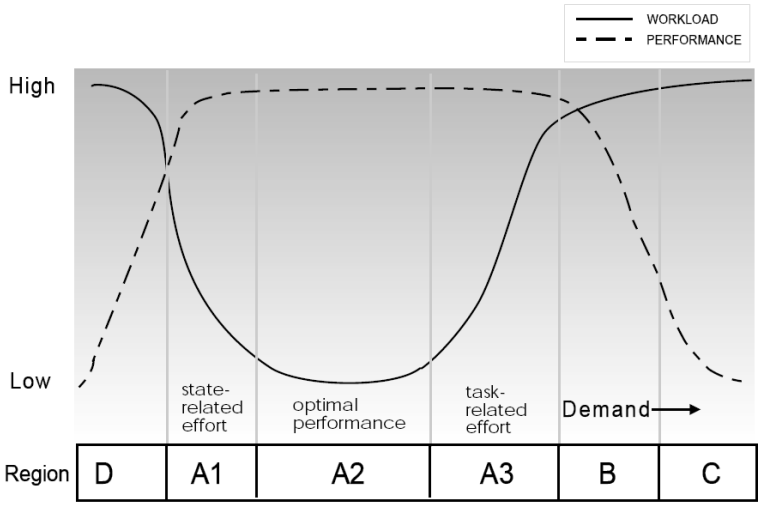


Figure 1.2.: Workload and performance as a function of the demand (from [16]).

The fact that people differ in physical and mental toughness will furthermore lead to a wide statistical spread of feature values. This results in overlapping classes in the *feature space*, which impedes the classification process.

Finally, the broad definition of the term "stress" and the fact that a "neutral" state is merely defined by the absence of stress (cp. 2.1.1) may cause a small overlap when intersecting optimum feature sets of different stress types, which is the technique to obtain a "transferable" set of features.

1.4. Explanation of Terms

Stress

Wherever the term "stress" appears in this thesis, it always refers to some kind of physical or psychological pressure influencing an individual. The linguistic meaning of "stress" (as the emphasis given to a syllable) is never the subject of discussion.

Emotional Classes

Since emotions and stress are highly related to each other (see 2.1.3), the term "emotional class" is used to describe acted emotions as well as workload levels or actual stress.

Phonemes

The *Encyclopædia Britannica* defines a phoneme to be "the smallest unit of speech distinguishing one word (or element) from another" [36]. Roughly speaking, every language of the world has its distinct set of phonemes from which speech is constructed by concatenating these single *unit sounds* to meaningful utterances.

In most western languages, it is not the case that a certain phoneme can be assigned to one corresponding letter in the alphabet. The phoneme /k/ can, e.g., be represented by the letters *k* and *c*, as in *kit* and *cat*.

Features and Feature Characteristics

To be able to differentiate between signal properties extracted directly from the speech signal (which are certainly "features") and descriptive scalar values derived from these properties, the latter is referred to as "feature characteristics". The n-dimensional feature vector used in the classification process consists of n feature characteristics, each representing a statistical property of a corresponding time series of feature values.

1.5. Outline

This thesis is organized as follows:

Chapter 2 starts off by looking at different types of stress and by pointing out how they influence the speech production process. A general survey of research on emotional speech and speech under stress concludes the introduction into the topic.

Chapter 3 outlines the presented work including the general analysis framework, speech databases used for the experiments, features taken into consideration, the task of phoneme boundary detection, numerical and graphical feature evaluation techniques, methods to reduce the feature space dimension, and the classification procedure.

Chapter 4 takes a closer look at the algorithms implemented in Praat and MATLAB, explaining parameters and settings used.

Chapter 5 presents the results of the two main experiments that have been conducted to investigate the *degree of transferability*, which is the main issue of this work.

Chapter 6 discusses the results presented in chapter 5, pointing out achievements as well as open tasks. Furthermore, an outlook on future research as a consequence of the presented work is given.

Chapter 2.

Analysis of Speech under Stress

2.1. Stress and its Influence on the Speech Production Process

2.1.1. A Definition of "Speech under Stress"

Stress is the response to physical or mental challenges and can be caused by a variety of reasons. The stimuli producing a stress response are referred to as *stressors* and have been classified by Hansen et al. [24]. Table 2.1 sets examples for each of the four categories.

Stressor Order	Description	Stressors
0	Physical	Vibration, Acceleration (G-force), Personal Equipment, Pressure Breathing, Breathing Gas Mixture
1	Physiological	Medicines, Narcotics, Alcohol, Nicotine, Fatigue, Sleep Deprivation, Dehydration, Illness, Local Anesthetic
2	Perceptual	Noise, Poor Communication Channel, Poor Grasp of Language
3	Psychological	Workload, Emotion, Task-related Anxiety, Background Anxiety

Table 2.1.: Taxonomy of Stressors (from [24]).

Due to this multitude of possible stressors, *speech under stress* is defined as speech showing any divergence from a pre-defined "neutral" state regarding speaking style, selection and usage of words, or duration of single utterances [22].

Stress can be seen as one of several factors affecting speech; what implies, of course, that a specific level of stress does not necessarily lead to comparable responses when affecting different individuals. Figure 2.1 depicts the relationship of speech, stress, and respective influences.

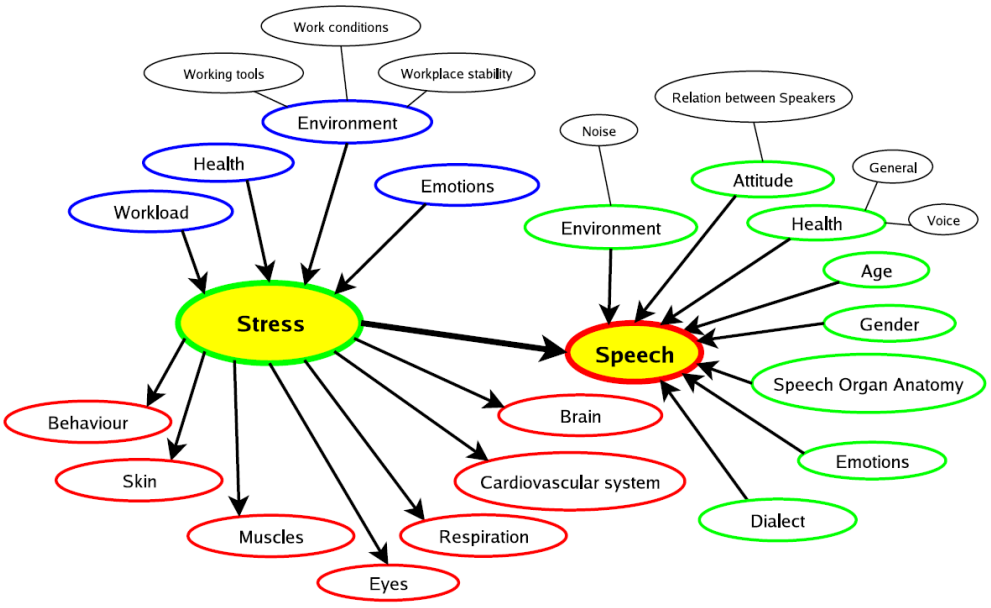


Figure 2.1.: Stress as one of several influences on the speech production process (from [19]).

2.1.2. An Extended Model of Speech Production

The established model of speech production can be sketched as follows: the lungs produce an air flow causing vibration of the vocal folds located within the larynx (see figure 2.2). As a consequence of this, the air flow is cut into audible pulses (the *source*), which are subsequently spectrally modified by the articulators (which include all parts of the vocal tract above the larynx – the *filter*), before being radiated from the lips.

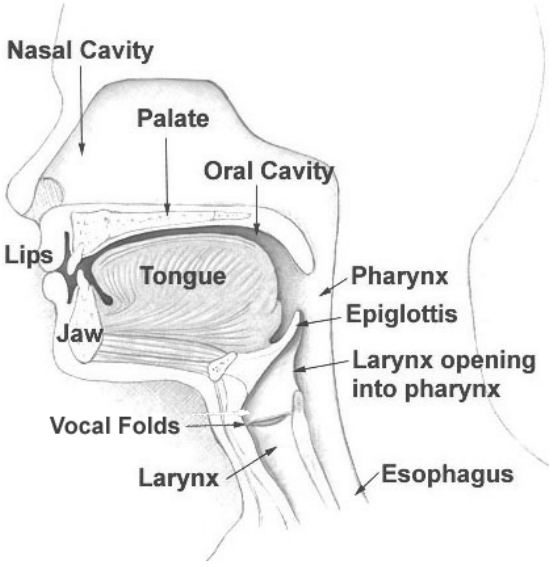


Figure 2.2.: The human vocal tract (from [46], adapted).

To study all possible effects of stress on the speech production process, this purely physical model has to be extended by the "control unit" workflow; i.e., what happens inside the brain before muscular actions are executed.

The speech production model shown in figure 2.3 was introduced by Hansen [24]. It starts with the *idea* what to say, followed by the creation of an appropriate sentence (*linguistic programming*). This sequence of words to produce has to be translated into a sequence of *articulatory targets* first, before appropriate *neuro-muscular commands* are created and transmitted to the muscles controlling the respiratory system and the vocal tract.

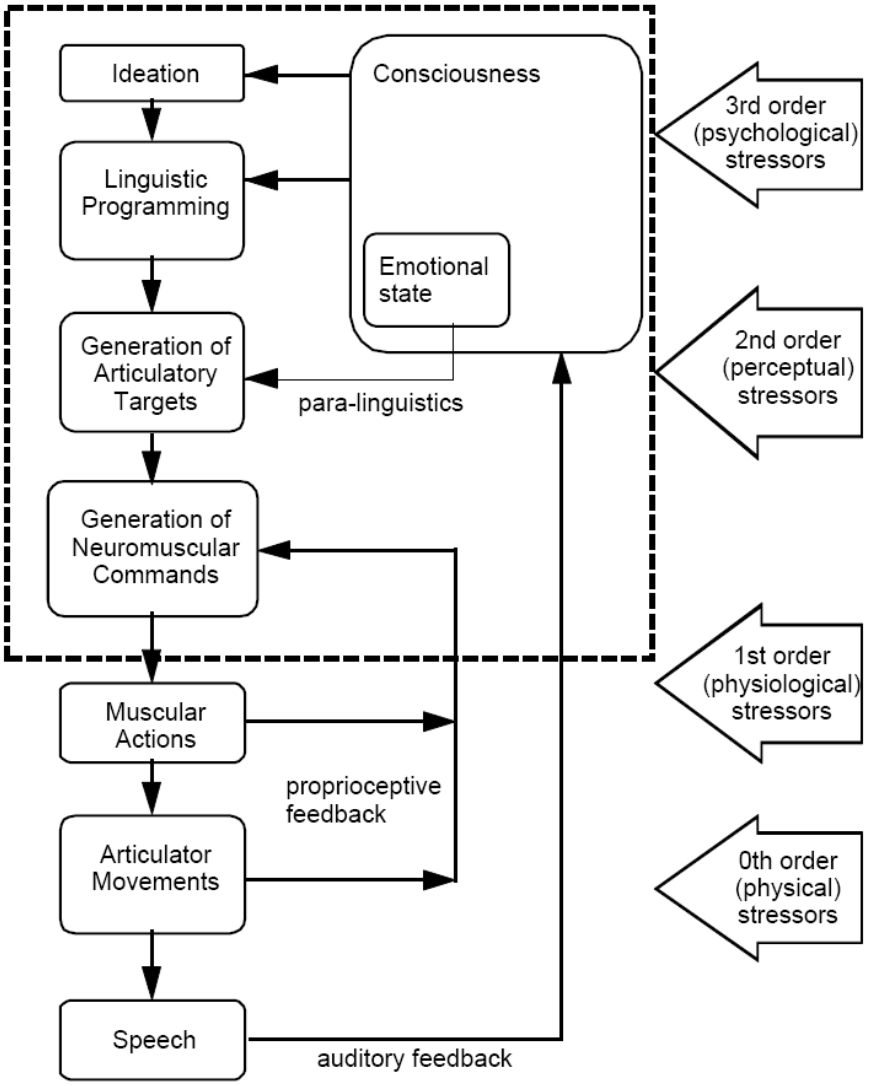


Figure 2.3.: An extended model of speech production and the influence of external stimuli (stressors) of various orders (from [24]).

Third-order stressors affect the speech production process at its highest level. A high workload or emotional states like anxiety or fear may affect the ideation process and the creation of the utterance; even the articulary target generation will be impacted via para-linguistics.

Second-order stressors have their effects at the conversion of the linguistic program into neuro-muscular commands, with noise being the most prominent stressor. The term *perceptual stressor* indicates that there is some kind of conscious interpretation of the stressor [35], but without involving higher-level emotions.

First-order stressors modify the neuro-muscular signal transduction process and thus provoke changes in articulator movement; the proprioceptive feedback loop may also be affected. Responsible are chemical effects in most instances, be it externally (e.g., medical or narcotic drugs) or internally triggered (e.g., illnesss or fatigue).

Zero-order stressors directly result in physical changes to the speech production system. The *mental stage* is not affected, but the articulator responses change due to some kind of force they are exposed to.

2.1.3. Stress and Emotions

Up to now, the terms *stress* and *emotions* have been used with equal meaning. This is, of course, not completely true, since emotions belong to the category of third-order stressors that influence consciousness and emotional state of an individual. At the same time we know that any kind of stress – even if caused by lower-order stressors – will always be accompanied by specific emotions [22].

However, the fact that we can not precisely distinguish between stress and emotions has little effect on the recognition task. Hansen and his group [22, 24, 50, 54] treat acted emotional speech in the same manner as speech under background noise and speech produced by performing a computer response task, and just distinguish between *simulated* and *actual* stress.

Throughout the literature, both terms are widely equated with each other; approaches focusing on emotion solely report similar findings and work with similar features as approaches concentrating on stress. An interesting point is made by Cowie and Cornelius [14], who state that it has not yet been verified if "positive" emotions such as happiness may produce comparable outcomes as stress (which is always associated with "negative" emotions such as anger), e.g., an increase in pitch or intensity.

This thesis uses the term *emotional class* globally; i.e., for acted emotions as well as for different workload levels and for situations in which people are exposed to "actual stress".

2.2. Related Work

Summarizing an ESCA-NATO workshop on speech under stress in 1995, Murray et al. [35] published a paper in which a basic definition of stress is given and a variety of stress models is proposed. An important issue is the distinction between the causes and the effects of stress and how they are related.

As a follow-up, Hansen et al. [24] authored a comprehensive technical report for the NATO research and technology organization in 2000. Although focusing on military speech technology, this paper serves as a prime summary of speech-under-stress related work by the time; including definitions, a database overview, a chapter on features for analysis and one on stress classification and detection methods.

A more general and compact work with similar content was published in the *Lecture Notes in Computer Science* series in 2007 [22], which forms the basis of investigations reported in this diploma thesis (cp. section 3.1.2).

Ververidis and Kotropoulos [48] overview the area of emotional speech recognition concerning available databases, typical features used for analysis, and classification techniques including artificial neural networks, multichannel Hidden Markov Models, and mixtures of HMMs.

Hagmüller et al. [19] concentrate on acoustic correlates of workload-induced stress in speech. While a bigger part of the paper corresponds to Hansen's work [24, 22], it still is a fundamental work concerning workload-induced stress.

These publications are the basic essentials from literature on speech under stress. Those papers dealing mainly with features considered in this thesis are mentioned in the respective *Literature* subsections in 3.3.

Chapter 3.

Feature Selection and Evaluation

3.1. Methodology

3.1.1. General Analysis Framework

A variety of low-level features is extracted from the buffered audio signal, resulting in one value per feature and frame. For each of these feature series, mean and variance as well as other specific *feature characteristics* are computed, which are then evaluated regarding their ability to separate between emotional classes.

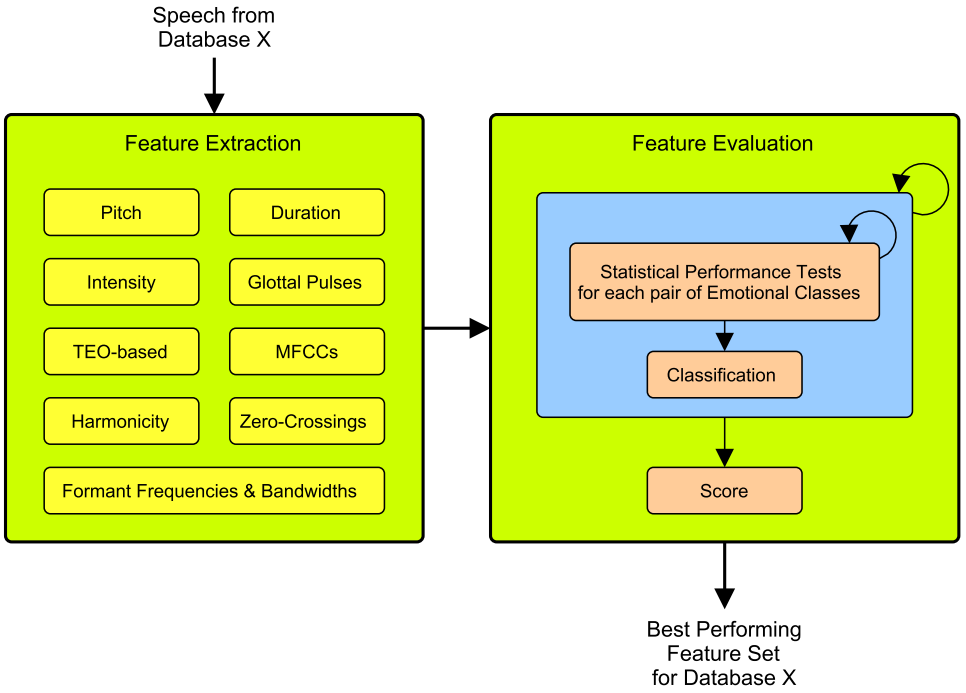


Figure 3.1.: The feature extraction and evaluation framework.

The discriminative power of single features is estimated by means of appropriate statistical tests (see section 3.5.1), resulting in a feature ranking list for a selected combination of two

emotional classes. The best performing set of features is then determined iteratively.

Subsequently, a supervised classification method (see section 3.5.4) is employed in a cross-validation process. Its outcome is the percentage of correctly assigned emotional classes, which is taken as a measure of performance.

Features and derived characteristics are described in detail in section 3.3. Figure 3.1 represents the two main analysis stages, extraction and evaluation, in a compact way.

3.1.2. Guidelines

Prof. John H.L. Hansen is the director and founder of the Center for Robust Speech Systems at the University of Texas at Dallas (USA). He is the author of more than 250 journal and conference papers and several books in the area of speech and signal processing; his first paper on stress and its acoustic correlates in the speech signal was published in 1989 [25].

In 2007, he published a chapter in the book series *Lecture Notes in Computer Science* [22] considering analysis and recognition of speech under stress, which summarizes the findings of many years of research in this area and serves as a benchmark for this thesis.

So the first step is to verify Hansen's results by performing feature extraction on the SUSAS database (see 3.2.1), using just those methods suggested in the above-mentioned article. The proposed features include fundamental frequency (pitch), duration of single phonemes, intensity level, spectral level and slope of glottal pulses, and frequency and bandwidth of the first two formants.

To broaden the spectrum of research, jitter and shimmer of glottal pulses, the first 12 mel-frequency cepstral coefficients (MFCCs), harmonicity, zero-crossings, and a feature based on the Teager Energy Operator (TEO-CB-AutoEnv) are extracted as potentially meaningful features in addition to the Hansen approach.

3.1.3. Verification of Transferability

Since the main intent of the presented work is to produce results which are applicable to speech under a broad spectrum of stress types, two other databases with differing focuses are employed (full particulars on these databases can be found in section 3.2).

The dependency on the spoken language and possible effects caused by analysis of continuous speech (rather than single-word utterances) is then studied by analyzing a German emotional database (see 3.2.2) in the same manner as described above. Results are compared for the emotional classes neutral and angry, which are contained in both SUSAS and the Emo-DB database.

Finally, this analysis framework is applied to the ATCOSIM corpus (see 3.2.3) in order to study the effects of variably induced taskload levels for the small, discrete word set versus a large set of non-prompted speech.

3.2. Databases

The presented work investigates speech data from three different databases, which are briefly presented in the following paragraphs.

3.2.1. SUSAS

The SUSAS database [21] comprises Speech Under Simulated and Actual Stress in five different domains and was especially created for investigating speech under stress. The vocabulary covers 35 single-word utterances from aircraft communication. Three of these five domains are selected for analysis; namely *Talking Styles* (neutral, slow, fast, soft, loud, clear, angry and question), *Single Tracking* and *Dual Tracking* computer response tasks.

SUSAS is employed in the majority of literature on speech under stress research and thus taken as a reference for the current work.

3.2.2. Emo-DB

The Emo-DB database [11] consists of 10 different German utterances produced in seven (acted) emotional states, each produced by 5 female and 5 male speakers. The simulated emotions include angry, anxious, bored, disgusted, joyful, sad, and neutral.

This database has been chosen in order to be tested against SUSAS' talking styles domain, which contains two equivalent emotional classes; neutral and angry.

3.2.3. ATCOSIM

The ATCOSIM speech corpus [27] contains about 10 hours of non-prompted, clean Air Traffic Control Simulation speech recorded during real-time simulations. In contrast to the remaining two databases, the data is not categorized into emotional classes, and no metadata in terms of label files with phoneme information is provided.

But since a list of appearing utterances exists, phoneme labels and boundaries can be nevertheless determined with the HTK toolkit (see 3.4.2), using a standardized lexicon file. With a check against the SUSAS computer response domains in mind, emotional classes are assigned approximately (described in section 4.3.2).

3.3. Selection of potentially meaningful Features

3.3.1. Pitch

Motivation. Pitch (or *fundamental frequency*, f_0), is the most prominent and the most widely used speech characteristic for stress classification. This is based on the fact that humans tend to increase their respiration rate in stressful situations, what causes an increase in subglottal pressure during speech, which itself leads to an increased pitch [41].

Literature. Hansen et al. [22, 24] performed t- and F-tests[1] on the SUSAS data and conclude that mean and variance can serve as good indicators over a wide variety of emotional classes, while higher-order moments and contours are no effective traits for stress classification.

Lively et al [33] investigated effects of cognitive workload on the speech production process and noticed that speakers under workload tend to produce an utterance with a monotone pitch (rather than reducing jitter); i.e., pitch variance is decreased. Similar results are reported by Brenner et al. [10], who monitored speakers performing a speeded arithmetic task while talking.

Burkhard and Sendlmeier [12] developed a speech synthesizer for emotional expression and report that pitch variance can especially be used to distinguish between the emotional states sadness and boredom.

Implementation. Pitch is extracted using Praat (see 4.1.2). The algorithm implemented in the software tool performs an acoustic periodicity detection based on an auto-correlation method presented in [5].

Adjustable parameters are: frame rate (set to $10ms$), pitch floor (set to $75Hz$), and pitch ceiling (set to $600Hz$). By specifying the minimum pitch, the analysis window length (which is set to the length of three maximum periods automatically) has implicitly been set to $40ms$. The algorithm further uses a Hanning window.

Pitch tracks for different emotional classes are depicted in figure 3.14 on page 27.

Characteristics. Statistical moments up to fourth order are computed to keep information on the shape of the distribution function. An exemplary pitch distribution is shown in figure 3.2 on page 14.

3.3.2. Intensity

Motivation. As for fundamental frequency, it seems obvious that the intensity of an utterance correlates with the speaking person's emotional state. A noisy environment, an experience or intense emotions are so-called *perceptual stressors* (cp. 2.1.1) which cause an increase in vocal effort to make oneself heard. [22].

[1] The *t-test* checks if the mean values of two normally distributed samples are equal, while the *F-test* provides a measure for the probability that two independent samples have the same variance.

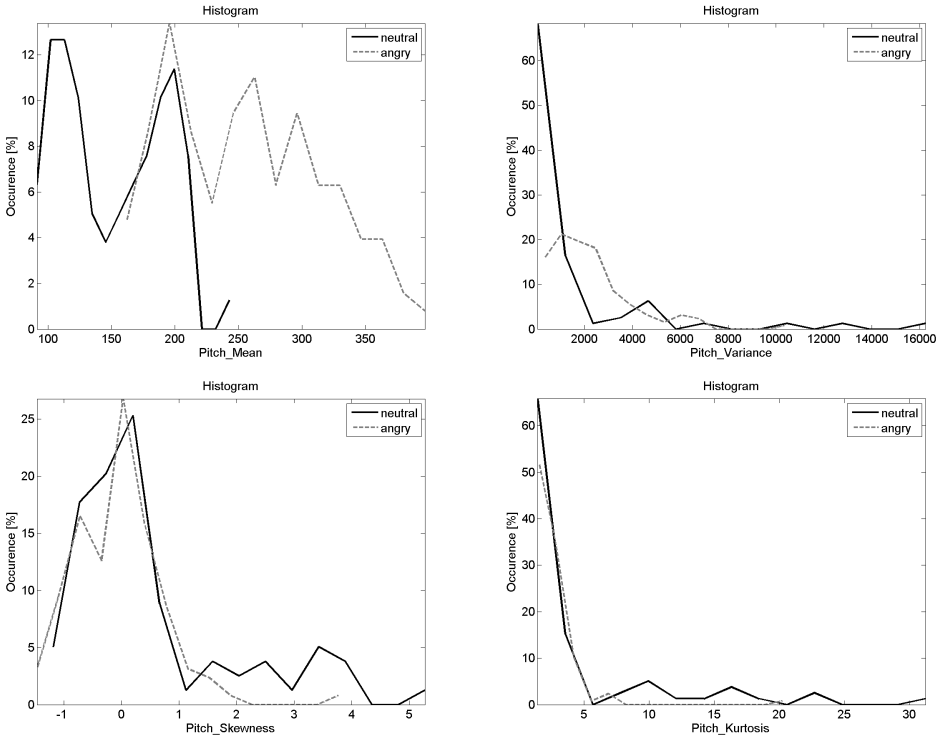


Figure 3.2.: Pitch histogram plots for talking styles angry vs. neutral (Emo-DB database)

Beyond, speakers tend to emphasize parts of an utterance containing the most important information while de-emphasizing others in time-critical situations [13, 30], which might result in increased variance of intensity.

Literature. Hansen at al. [22, 24] extracted RMS energy of single phonemes as well as of complete words, finding that mean intensity possesses a good level of stress discriminating ability. In opposition to the initial assumption, intensity variance is not consistently successful for stress detection, however.

Implementation. The intensity of a discrete-time signal[2] $x[n]$ is computed relative to the human auditory threshold using

$$I = 10 \cdot \log_{10} \left(\frac{1}{N \cdot p_0^2} \sum_{n=1}^{N} x^2[n] \right) \quad [dB\ SPL] \tag{3.1}$$

where N stands for the analysis window length and $p_0 = 2 \cdot 10^{-5} Pa$ is the normative auditory threshold for a $1000 Hz$ sine wave.

Praat's intensity algorithm uses a Kaiser window to ensure sidelobes below $-190 dB$. Adjustable parameters include frame rate and pitch floor, which are again set to $10 ms$ and $75 Hz$,

[2]The signal $x[n]$ is assumed to be a "pure" speech signal. Since the speech data used does not contain any *non-speech* segments in between and pauses at the beginning or at the end of an utterance have been removed via end-point detection, this assumption is valid and there is no need for a Voice Activity Detection algorithm (yet).

respectively. The algorithm choses the analysis window length to be 3.2 (rather than 3) divided by the minimum pitch in order to keep the pitch-synchronous intensity ripple low [8].

Intensity tracks for different emotional classes are depicted in figure 3.14 on page 27.

Characteristics. Mean and variance are computed for each phoneme class individually; i.e., we end up with a total of 8 feature characteristics; two for each of the *phoneme classes* vowel, semivowel, consonant, and diphthong. Overall intensity distributions are shown in figure 3.3.

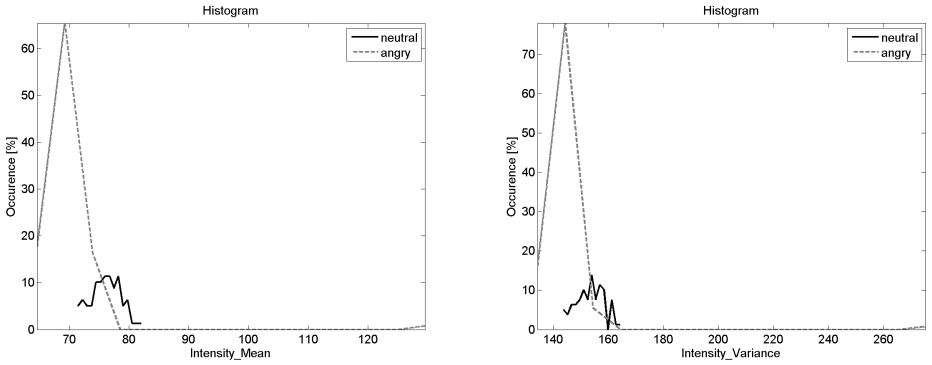


Figure 3.3.: Intensity histogram plots for talking styles angry vs. neutral (Emo-DB database)

3.3.3. Duration

Motivation. An increased respiration rate does not only lead to an increased pitch (cp. 3.3.1), but also affects the temporal pattern when the same amount of words is to be produced within shorter time windows between consecutive breaths.

Furthermore, the above-mentioned intensified emphasis and de-emphasis of more and less important fractions of the utterance (cp. 3.3.2) may also be expressed in terms of changes in (sub-word) duration.

Literature. According to Hansen et al. [22], induced stress may lead to shifts in duration between consonants and vowels while the overall word duration remains constant. Thus, they propose the *consonant-to-vowel duration ratio* (CVDR), the *consonant-to-semivowel duration ratio* (CSVDR), and the *vowel-to-semivowel duration ratio* (VSVDR) as potential discriminating characteristics for stress classification.

Implementation. The duration feature is calculated in a secondary feature extraction step, where a phoneme boundary detection algorithm (see section 3.4) has been executed previously. With this *phoneme time grid* available, phoneme durations are obtained by simply computing the first-order difference.

Characteristics. Mean and variance are computed for each phoneme class individually. In addition, average CVDR, CSVDR, and VSVDR are determined as described above. Distribution

plots for mean and variance of overall durations as well as CVDR can be found in figure 3.4 on page 16.

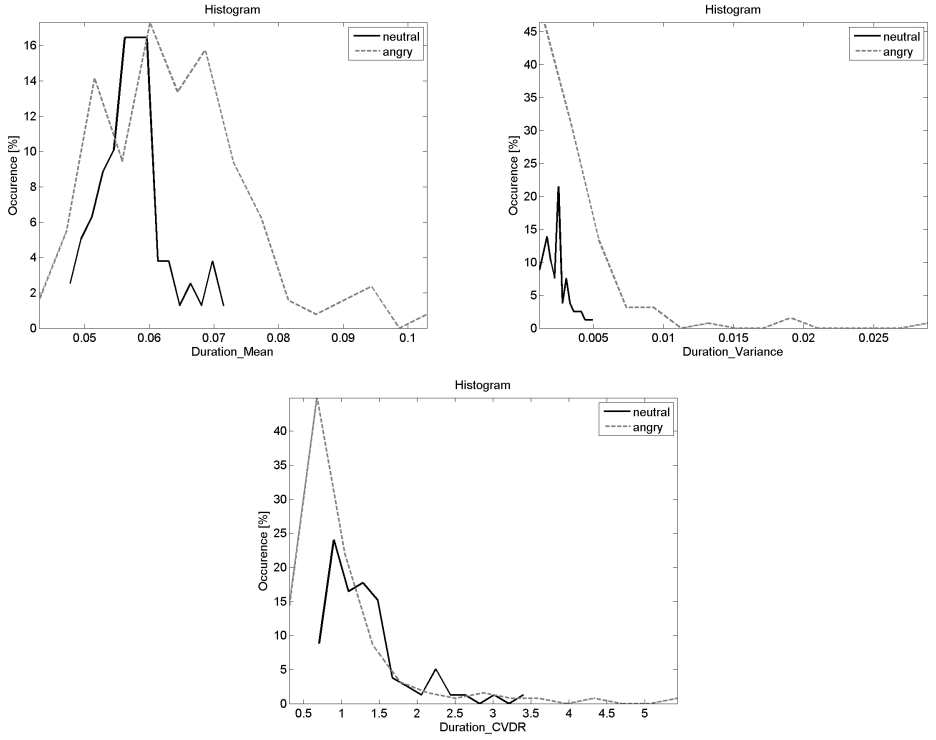


Figure 3.4.: Duration histogram plots for talking styles angry vs. neutral (Emo-DB database)

3.3.4. Glottal Source Characteristics

Characteristics derived from the glottal source include jitter, shimmer, spectral mean, and spectral slope.

Explanation of Terms. *Jitter* is the variation of a periodic signal property; which in this case is the frequency of the glottal pulses. *Shimmer* refers to the variation of the amplitude of single pulses.

Motivation. Responses to induced stress include changes in subglottal air pressure or vocal-fold tension as well as irregular closures of the vocal folds during phonation. Especially the dryness of the mouth in situations of excitement or anger can effect the condition of the vocal folds [22]. These non-uniformities are suspected to affect a variety of measurable characteristics.

Literature. Clary and Hansen [13] investigated spectral glottal features and found out that differentiating features include spectral slope and amplitude; the latter especially in the $2kHz$ to $4kHz$ band [22].

Implementation. Extracting glottal pulses requires prior pitch analysis, which delivers voiced/

unvoiced decisions and frequency information for voiced segments. For each of these voiced segments, a number of glottal pulses is found by detecting the absolute extremum of the amplitude around the interval midpoint. Once found, a recursive cross-correlation searching method is employed which searches for additional points towards the interval edges [8]. The pitch extraction parameters equal those listed in 3.3.1.

The *relative local jitter* is computed once per utterance, which is the average absolute difference between consecutive periods divided by the average period. Similarly, the *relative local shimmer* is calculated by dividing the average absolute difference between the amplitude of consecutive pulses by the average amplitude. Both Features are extracted using Praat.

For features derived from the glottal spectrum, a pitch-corrected Long-Term Average Spectrum (LTAS) is computed, which represents the logarithmic power spectral density (PSD) as a function of frequency relative to the normative auditory threshold. For details on the pitch-corrected method, see [6]. Parameters are set as follows: maximum frequency ($4kHz$), subband width ($50Hz$), shortest and longest period ($0.1ms$ and $20ms$, respectively), and maximum period factor (default setting of 1.3 kept).

Characteristics. All four glottal source-related features are scalar values and thus directly taken as feature characteristics. Exemplary distributions of glottal source-relate features are depicted in figure 3.5.

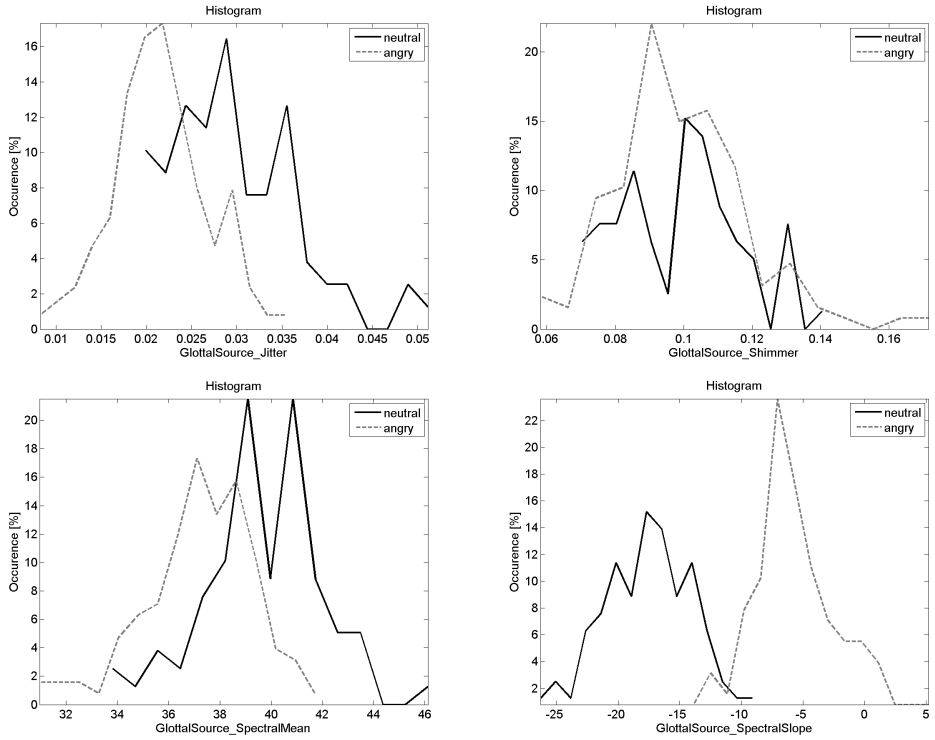


Figure 3.5.: Histogram plots of jitter, shimmer, spectral mean, and spectral slope for talking styles angry vs. neutral (Emo-DB database)

3.3.5. Vocal Tract Spectrum

Motivation. Formants are the characteristic frequency components of human speech. Physical and physiological stressors impact the muscles which control the articulators or the articulator movement itself (cp. figure 2.3 on page 7), so the presence of stress will traceable in the speech signal.

Literature. Hansen and his research group [22] analyzed mean and variance of formant frequencies and bandwidths. They found out that stress certainly does effect typical vocal tract movement, with greater effects on F1 and F2 than on F3 and F4. They performed a series of t-tests, assuming both equal and unequal variance, but arrived at the conclusion that no general statement could be made concerning the quality of single parameters.

Implementation. F1 and F2 are extracted using one of several analysis methods provided by the Praat tool. It resamples the audio file to twice the maximum formant frequency before its spectrum is "flattened" by applying pre-emphasis[3]. Subsequently, LPC coefficients are computed for Gaussian-windowed frames using the Burg method. F1 and F2 are then tracked within a specified frequency range.

Parameters have been set as follows: frame rate to $10ms$, maximum number of formants to 2, maximum formant frequency to $4kHz$ (which is half the sampling frequency for the SUSAS database), effective analysis window length[4] to $25ms$, and the pre-emphasis filter cutoff frequency to $50Hz$.

Extracted formant tracks for different emotional classes can be found in figure 3.14 on page 27.

Characteristics. Mean and variance of center frequencies and bandwidths are computed for F1 and F2, respectively. Histograms of these feature characteristics are shown in figures 3.6 and 3.7.

[3] Vowel spectra have an average spectral slope of $-6dB/octave$. Since formants should match the local peaks rather than the global spectral slope, an inverted low-pass filter with a slope of $+6dB/octave$ is employed to equalize this trend.

[4] The actual Gaussian window has twice the effective length, but values outside the [25%..75%] interval are lowered by more than 96%, and the side-lobe attenuation is three times higher as for a "standard" Hamming window.

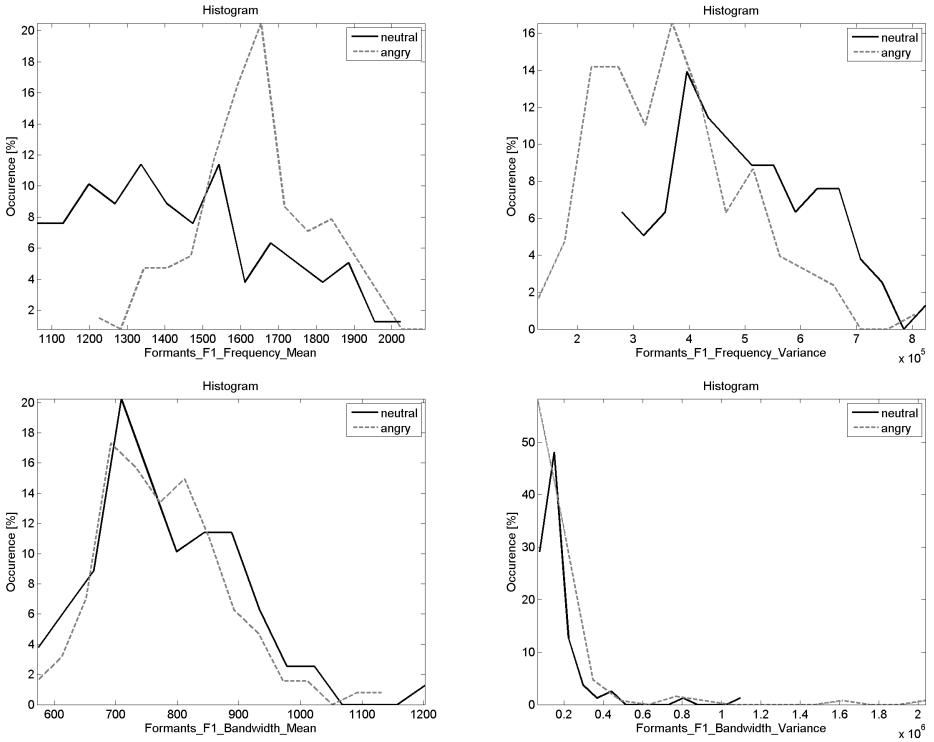


Figure 3.6.: F1 histogram plots for talking styles angry vs. neutral (Emo-DB database)

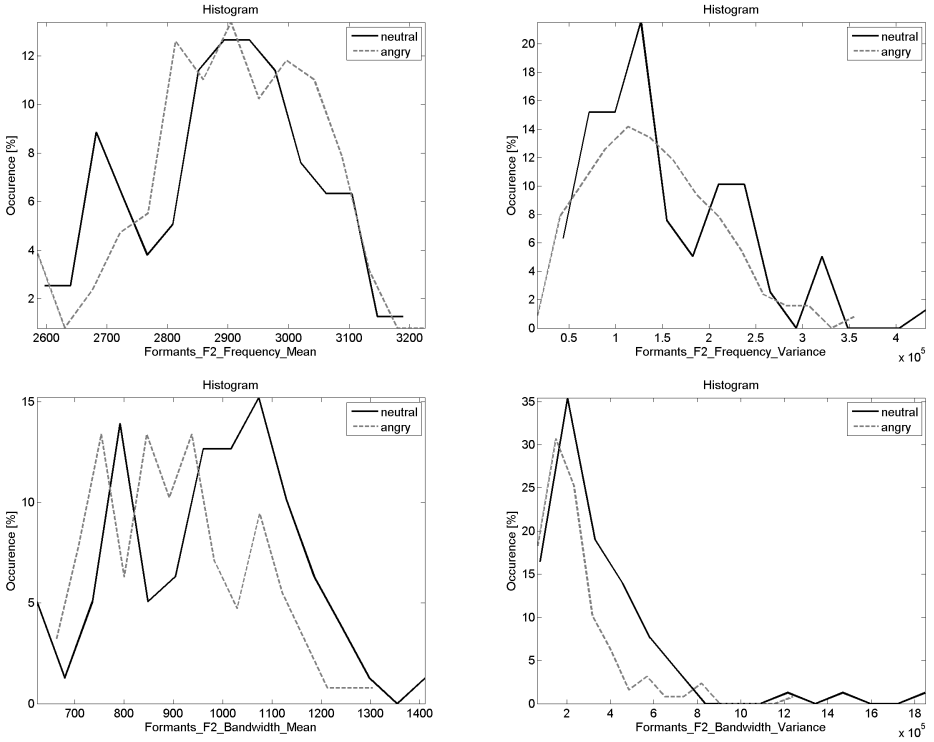


Figure 3.7.: F2 histogram plots for talking styles angry vs. neutral (Emo-DB database)

3.3.6. A TEO-based Feature: TEO-CB-AutoEnv

Theory. All the features mentioned so far are derived from a linear speech production model, assuming that the airflow propagates in the vocal tract as a plane wave. According to studies by Teager [45], however, the flow actually consists of separate and simultaneous vortices distributed throughout the vocal tract, as depicted in figure 3.8.

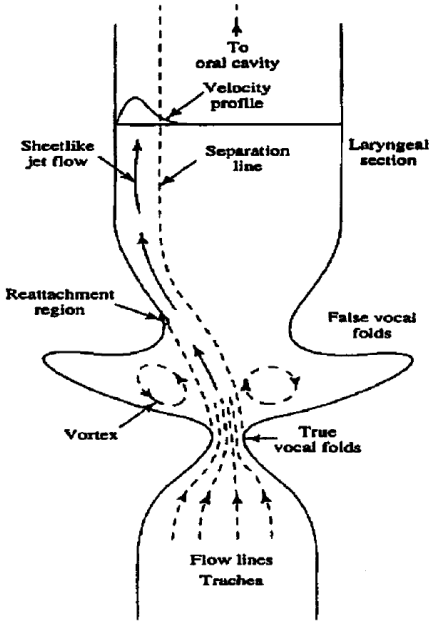


Figure 3.8.: The nonlinear wave propagation model after Teager (from [54]).

Responsible for the "characteristic sound" of a phoneme is, following Teager, not the shape of the vocal tract itself, but rather the resulting airflow properties exciting this resonator. Stating that hearing could be viewed as the process of detecting the energy, Teager developed an energy operator to reflect the instantaneous energy of the nonlinear vortex-flow interactions [54].

Kaiser [31] introduced an elegant form of this *Teager Energy Operator* as

$$\Psi_c = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) \tag{3.2}$$

or, for discrete-time signals,

$$\Psi_d = x^2[n] - x[n+1]x[n-1]\ . \tag{3.3}$$

In both cases, Ψ denotes the TEO, while $x(t)$ and $x[n]$ are the continuous and the sampled speech signal, respectively.

Zhou, Hansen and Kaiser [54] perform critical-band filtering on the speech signal before applying the TEO operator. In a further step, these bandpass TEO profiles are segmented into frames with subsequent autocorrelation envelope analysis performed. For each frame and critical band, the area under this autocorrelation envelope is calculated and normalized by half the frame

length (which equals the area under the "ideal" envelope in consequence of 0% pitch variation within the frame) [54].

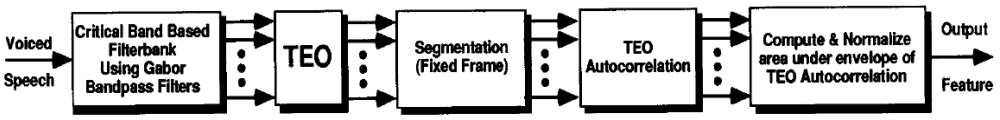


Figure 3.9.: Extraction of the TEO-CB-AutoEnv feature (from [54]).

The idea behind this approach sketched in figure 3.9 is that the autocorrelation envelope can track the variability of the fine energy structure reflected in the TEO critical band partition.

Motivation. As an alternative, nonlinear approach, features derived from the TEO are without any competitor and may reflect properties of the speech production process that are not covered by features derived from the linear model of speech production.

Literature. The research group around Hansen published a number of papers on the usage of the TEO as a new feature for stress recognition and classification [52, 53, 54, 23, 40, 39, 42], most of them demonstrating the discriminative power of the TEO-CB-AutoEnv feature in stress clasification or discussing methods for transforming the [frames $\times$ subbands] matrix (which is the outcome of their proposed TEO feature analysis) into a scalar value.

Jabloun [29] and Fernandez [18] applied an inverse DCT to the logarithm of average Teager energy in spectral subbands and in so doing extracted "TEO-based cepstral coefficients" as features for stress detection.

Implementation. The TEO-CB-AutoEnv feature is implemented in MATLAB, using vowel parts as input. The implementation does exactly follow the approach described above, except that 24 critical bands are used instead of 16 as it is the case in [54].

Characteristics. Mean and variance are computed for the average energy over all critical bands. To reflect the dynamics between adjacent frames, the absolute value of average first-oder differences (*TEO Mean Difference*) is computed. Last, following [39], a characteristic score is obtained using

$$S = \sum_{n=1}^{N} E(n) \cdot W_n(n) - \sum_{n=1}^{N} E(n) \cdot W_{st}(n) \tag{3.4}$$

where $n = 1 \ldots N$ denotes the subband index, W_n and W_{st} are specific weighting schemes for "neutral" and "stressed", respectively, and $E(n)$ represents the overall energy in the n-th critical band. This feature will be referred to as the *TEO Weighted Score*.

Distributions of TEO-derived feature characteristics are shown in figure 3.10 on page 22.

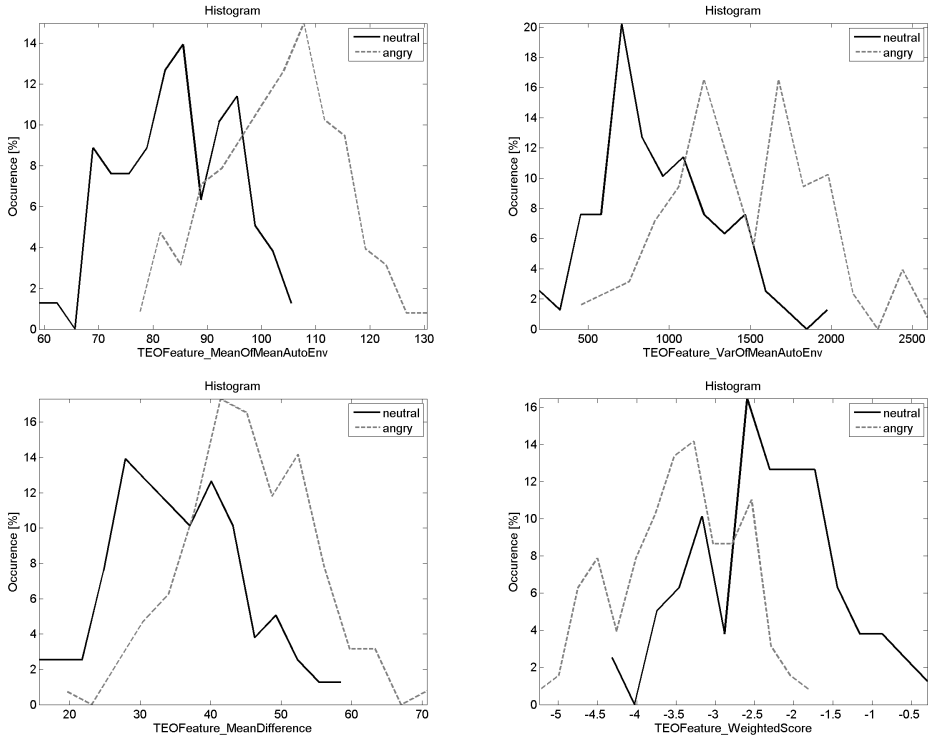


Figure 3.10.: TEO-derived feature characteristic histogram plots for talking styles angry vs. neutral (Emo-DB database)

3.3.7. Mel-Frequency Cepstral Coefficients (MFCCs)

Motivation. Mel-Frequency Cepstral Coefficients represent the spectrum of a signal in a compact way, at which the resolution of this representation depends on the number of coefficients chosen. MFCCs are used widely in the area of speech processing; mainly in speech recognition and speech coding algorithms, where the cepstral domain is useful for the extraction of the spectral envelope and the separation of signals [47].

Literature. Zhou et al.[54] take MFCCS and pitch information as reference features for investigations on stress classification performance and report that the TEO-based features perform significantly better; although it is important to note that a very small subset of six (!) words out of the SUSAS vocabulary was used for testing, five of them consisting of one single syllable only. So the question is if these results are valid for other kinds of data as well; but this is exactly what this thesis is about.

Hansen [26] did not employ MFCCs directly as a feature for stress classification, but rather makde use of the coefficient's sensitivity to noise to develop a stress compensation scheme, which is used in a preprocessing stage to speech recognition systems.

Bou-Ghazale [9] points out that, depending on the way of computation (linear-perdiction based or DFT based), MFCC performance may vary in terms of robustness in noise or for speech

under stress.

Implementation. For each frame, a MFCC vector is computed as follows: the Fourier transform of a windowed signal block is mapped onto the Mel scale (which is a quasi-logarithmic scale closely related to the critical band scale) using triangular filters. The logarithm of the filterbank output is then again transformed by means of a Discrete Cosine Transform (DCT), such that the ith MFCC ($i = 1 \ldots N$) can be obtained by [15]:

$$c_i = \sum_{j=1}^{N} P_j \cos\left(\frac{\pi \cdot i}{N(j-0.5)}\right) , \tag{3.5}$$

where P_j denotes the power in the jth filter (in dB). Discarding higher-order coefficients ensures a certain smoothness of the mel-frequency spectrum.

The Praat implementation accepts the following parameters: number of coefficients (12 plus c_0), frame rate ($10ms$), analysis window length ($25ms$), position of first filter ($100mel$), and distance between filters ($100mel$).

Harmonicity tracks for different emotional classes are depicted in figure 3.14 on page 27.

Characteristics. The first four coefficients are each averaged over the whole utterance. As an additional characteristic, the MFCC variances are calculated for each frame and subsequently averaged as well. This will be referred to as the *MFCC mean variance*. Exemplary distributions of MFCC-derived feature characteristics are shown in figure 3.11 on page 24.

3.3.8. Harmonicity

Motivation. Harmonicity (also known as *Harmonics-to-Noise Ratio*, HNR) is a measure of the degree of acoustic periodicity. It can be used as a measure of voice quality, as a hoarse voice will show lower harmonicity values than a "healthy" voice.

Literature. Alter et al. [1] examine harmonicity as one of several indicators for the perceptional features *breathiness* and *roughness* on a self-recorded database comprising the acted emotional states happy, neutral and cold anger. They report slightly higher values of harmonicity for the angry talking style than for the others; it is furthermore noticed that harmonicity correlates with the *accentuation type of the sentence*, i.e., which syllable is emphasized.

Implementation. The Praat algorithm performs a periodicity detection using a forward cross-correlation analysis method described in [5]. Harmonicity is expressed in dB: a value of, e.g., $20dB$ indicates that 99% of the signal energy is contained in the periodic part and just 1% in the stochastic part (= noise), since $10 \log_{10}(\frac{99}{1}) \approx 20$.

Characteristics. For harmonicity, lower-order statistical moments (mean and variance) are calculated. Figure 3.12 on page 24 shows distributions for mean and variance of the extracted harmonicity tracks.

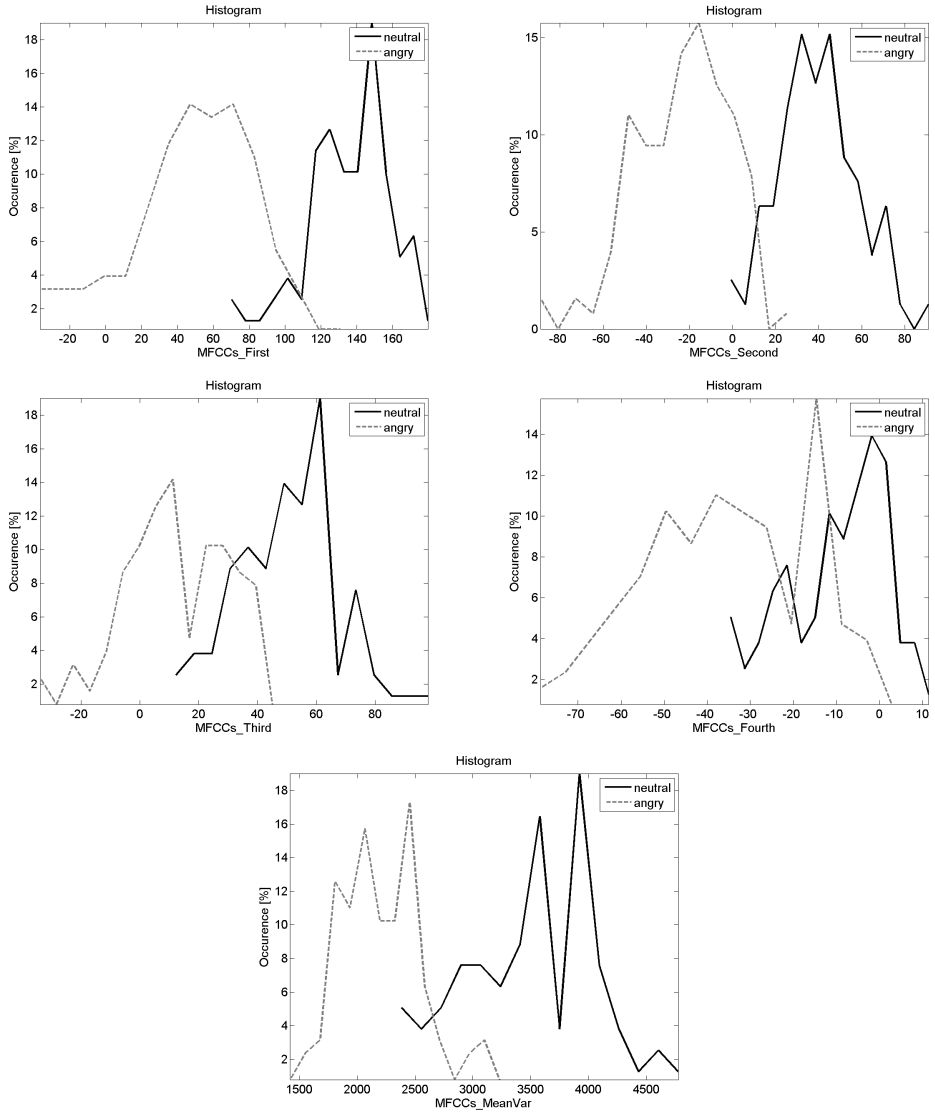


Figure 3.11.: MFCC histogram plots for talking styles angry vs. neutral (Emo-DB database)

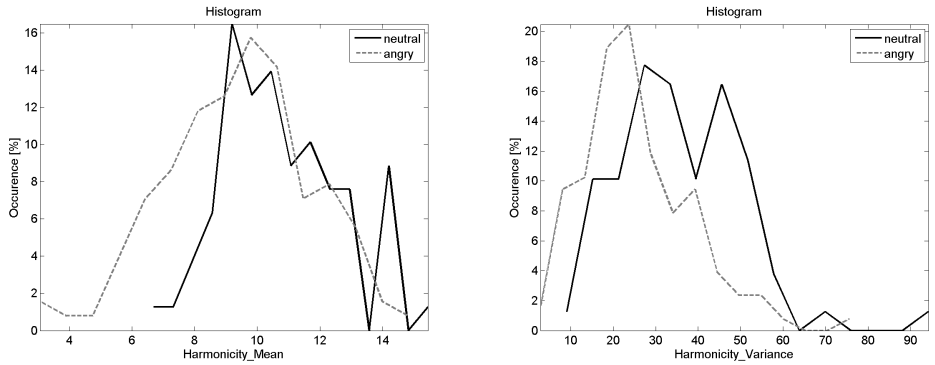


Figure 3.12.: Harmonicity histogram plots for talking styles angry vs. neutral (Emo-DB database)

3.3.9. Zero-Crossing Rate

Motivation. The zero-crossing rate measures the number of sign changes in the time domain. This can be interpreted as a measure of noisiness or tonality (in musical terms); it may even be used as a coarse approximation of pitch.

Literature. Junqua [30] compared speech produced in the presence of noise with clean speech and arrived at the conclusion that the zero-crossing rate increases for most phoneme classes – but, surprisingly, for female speakers only.

Implementation. For each windowed block extracted from the time-domain speech signal $x[n]$, the number of zero-crossings is computed with

$$ZC = \frac{1}{2}\sum |sgn(x[n]) - sgn(x[n-1])| \quad \text{with } sgn(x) = \begin{cases} -1 & \text{for} \quad x < 0 \\ 1 & \text{for} \quad x \geq 0 \end{cases}, \tag{3.6}$$

which has to be subsequently normalized by the block size in samples to yield the zero-crossing rate. This feature has been implemented in Matlab.

Characteristics. As for harmonicity, mean and variance are calculated as feature characteristics. The corresponding exemplary distributions are shown in figure 3.12.

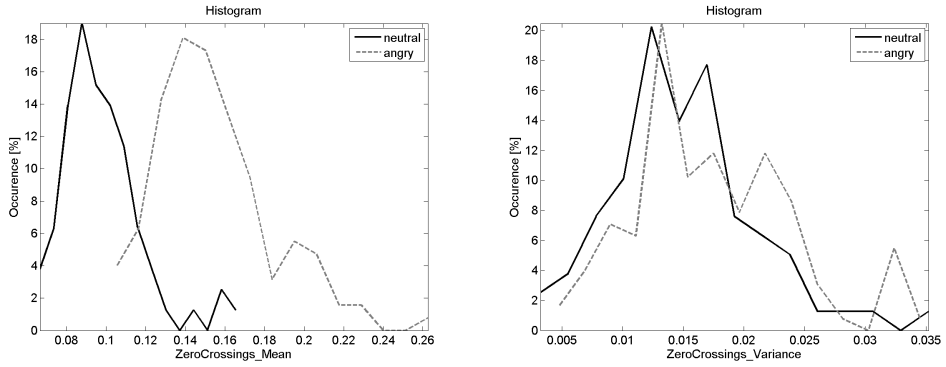


Figure 3.13.: Histogram plots of zero-crossing rates for talking styles angry vs. neutral (Emo-DB database)

3.3.10. Summary

A total of 29 feature characteristics is derived from the low-level features listed in this section. For the sake of clarity, the characteristics are summarized in table 3.1 on page 26

Feature	Derived Characteristic
Pitch	Mean value
Pitch	Average deviation
Pitch	Standard deviation
Pitch	Variance
Pitch	Skewness
Pitch	Kurtosis
Intensity	Mean (for each phoneme class individually)
Intensity	Variance (for each phoneme class individually)
Duration	Mean (for each phoneme class individually)
Duration	Variance (for each phoneme class individually)
Duration	Consonant/vowel duration ratio
Duration	Consonant/semivowel duration ratio
Duration	Vowel/semivowel duration ratio
Jitter	
Shimmer	
Mean of glottal pulse spectrum	
Slope of glottal pulse spectrum	
F1 Frequency	Mean
F1 Frequency	Variance
F1 Bandwidth	Mean
F1 Bandwidth	Variance
F2 Frequency	Mean
F2 Frequency	Variance
F2 Bandwidth	Mean
F2 Bandwidth	Variance
Harmonicity	Mean
Harmonicity	Variance
MFCC	c_1, c_2, c_3, c_4
MFCC	Mean variance
TEO-CB-AutoEnv	Mean of mean
TEO-CB-AutoEnv	Variance of mean
TEO-CB-AutoEnv	Mean difference
TEO-CB-AutoEnv	Weighted score
Zero-Crossings	Mean
Zero-Crossings	Variance

Table 3.1.: Feature characteristics taken for evaluation.

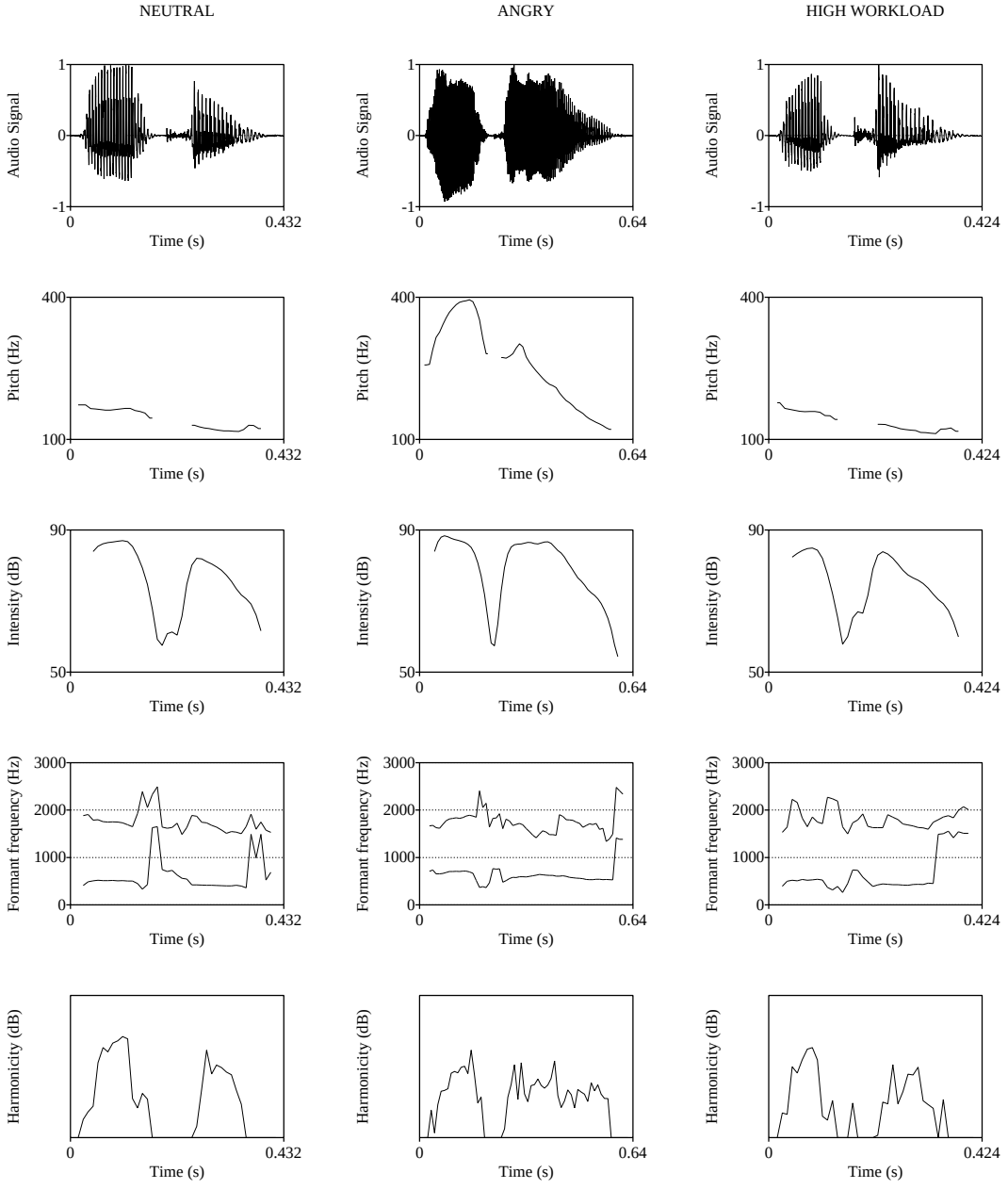


Figure 3.14.: PCM waveforms and a variety of extracted features for the word *enter* (from the SUSAS database) spoken under three different stress conditions.

3.4. Phoneme Boundary Detection

3.4.1. A Text-independent Approach

In order to determine the phoneme boundaries automatically, a *text-independent* approach presented in [3] is followed. Independence of text means that phonetic segmentation of speech can be performed without prior knowledge of the phoneme sequence. The algorithm consists of two main steps, which are presented in the following.

Preprocessing Step

Signal preprocessing includes critical-band filtering, equal loudness pre-emphasis, and spectral intensity-loudness compression.

Critical-band filtering is realized by means of an *equivalent rectangular bandwidth* (ERB) filterbank[5] using an efficient implementation of the Patterson-Holdsworth gammatone filter bank described by Malcolm Slaney in [44].

For equal loudness pre-emphasis, a (discrete) approximation of an equal loudness curve for frequencies up to $5kHz$ (found in [49]) given by

$$E[k] = \frac{k^4(k^2 + 56.8 \cdot 10^6)}{(k^2 + 6.3 \cdot 10^6)^2(k^2 + 0.38 \cdot 10^9))} \tag{3.7}$$

is used to pre-emphasize the short-time spectra $X_{CB_i}[n,k]$ (with critical-band index i, frame index n, and frequency bin index k) over all critical bands. This is done by simple multiplication:

$$X_{CB_i,PE}[n,k] = X_{CB_i}[n,k] \cdot E[k] \; . \tag{3.8}$$

Intensity-loudness compression is nothing but a cubic root compression of the spectral magnitude:

$$X_{CB_i,PE,comp}[n,k] = (X_{CB_i,PE}[n,k])^{\frac{1}{3}} \; . \tag{3.9}$$

After subsequent summation over all frequency bins,

$$x_i[n] = \sum_k X_{CB_i,PE,comp}[n,k] \; , \tag{3.10}$$

the outcome of this preprocessig stage is the spectral energy within critical bands for each frame. It is represented by a collection of M time sequences,

$$\tilde{X} = \{x_i[n]\} \quad \text{with} \quad n = 1 \ldots N, \; i = 1 \ldots M \; , \tag{3.11}$$

where N is the total number of frames and M corresponds to the number of critical bands[6].

[5] The ERB scale is closely related to the Bark scale, but defined analytically instead of being measured. For detailled information on this topic, see [28].

[6] In the presented approach, the number of time sequences is further reduced by summing over groups of three Bark bands. Since no special point was found in that, this step is omitted in the implementation.

Boundary Detection Step

For each single time sequence $x_i[n]$, a set $\{J_i^a[n]\}$ of so-called *jump functions* is defined as follows:

$$J_i^a[n] = \left| \sum_{m=n-a}^{n-1} \frac{x_i[m]}{a} - \sum_{m=n+1}^{n+a} \frac{x_i[m]}{a} \right| \tag{3.12}$$

where the parameter a controls the analysis interval of this peak detection function. In words: the averages over a previous and following frames are compared with each other, and great differences lead to high peaks in the jump function. An example is given in fig. 3.15.

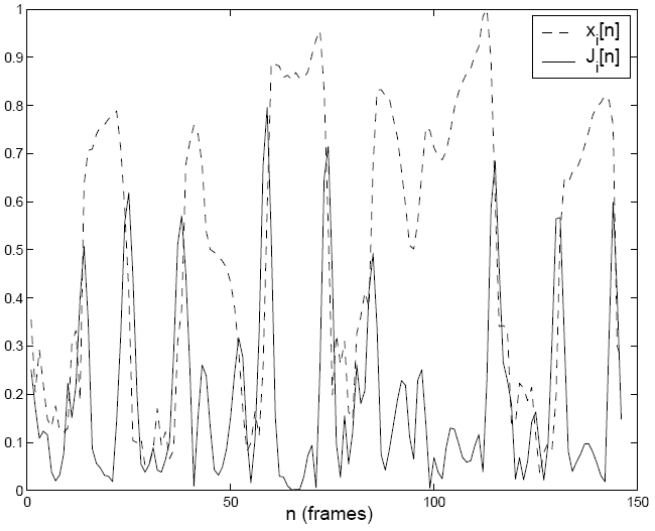


Figure 3.15.: A sequence $x_i(n)$ and the corresponding jump function $J_i(n)$, computed with a span of ± 5 frames (from [3]).

Next, the problem of non-simultaneous jump occurances wthin each of the M sequences is to be solved. A fitting algorithm places the segmentation boundary in the center of a cluster of quasi-simultanoues jumps. The output of this procedure, which is in detail described in [3], is an *accumulation function* $acc[n]$ containing defined peaks which are easy to detect. An example is depicted in figure 3.16.

The algorithm is regulated by three parameters; a, b, and c. The role of a as being the $\pm$span of the jump detection function and thus setting the height of the peaks has already been addressed (cp. equation 3.12). The parameter b serves as a threshold for the peak heighths; if exceeded, the corresponding peak is taken into consideration for the accumulation function. Finally, c adjusts the width of the "jump clusters" in which the algorithm searches for a center.

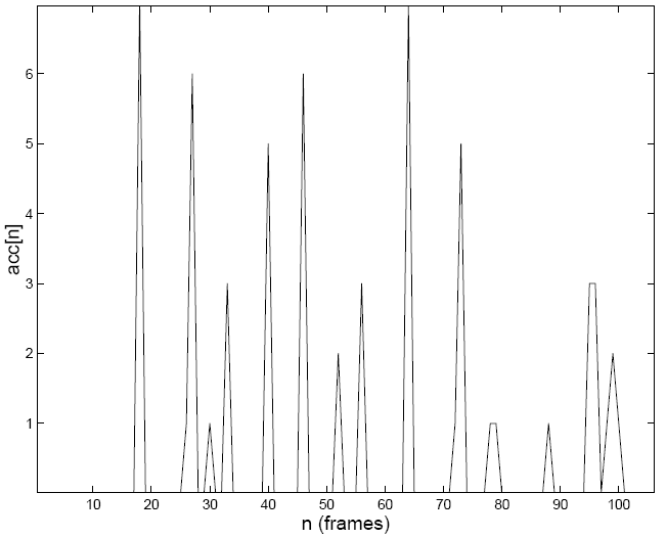


Figure 3.16.: A typical accumulation function (from [3]).

Performance

The authors evaluated their algorithm on the DARPA-TIMIT database[7] and report results of $\sim 74\%$ correctly detected boundaries when allowing no *over-segmentation*; i.e., the number of detected points must not exceed the number of true phoneme boundaries. The term "correctly detected" includes an allowed deviation of $\pm 20ms$ in this case.

Details on the usage within this thesis can be found in section 4.1.3.

3.4.2. Detection using Hidden Markov Models (HMMs)

The HTK Toolkit

Since the ATCOSIM database does not provide labelled data, but transcriptions of the (non-prompted) utterances, an additional phoneme boundary detection method is used. It employs the *HTK Toolkit* developed at the Engineering Department of Cambridge University, which forms a very flexible framework for building and manipulating HMMs. HTK is freely available as a set of library modules and tools in C and can be downloaded from the HTK website[8].

An excellent tutorial on Hidden Markov Models can be found in [38]. For detailed information on the HTK toolkit, the interested reader is recommended to study the HTK book [51]. This section deals just with the parameters and settings used.

[7] The DARPA database for continuous speech recognition is described in [37].

[8] `http://htk.eng.cam.ac.uk`

Parameters and Settings

The task to be performed is called *forced alignment*, i.e., a phoneme label is to be aligned to a determined position in the speech signal. This requires the sequence of words to be available as well as a lexicon file containing phonectic transcriptions for each single word.

Features used for analysis include the first 13 MFCCs, where the zeroth coefficient serves as a measure for the energy contained in the signal, while the rest approximates the spectral shape. In addition, the *delta* and *acceleration coefficients* are taken into consideration, which are the first- and second-order differences of coefficients of consecutive frames. Employing these widely used settings, we end up with a 39-dimensional feature vector extracted from each signal frame.

Phonemes are modeled by three-state HMMs (one state for the beginning, one for the middle, and one state for the end), while for *silence* and *short pause*, two states are sufficient.

3.5. Feature Evaluation and Classification

3.5.1. Numerical Feature Evaluation

Having computed the feature characteristics listed in section 3.3, two alternative statistical tests are performed, which are briefly overviewed in the following paragraphs. The test result is in both cases a feature ranking list for a selected combination of emotional classes. Out of the 10 top-ranked feature characteristics, the best performing combination is determined iteratively.

Criterion 1: Fisher's Ratio

Fisher's Ratio is the ratio between inter-class scatter (distance of mean values) and intra-class scatter (combination of variances) and is given by [4]

$$FR(n) = \frac{(\mu_{C_1}(n) - \mu_{C_2}(n))^2}{\sigma^2_{C_1}(n) + \sigma^2_{C_2}(n)} \; . \tag{3.13}$$

It reflects the discriminative power of feature n between the two classes C_1 and C_2.

Criterion 2: Area Under ROC Curve

A *Receiver Operating Characteristics* (ROC) graph is a way to visualize the performance of a binary classifier by plotting the *true positives* rate over the *false positives* rate for a varying decision threshold (an example is shown in figure 3.17). The area under this ROC curve (in the unit square) is equivalent to the probability that a randomly chosen sample will be classified correctly. It is a non-parametric statistical test and further equivalent to the Wilcoxon test of ranks [20].

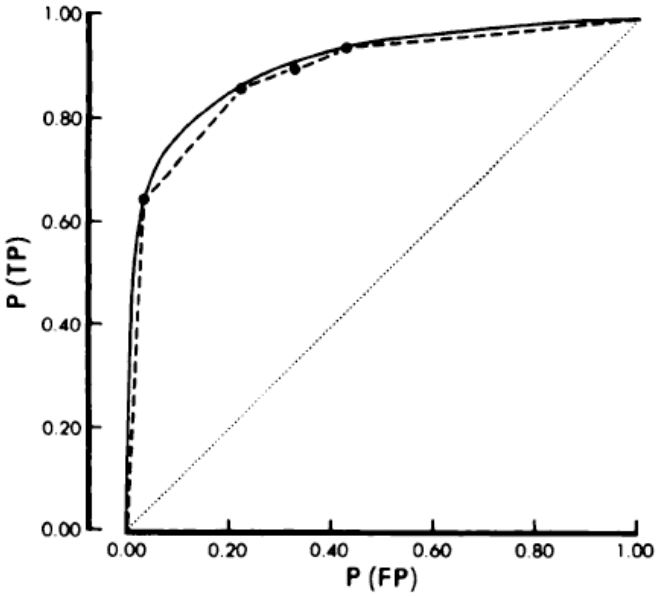


Figure 3.17.: A reciever operating characteristics curve (from [20]).

3.5.2. Graphical Feature Evaluation

To be able to comprehend and to verify the results achieved by numerical evaluation (see 3.5.1), the MATLAB implementation also provides a graphical user interface that visualizes the feature space in up to 3 dimensions for selected database and emotional classes. A screenshot is shown in figure 3.18.

3.5.3. Dimension Reduction

Principal Component Analysis (PCA)

Principal Component Analysis, also known as *Karhunen-Loève-Transformation*, is a method to reduce the dimensionality of a feature space. It works based on the hypothesis that those directions showing the greatest variance contain the most information. (It must be carefully checked if this is true, especially for bigger data sets!)

Mathematically speaking, PCA is an orthogonal linear transformation mapping the data to a new coordinate system such that the direction of the greatest variance is projected onto the first coordinate, the second greatest variance onto the second coordinate and so on. An example for three dimensions is depicted in figure 3.19 on page 33.

A lower-dimensional representation of the data can be obtained by ignoring a specific number of higher-order dimensions, depending on the amount of variance to be maintained. PCA is an unsupervised technique and as such does not include label information of the data.

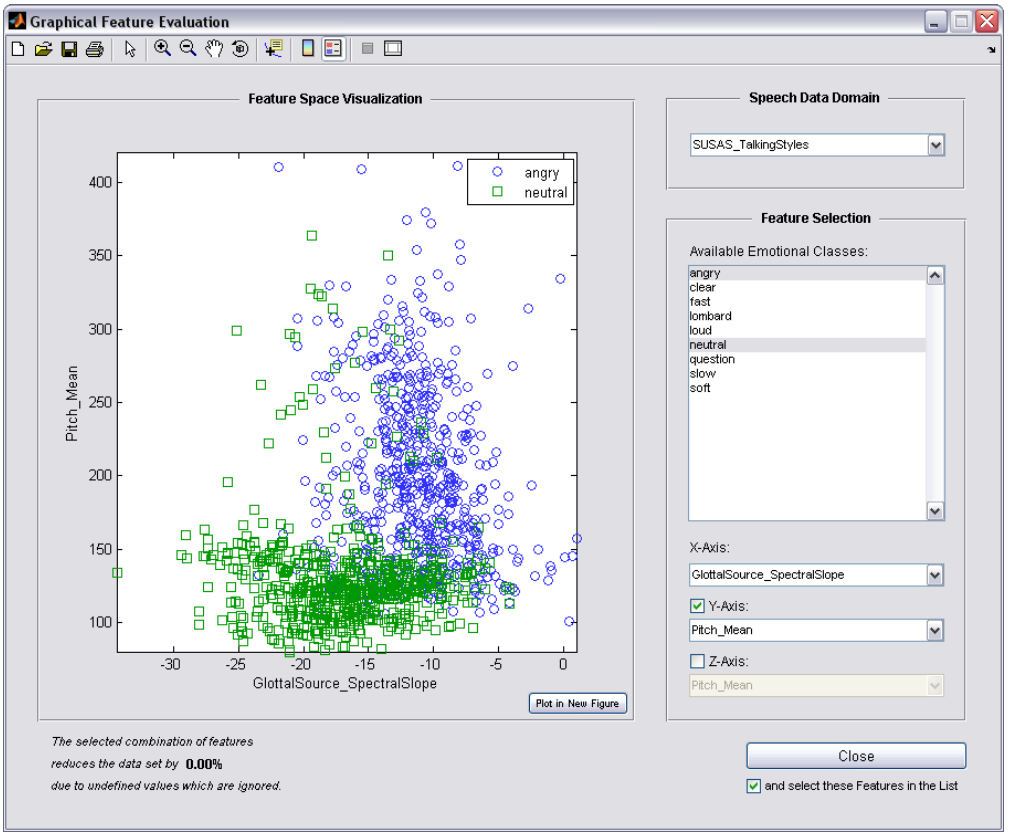


Figure 3.18.: Feature space visualization as a part of the MATLAB tool.

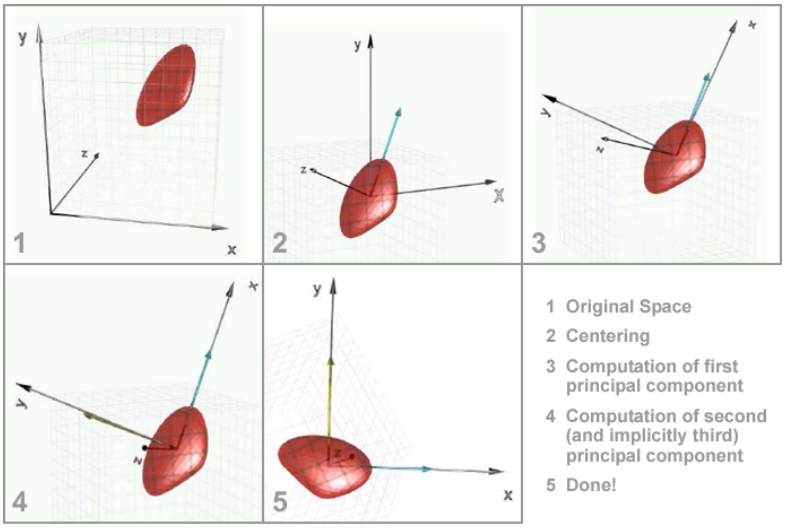


Figure 3.19.: Consecutive steps of PCA (from [43]).

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is related to Principal Component Analysis in that both look for linear combinations of variables which best explain the data. But contrary to PCA, LDA is a supervised technique incorporating class affiliation and tries to yield the largest mean differences between classes. Again, the hypothesis that the mean is an appropriate discriminating factor has to be verified.

Figure 3.20 sketches the two extreme cases, where either one of the discussed methods fails although the data are obviously well-separated.

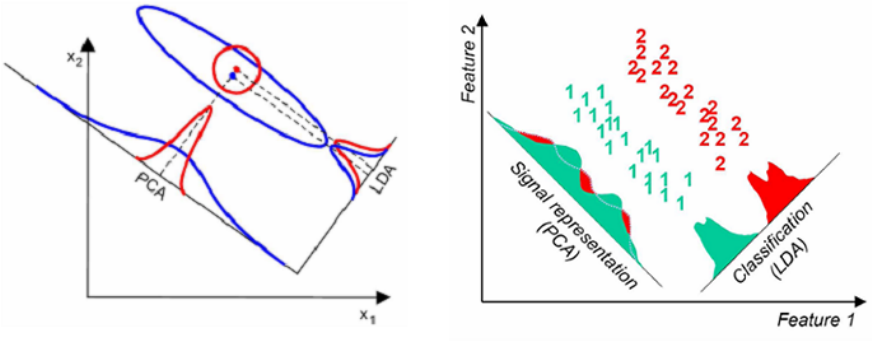


Figure 3.20.: PCA vs. LDA: two extreme cases (from [32]).

A comparison of PCA and LDA concerning performance on databases of different sizes can be found in [34].

3.5.4. Classification

k-Nearest Neighbours Algorithm

In order to set a baseline of performance, the (rather simple) *k-nearest neighbours* algorithm is chosen for classification. The basic principle is as follows: an object to classify is assigned to the most frequent class among the k nearest training samples.

An example is shown in figure 3.21 on page 35: the blue circle is our unknown object and to be assigned either to the "green triangles" or the "orange squares" class. For $k = 3$, the decision will be "orange squares"; for $k = 5$, it will be "green triangles".

n-Fold Cross Validation

The scenario is as follows: emotional classes are known for the data samples and the aim is to estimate how accurately our feature set will perform on an unknown data set (this is called *supervised learning*). To do so, the data set is partitioned into n segments, from which $n - 1$ are used as training data for the classification algorithm, while 1 partition serves as "unknown" data.

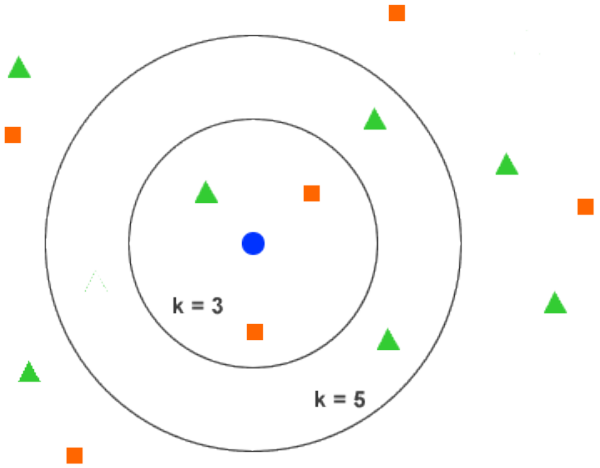


Figure 3.21.: The basic principle of k-nearest neighbours classification.

To ensure that the result (in terms of the correct classification rate) does not depend on the sample distribution in the training data set, this process is repeated n times. Figure 3.22 sketches how n-fold cross validation is performed in a loop, with each of the n data set partitions used exactly once as the validation set.

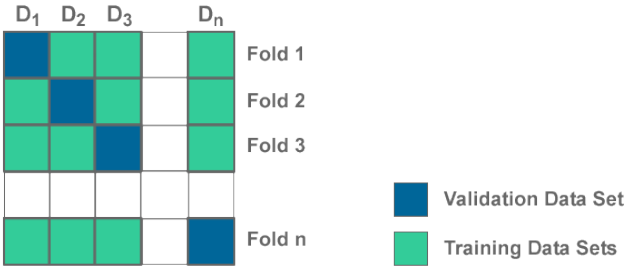


Figure 3.22.: The basic principle of n-fold cross validation.

Chapter 4.

Implementation

This chapter takes a closer look on how the presented methods are realized. As already depicted in figure 3.1 (on page 10), the analysis framework comprises two main stages: feature extraction and feature evaluation.

4.1. Feature Extraction

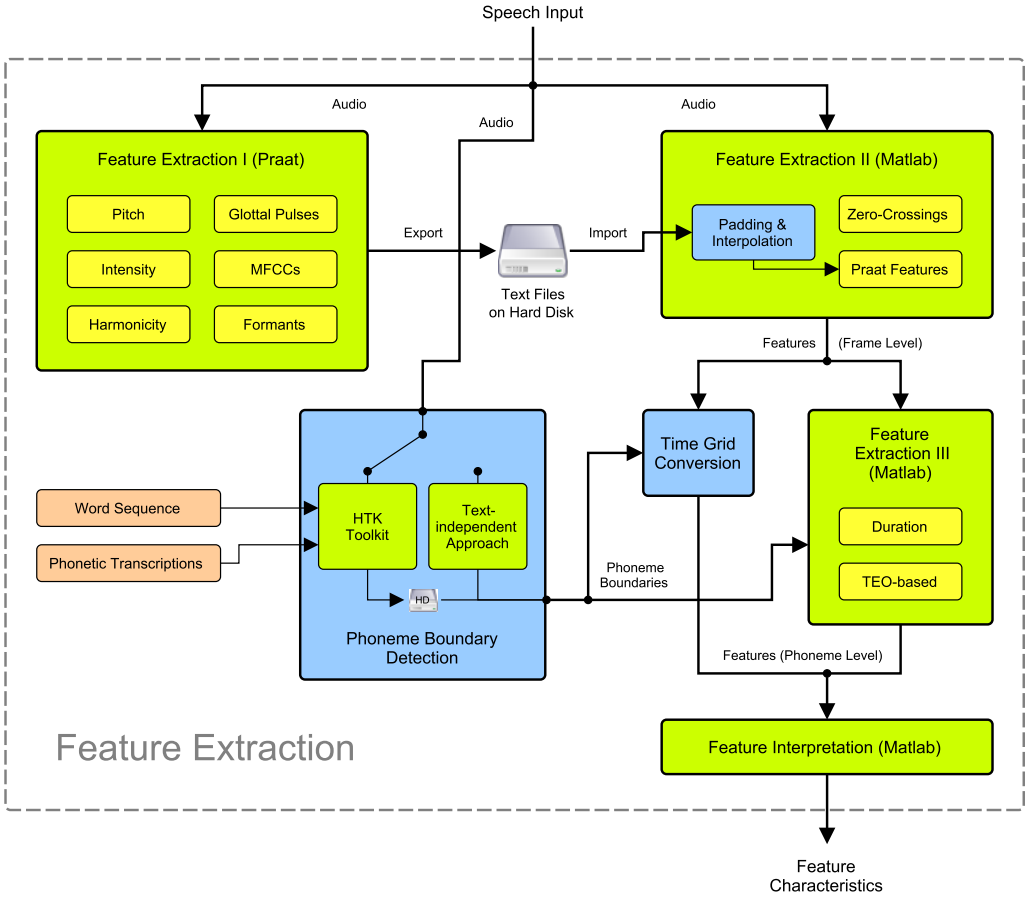


Figure 4.1.: Flow chart of the feature extraction stage.

4.1.1. Overview

To understand the quite complex analysis workflow, have a look at the flow chart above. In the current implementation, three different tools are involved into the analysis process; namely Praat, HTK, and Matlab – which does not pose any problem, since real-time compatibility has never been part of the requirements.

All features are extracted for a series of overlapping frames of varying length depending on the respective feature; but with a constant frame rate (cp. 4.3.1).

The diagram in figure 4.1 can be read from top to bottom, which traces the chronology from the PCM waveform to a set of feature characteristics. At the same time, it can be read from left to right, indicating the sequence of computational steps:

- Extraction of basic features is implemented in Praat by means of an analysis script for each of the three databases. Individual scripts are necessary due to different speech file formats and directory structures within the databases.
- If phoneme boundaries are to be detected by HTK, this task has also to be accomplished before running the Matlab analysis functions. The text-independent approach (cp. 3.4.1), however, is implemented in Matlab.
- Matlab analysis depends on data delivered by the previous items in terms of text files stored on hard disk.

The advantage of this method is the independence of computational steps, which facilitates testing and de-bugging.

4.1.2. Basic Features (Praat)

Praat (Dutch for "talk") is an open-source computer program for speech analysis and synthesis [7] developed by Paul Boersma and David Weenink from the Institute of Phonetic Sciences at the University of Amsterdam in the Netherlands. It is freely available from the internet[1] and is used by many researchers in the field of speech and communication sciences.

Pitch, intensity, F1/F2 frequency and bandwidth, harmonicity, MFCCs and glottal spectral features are extracted using a script which writes all extracted information to text files (one file per speech file and feature). An example can be found in appendix A.1; parameters and settings used are discussed in section 3.3 for each feature, respectively.

4.1.3. Further Feature Extraction (Matlab)

The main part of the analysis stage is implemented in various Matlab functions, all being controlled by *SpeechAnalysis.m*. The following pseudo-code sketches the main issues:

[1] `http://www.fon.hum.uva.nl/praat/`

```
% SpeechAnalysis.m

determine speech data locations
open analysis settings GUI

for 1 to (number of selected domains)
    for 1 to (number of files in current domain)
        read in speech file
        read metadata file
        extract features on frame level
        perform phoneme boundary detection
        (or use HTK results instead)
        extract features on phoneme level
        compute feature statistics
        store results
    end
    save results of current domain
end
```

Feature Extraction on Frame Level

Feature extraction on *frame level*, i.e., based on frame-wise signal analysis, includes Praat text file import and the computation of the zero-crossing rate (cp. 3.3.9). Results are stored in the MATLAB structure `FEATURES.FrameLevel` as depicted in figure 4.2.

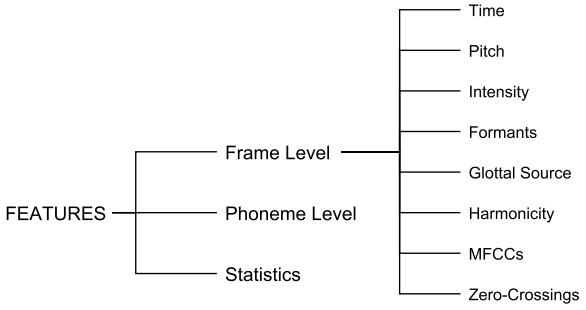


Figure 4.2.: MATLAB structure containing features extracted on frame level.

Phoneme Boundary Detection

Unfortunately, the quality of label data provided with the SUSAS database leaves a lot to be desired (see section 4.4.1). This is why phoneme boundary detection performance could be reliably evaluated on Emo-DB labels solely. (As already mentioned, ATCOSIM does not provide any label files at all.)

The implementation of the text-independent approach has been tested using those parameter settings proposed in the paper to prevent over-segmentation (cp. 3.4.1). An example is depicted in figure 4.3.

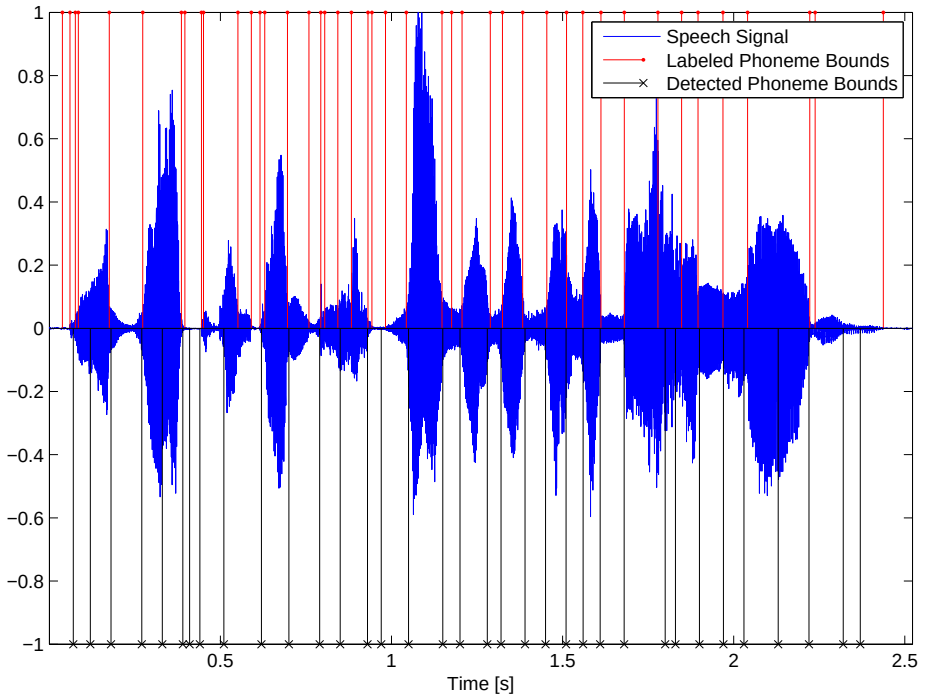


Figure 4.3.: Phoneme boundary detection using the text-independent approach.

Surprisingly, phoneme segmentation using the text-independent approach performs nearly equally well as the HMM-based detection method, as shown in table 4.1. Using generous values for $\Delta_{allowed}$, the text-independent method even outperforms the HTK toolkit.

$\pm\Delta_{allowed}$	20 ms	30 ms	40 ms
HTK Toolkit	71.17%	80.83%	86.15%
Text-independent approach	63.78%	80.93%	90.78%

Table 4.1.: Correct detection rates (CDR) of correctly identified phoneme boundaries within specified uncertainty intervals Δ for both detection methods.

The *correct detection rates* (CDR) are computed by

$$CDR = 100 \cdot \frac{TP}{TP + FP + FN} \quad [\%] \,, \tag{4.1}$$

where TP indicates the number of *true positives* (correct detections), FP stands for the number of *false positives* (incorrect detections), and FN represents the number of *false negatives* (missed detections).

Aversano and his colleagues [3] calculate the same quantity by

$$CDR_{Av} = 100 \cdot \frac{\text{\# correctly detected segmentation points}}{\text{\# “true” segmentation points}} = 100 \cdot \frac{TP}{TP + FN} \quad [\%] \,. \tag{4.2}$$

Using their formula, the algorithm as implemented would yield a CDR of 77.52% for a Δ of $\pm 20ms$ and thus outperform their best result obtained, which is reported to equal 73.58% (13193 correctly detected phoneme bounds out of 17930 true phoneme bounds over the whole DARPA-TIMIT database)[2].

An alternative idea for improving the percentage of correct detections would be to tune the parameters in such a way that a significant amount of over-segmentation is produced, before subsequent discontinuity analysis of extracted low-level features verifies or weakens the case made that each of the outcoming points of time represent a phoneme boundary. In other words, a post-processing stage would be introduced that removes the oversupplied amount of segmentation points by means of discontinuity detection in those features which are extracted anyway.

Feature Extraction on Phoneme Level

After having determined phoneme classes and their temporal boundaries using the HTK toolkit [51], the phoneme durations are calculated (cp. 3.3.3). This is done separately for each class; i.e., vowels, semivowels, consonants, and diphthongs. Additionally, the *Critical Band Based TEO Autocorrelation Envelope* (cp. 3.3.6) is computed on the basis of this phoneme time grid, taking vowel parts as input.

Frame-based features are averaged within phoneme bounds to end up with consistent analysis results. Features derived from the glottal source are extracted once per utterance and thus require no further processing. The structure `FEATURES.PhonemeLevel` is organized as shown in figure 4.4.

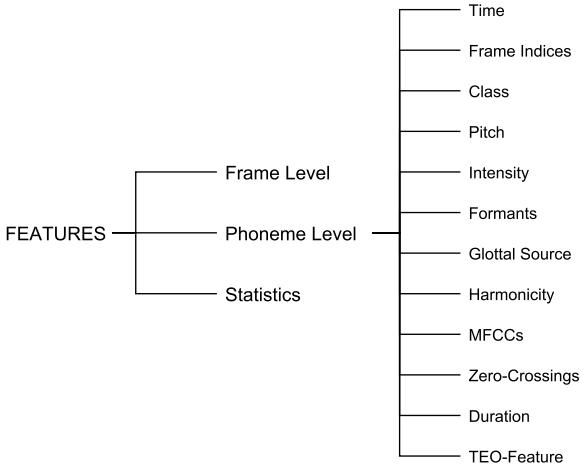


Figure 4.4.: MATLAB structure containing features extracted on phoneme level.

[2]The Emo-DB database shows a total of 26909 labeled phoneme bounds, of which 20870 were detected correctly. The corresponding ratio equals 77.56%, which is approximately the same as 77.52% taking roundoff errors into account – q.e.d.

Feature Interpretation and Statistics

Having extracted all features on basis of the phoneme time grid, the corresponding feature characteristics are calculated as described in section 3.3. For quasi-historical reasons, the MATLAB structure is entitled “statistics”, which does not describe all of its content correctly, however.

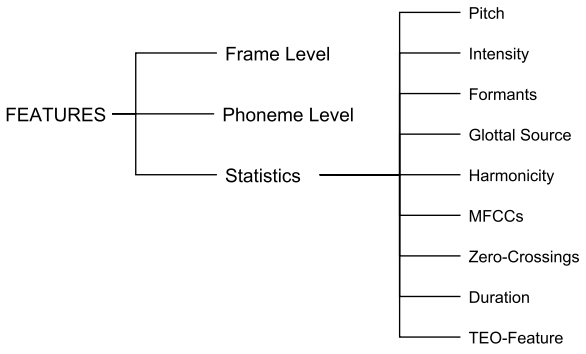


Figure 4.5.: MATLAB structure containing feature characteristics.

4.2. Feature Evaluation

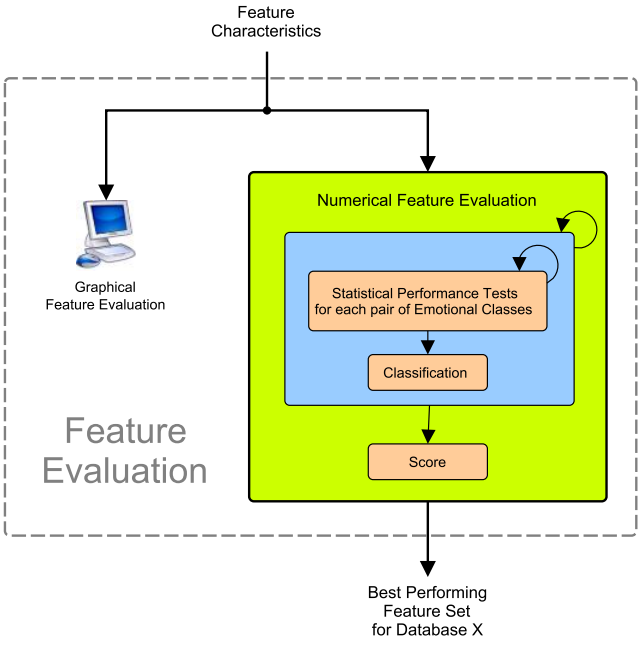


Figure 4.6.: Flow chart of the feature evaluation stage.

The feature evaluation process, as depicted in figure 4.6, is controlled by *Classification.m*, which comprises the following steps:

```
% Classification.m

import analysis results
re-organize data structure
open classification settings GUI
determine "groups" for cross-validation

for 1 to (number of groups)
    divide data set into "training" and "test"
    train KNN classifier with training data
    perform KNN classification with test data
    evaluate classification results
end
average over classification results
```

4.2.1. A Note on Iterative Feature Evaluation

It is to note that an implicit constraint was made by iteratively decreasing the number of features from 10 to 1 during feature evalution (cp. chapter 5): the hypothesis that a combination of individually best performing features at highest ranks always outperforms a combination of lower-ranked features is not necessarily true. It might be the case, e.g., that a feature set [#2, #4, #5] results in a higher CCR than [#1, #2, #3].

This task could be accomplished more accurately by performing *sequential backward selection* [4], i.e., starting with a full set of features and calculating Fisher's ratio for this combinaton of features ($\mu \rightarrow \underline{\mu}$, $\sigma^2 \rightarrow \Sigma$). In the next steps, one feature is eliminated and the ratio is computed for the $n-1$ remaining features. This is done n times in a loop, until the minimum ratio is found and the corresponding feature is discarded, before the whole thing starts again, now from $n-1$ features.

This approach is not subject to the above-mentioned constraint, but still suffers from nesting: once a feature has been discarded, it can not be re-considered.

4.2.2. Classification

In the k-nearest neighbours implementation, distances are determined by computing the Euclidean distance between two points; the factor k is approximated by $k = \sqrt{N}$ (with N being the number of feature vectors in the validation data set).

For n-fold cross validation, it is ensured that the classes are equally represented in the samples for both the training and the validation data set. The number of folds is set to 10.

4.3. Other Issues

4.3.1. Frame Rate

The global analysis frame rate is set to $10ms$, which corresponds to the shortest phoneme duration observed in the data.

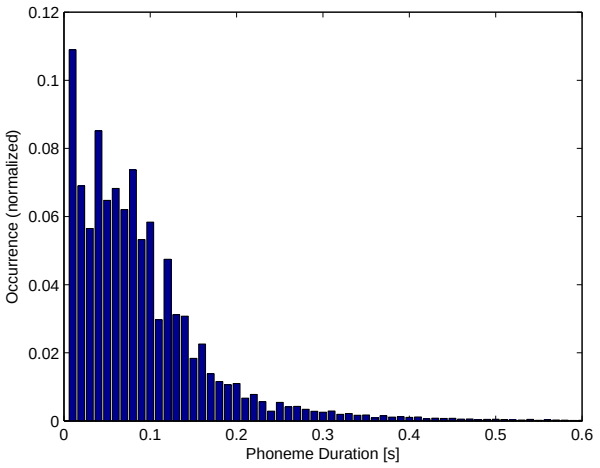


Figure 4.7.: Normalized histogram over phoneme durations (SUSAS and Emo-DB) from 10 to $600ms$.

4.3.2. ATCOSIM workload calibration

The ATCOSIM corpus has not been designed for the purpose of workload-induced stress recognition, so that assignments of utterances to a specific emotional class are missing. There are two possible ways of approximating the instantaneous workload at a given point in time:

- Determining the current taskload by counting the "utterances per minute" or
- Assuming that the operator's screen is initially blank and the workload level rises by-and-by, reaching its climax at the end of the recording session.

Furthermore, the latter accounts for the fact that fatigue grows the longer a demanding task has to be performed; which introduces an additional stressor.

We decided to group the utterances of each session into three categories; namely *low* workload ($0 \ldots 10min.$), *medium* workload ($10 \ldots 40min.$), and *high* ($40 + min.$) workload. Thus, we obtain appropriate counterparts to SUSAS' computer response task domains, which are categorized into the same classes.

4.4. Occured Problems

4.4.1. Bad SUSAS Label Files

During HTK performance analysis, it was found that many of the SUSAS label files are somewhat strange in content. As an example, the SUSAS labfile for "break" (speaker with Boston accent, emotional class angry) is shown, which looks as follows:

```
0 260 sil
260 460 b
460 940 r
940 2500 ey
2500 3660 k
3660 4096 sil
```

The HTK labfile for the same speech file has the following content:

```
seperator ;
nfields 1
#
  0.208 26  sil ; score -70.372986 ;
  0.242 26  b_cl ; score -68.599457 ;
  0.3   26  b ; score -45.037262 ;
  0.362 26  r ; score -32.933506 ;
  0.502 26  ey ; score -35.310688 ;
  0.696 26  k_cl ; score -75.229362 ;
  0.95  26  k ; score -67.300201 ;
  0.954 26  sil ; score -74.975967 ;
```

Dividing the sample-based values in the SUSAS labfile by f_s should yield the phoneme boundaries in seconds and approximately match those bounds detected by the HMM-based algorithm – which is not the case at all. To visualize the difference, SUSAS labels have been scaled in such a way that the last entry matches the last label defined by HTK (the first entry is zero in both cases). The result is depicted in figure 4.8, showing the PCM waveform and the phoneme boundaries extracted using HTK and as provided with the SUSAS database.

4.4.2. Reduced Number of ATCOSIM Files

Due to the fact that we deal with non-prompted speech, some of the ATCOSIM utterances contain vocabulary which is not part of the standardized English lexicon file that is used for HTK analysis:

- It seems to be common practice that operators say hallo and goodbye to the pilots in their national language.
- The respective airline company names are, of course, not part of the lexicon. Besides, names are not pronounced consistently; this may be due to the fact that the operators have German, Swiss German or Swiss French native tongue.

For this reason, only about 5800 utterances out of 10000 are taken for analysis.

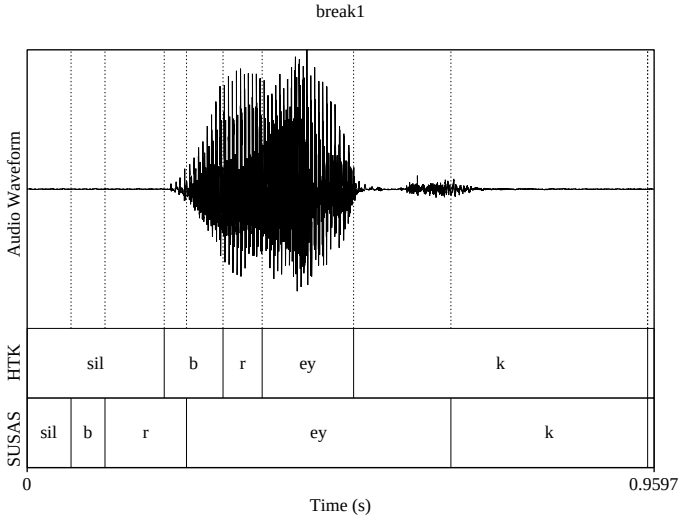
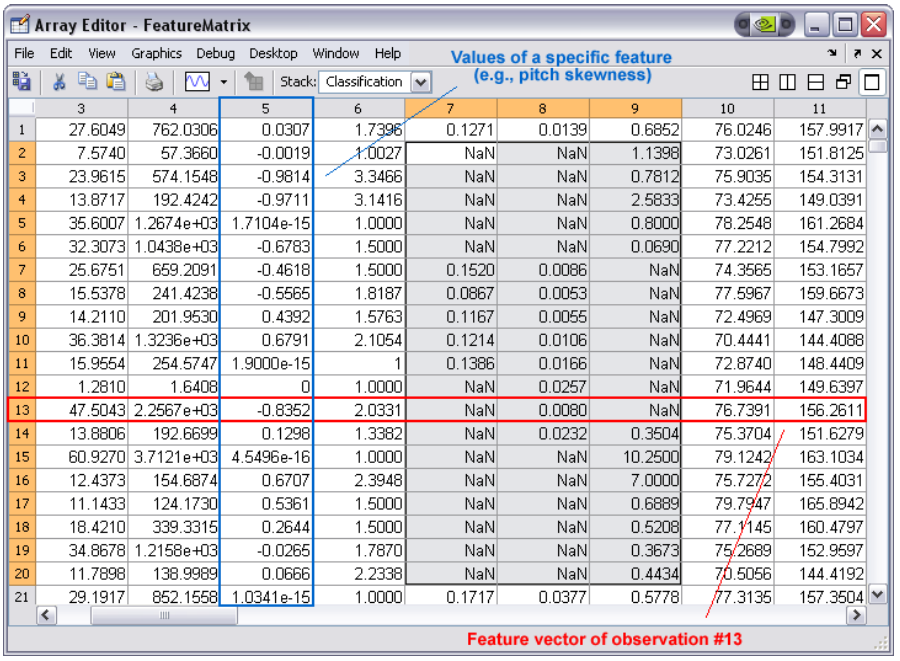


Figure 4.8.: Example of bad SUSAS label file.

4.4.3. Invalid Values in Feature Matrix

For classification, the feature vectors for the two selected emotional classes are "stacked" in a *feature matrix*, where each row corresponds to a single observation and each column contains all values of a specific feature characteristic over all observations (as depicted in figure 4.9).



	3	4	5	6	7	8	9	10	11
1	27.6049	762.0306	0.0307	1.7396	0.1271	0.0139	0.6852	76.0246	157.9917
2	7.5740	57.3660	-0.0019	1.0027	NaN	NaN	1.1398	73.0261	151.8125
3	23.9615	574.1548	-0.9814	3.3466	NaN	NaN	0.7812	75.9035	154.3131
4	13.8717	192.4242	-0.9711	3.1416	NaN	NaN	2.5833	73.4255	149.0391
5	35.6007	1.2674e+03	1.7104e-15	1.0000	NaN	NaN	0.8000	78.2548	161.2684
6	32.3073	1.0438e+03	-0.6783	1.5000	NaN	NaN	0.0690	77.2212	154.7992
7	25.6751	659.2091	-0.4618	1.5000	0.1520	0.0086	NaN	74.3565	153.1657
8	15.5378	241.4238	-0.5565	1.8187	0.0867	0.0053	NaN	77.5967	159.6673
9	14.2110	201.9530	0.4392	1.5763	0.1167	0.0055	NaN	72.4969	147.3009
10	36.3814	1.3236e+03	0.6791	2.1054	0.1214	0.0106	NaN	70.4441	144.4088
11	15.9554	254.5747	1.9000e-15	1	0.1386	0.0166	NaN	72.8740	148.4409
12	1.2810	1.6408	0	1.0000	NaN	0.0257	NaN	71.9644	149.6397
13	47.5043	2.2567e+03	-0.8352	2.0331	NaN	0.0080	NaN	76.7391	156.2611
14	13.8806	192.6699	0.1298	1.3382	NaN	0.0232	0.3504	75.3704	151.6279
15	60.9270	3.7121e+03	4.5496e-16	1.0000	NaN	NaN	10.2500	79.1242	163.1034
16	12.4373	154.6874	0.6707	2.3948	NaN	NaN	7.0000	75.7272	155.4031
17	11.1433	124.1730	0.5361	1.5000	NaN	NaN	0.6889	79.7947	165.8942
18	18.4210	339.3315	0.2644	1.5000	NaN	NaN	0.5208	77.1145	160.4797
19	34.8678	1.2158e+03	-0.0265	1.7870	NaN	NaN	0.3673	75.2689	152.9597
20	11.7898	138.9989	0.0666	2.2338	NaN	NaN	0.4434	70.5056	144.4192
21	29.1917	852.1558	1.0341e-15	1.0000	0.1717	0.0377	0.5778	77.3135	157.3504

Figure 4.9.: Extract of a feature matrix. The gray area indicates mutually exclusive feature characteristics.

Some low-level features may contain special non-numeric values that indicate a frame where this feature is not defined; e.g., pitch information in unvoiced frames. MATLAB uses the identifier `NaN` (Not a Number) for this purpose. When computing feature characteristics, these values

are neglected, of course. But it may still happen that a whole feature characteristic for a single observation is characterized with `NaN`, as it is the case with features extracted for individual phoneme classes.

Since this special value is no number by definition, it is not possible to calculate a distance to it, which is a prerequisite for k-nearest neighbours classification. As a consequence of this, an observation is discarded whenever it contains a single `NaN`. Unfortunately, this happens extremely often when incorporating characteristics derived from duration or intensity, since especially the phoneme class of semivowels occurs rather infrequently. Beyond, the short single-word utterances from SUSAS do never contain a vowel, a consonant, a semivowel, and a diphthong at the same time, which makes duration and intensity characteristics *mutually exclusive*.

This is why the idea of phoneme-class dependent features is revised with hindsight and duration as well as intensity are now just represented in terms of mean and variance – with the exception of the consonant-to-vowel duration ration (CVDR), which is still quite well-represented.

Chapter 5.

Results

5.1. Experiment 1: Talking Styles

The first experiment compares optimal performing feature sets for classification of the talking styles *angry* and *neutral*, employing acted emotional speech from the SUSAS and Emo-DB databases.

Feature evaluation results are summarized in tables 5.1 and 5.3, respectively. All feature characteristics are fairly normal distributed; with exception of the variances and pitch kurtosis, unsurprisingly. This is why evaluation using Fisher's Ratio does not include characteristics derived by calculating the variance at higher ranks[1].

5.1.1. Individual Results for SUSAS

Table 5.1 shows the "Top 10" of the ranking lists produced by the two statistical tests described in section 3.5.1.

In this table, columns 3 and 5 contain the *relative score* for the particular features, i.e., the normalized value of the respective criterion as an output of the statistical test applied:

$$rel.\ score(k) = \frac{test\ score(k)}{max(test\ scores)} \tag{5.1}$$

Since these values have been calculated individually, the (multivariate) performance of a combination of features is still to be determined. Table 5.2 lists the classification results for the first n features ($n = 1 \ldots 10$) using the k-nearest neighbours method in a 10-fold cross validation process. The parameter k is always set to $round(\sqrt{N})$, with N being the number of valid feature vectors which form the input to the classification algorithm.

Although the best performing feature set is the "full" 10-dimensional set, reducing the number of features to 3 (or even 2, taking the features chosen by Fisher's Ratio) would not lower the CCR significantly.

[1] Variance and kurtosis only reach positive values and thus are no two-sided functions. It is to note that these curves could theoretically be projected in such a way that a their shape resembles a Gaussian normal distribution

	Fisher's Ratio		**Area under ROC**	
Rank	Feature	Rel. Score	Feature	Rel. Score
1	Spectral slope	1.00	Mean Pitch	1.00
2	Mean Pitch	0.87	Pitch variance	0.91
3	1. MFCC	0.71	Spectral slope	0.87
4	TEO Mean Diff.	0.59	1. MFCC	0.81
5	Spectral Mean	0.54	TEO Var. of Mean	0.78
6	TEO Mean of Mean	0.50	TEO Mean Diff.	0.75
7	MFCC Mean Var.	0.42	Spectral Mean	0.73
8	TEO Var. of Mean	0.38	TEO Mean of Mean	0.71
9	Mean Intensity	0.27	MFCC Mean Var.	0.68
10	3. MFCC	0.23	3. MFCC	0.63

Table 5.1.: The 10 best performing features (SUSAS Talking Styles).

No.	**Fisher's Ratio**		**Area under ROC**	
Features	CCR	No. FVs	CCR	No. FVs
10	86.63%	1092	**88.09%**	1092
9	86.72%	1092	86.73%	1092
8	87.36%	1092	86.72%	1092
7	86.52%	1092	85.71%	1092
6	86.81%	1092	83.45%	1092
5	85.54%	1092	82.50%	1246
4	83.97%	1092	82.85%	1246
3	82.45%	1248	84.05%	1248
2	84.22%	1248	80.60%	1248
1	73.80%	1248	80.06%	1248

Table 5.2.: Percentages of correct classification rate (CCR) for varying numbers of features (SUSAS Talking Styles).

5.1.2. Individual Results for Emo-DB

An optimum feature set for the Emo-DB data is determined in the same manner as described above. It is notable that both statistical tests deliver identical results for the top 7 features and that 9 out of 10 features are the same. These results can be found in tables 5.3 and 5.4 on page 49.

The fact that scores of more than 94% correct classification rate are obtained even when using one single feature characteristic indicate that, in this case, only little discriminative power is contained in the remaining features. Using more than the three top-ranked features definetely weakens the result.

	Fisher's Ratio		**Area under ROC**	
Rank	Feature	Rel. Score	Feature	Rel. Score
1	Spectral slope	1.00	Spectral slope	1.00
2	2. MFCC	0.82	2. MFCC	0.99
3	1. MFCC	0.68	1. MFCC	0.97
4	MFCC Mean Var.	0.64	MFCC Mean Var.	0.95
5	Mean Zero-Crossings	0.48	Mean Zero-Crossings	0.92
6	3. MFCC	0.41	3. MFCC	0.89
7	Mean pitch	0.40	Mean Pitch	0.88
8	TEO Mean of Mean	0.27	Mean Intensity	0.81
9	4. MFCC	0.20	TEO Mean of Mean	0.78
10	TEO Weighted Score	0.18	4. MFCC	0.71

Table 5.3.: The 10 best performing features (Emo-DB).

No.	**Fisher's Ratio**		**Area under ROC**	
Features	CCR	No. FVs	CCR	No. FVs
10	88.75%	158	89.29%	158
9	88.57%	158	87.95%	158
8	89.29%	158	88.66%	158
7	87.95%	158	87.95%	158
6	87.41%	158	87.41%	158
5	88.04%	158	88.04%	158
4	96.79%	158	96.79%	158
3	**98.10%**	158	**98.10%**	158
2	97.50%	158	97.50%	158
1	94.29%	158	94.29%	158

Table 5.4.: Percentages of correct classification rate (CCR) for varying numbers of features (Emo-DB).

5.1.3. Combination of Results

To get an idea of transferability, cross-evaluation is performed in an intermediate step: the best performing feature set for data set A (SUSAS) is applied on data set B (Emo-DB) and vice versa. Results are presented in table 5.5, where $F(\cdot)$ denotes the best performing feature set for the respective data set and the arrow points at the data set on which the classification is performed.

F(A) → A	**F(B) → A**	**F(B) → B**	**F(A) → B**
88.09%	78.80%	98.10%	98.04%

Table 5.5.: Correct classification rates for talking styles angry vs. neutral; individual feature sets.

Finally, the performance of the shared feature set is to be determined. F(A) ∩ F(B) yields that

- the slope of the glottal pulse spectrum and
- the first MFCC

are the features of interest. Results are shown in table 5.6, pointing out that SUSAS data suffers more from the absence of pitch as a feature. It stands to reason that the short length of the SUSAS utterances causes the extracted features to be less steady than those derived from Emo-DB data (cp. chapter 6).

F(A) ∩ F(B) → A	**F(A) ∩ F(B) → B**
75.96%	97.50%

Table 5.6.: Correct classification rates for talking styles angry vs. neutral; shared feature set.

Classification using this set of common top-performing features leads to an equal classification rate for Emo-DB data and to a decrease of 12% for the SUSAS database.

5.2. Experiment 2: Workload Tasks

In the second experiment, the SUSAS dual tracking domain is tested against the ATCOSIM database. Both data sets contain three emotional classes, namely *neutral/low stress*, *medium stress*, and *high stress*. In order to maximize results, the first category is classified against the third one, respectively.

5.2.1. Individual Results for SUSAS

Tables 5.7 and 5.8 on page 51 contain the results for the SUSAS dual tracking domain. All explanatory notes form above hold likewise.

	Fisher's Ratio		**Area under ROC**	
Rank	Feature	Rel. Score	Feature	Rel. Score
1	TEO Weighted Score	1.00	TEO Weighted Score	1.00
2	MFCC Mean Var.	0.65	Spectral Mean	0.80
3	Spectral Mean	0.44	1. MFCC	0.79
4	1. MFCC	0.42	Intensity Var.	0.76
5	3. MFCC	0.41	MFCC Mean Var.	0.75
6	Mean Intensity	0.33	3. MFCC	0.72
7	Spectral Slope	0.31	Mean Intensity	0.67
8	Mean Zero-Crossings	0.29	Pitch Variance	0.67
9	4. MFCC	0.29	Mean Pitch	0.67
10	TEO Mean of Mean	0.29	Mean Zero-Crossings	0.66

Table 5.7.: The 10 best performing features (SUSAS Dual Tracking).

No.	**Fisher's Ratio**		**Area under ROC**	
Features	CCR	No. FVs	CCR	No. FVs
10	60.81%	1398	58.94%	1386
9	59.71%	1398	65.52%	1386
8	62.01%	1398	65.15%	1386
7	**69.95%**	1398	66.68%	1386
6	63.67%	1398	65.22%	1386
5	62.87%	1398	63.42%	1386
4	60.08%	1398	64.21%	1386
3	58.09%	1398	60.45%	1398
2	55.65%	1398	58.23%	1398
1	57.65%	1398	57.65%	1398

Table 5.8.: Percentages of correct classification rate (CCR) for varying numbers of features (SUSAS Dual Tracking).

5.2.2. Individual Results for ATCOSIM

Tables 5.9 and 5.10 show the feature ranking list and the classification results for the ATCOSIM data, repectively.

	Fisher's Ratio		**Area under ROC**	
Rank	Feature	Rel. Score	Feature	Rel. Score
1	Spectral Mean	1.00	Glottal Jitter	1.00
2	Mean Intensity	0.88	Spectral Mean	0.78
3	Glottal Jitter	0.60	Mean Intensity	0.72
4	TEO Var. of Mean	0.42	TEO Var. of Mean	0.64
5	Mean Harmonicity	0.21	Mean Pitch	0.63
6	Zero-Crossings Var.	0.19	Intensity Var.	0.61
7	Glottal Shimmer	0.19	Mean F2 Bandwidth	0.46
8	F2 Bandwidth Var.	0.17	Duration Variance	0.45
9	3. MFCC	0.17	Mean Harmonicity	0.40
10	Mean F2 Frequency	0.16	F2 Bandwidth Var.	0.36

Table 5.9.: The 10 best performing features (ATCOSIM).

No.	**Fisher's Ratio**		**Area under ROC**	
Features	CCR	No. FVs	CCR	No. FVs
10	62.62%	888	57.37%	884
9	63.19%	888	59.72%	884
8	63.29%	888	57.69%	884
7	**63.96%**	888	54.61%	884
6	58.18%	890	55.77%	884
5	57.63%	890	55.27%	890
4	54.85%	890	55.61%	890
3	55.61%	890	55.40%	890
2	56.67%	890	55.84%	890
1	53.50%	890	54.74%	890

Table 5.10.: Percentages of correct classification rate (CCR) for varying numbers of features (ATCOSIM).

As for the SUSAS computer response task, it is a set of seven features that performs best.

5.2.3. Combination of Results

Appplying the best performing feature on the respective opposite data set leaeds to results presented in table 5.11. Here, the SUSAS dual tracking task is referred to as data set "C", while the

ATCOSIM data are denoted with the letter “D”.

F(C) → C	**F(D) → C**	**F(D) → D**	**F(C) → D**
69.95%	61.44%	63.96%	57.73%

Table 5.11.: Correct classification rates for neutral/low stress vs. high stress; individual feature sets.

Finally, the performance of the shared feature set is to be determined. F(C) ∩ F(D) consists of

- the average level of the glottal pulse spectrum and
- the average intensity.

Results can be found in table 5.12.

F(C) ∩ F(D) → C	**F(C) ∩ F(D) → D**
59.45%	57.64%

Table 5.12.: Correct classification rates for neutral/low stress vs. high stress; shared feature set.

5.3. What about the Benchmark?

As mentioned in section 3.1.2, research results by Hansen and his colleagues are taken as reference. Although their evaluation methods differ from the approach presented in this thesis, it is nevertheless interesting to compare results.

In [54], they performed pairwise tests on isolated phonemes using a HMM-based classifier. Amongst others, talking styles neutral and angry from the SUSAS database were tested against each other in the following way: a cross validation loop (cp. 3.5.4) was executed, in which the reference models were trained from 17 tokens each, while models trained from token #18 were taken as input.

Performing a text-dependent test[2], the authors report an average CCR of 95.4% for neutral and 89.8% for angry talking style when employing the TEO-CB-AutoEnv feature alone. MFCCs and pitch yield 94.4% / 89.8% and 96.5% / 82.4%, respectively.

Results degrade slightly when using models trained on different data for testing and validation: the TEO-CB-AutoEnv feature classifies neutral and angry with 90.8% and 93.0%, respectively; using MFCCs as features, 59.6% and 80.0% are reached. Pitch yields 92.6% CCR for neutral and 83.0% for angry.

[2]This is also known as an “in-vocabulary” test; meaning that the models for training and validation have been trained on the same data (sub)set.

Chapter 6.

Discussion of Results and Perspective

Regarding acted emotional speech, classification results are more than convincing. A shared feature set has been found which reduces the performance only slightly for continuous speech samples, but after all by 12% for isolated words. Still, it can be concluded that the presented analysis framework is, on principle, transferable to another language without dramatic changes in performance; being aware of the fact that English and German are related languages up to a certain degree (as both are Germanic languages) and things might be different considering, e.g., Asian languages.

Another point regarding performance is that the emotional classes angry and neutral, which are classified against each other, surely reside at opposite positions in the emotional spectrum (when discarding "positive" emotions) and thus are likely to differ in descriptive parameters. The real cause for selecting exactly this pair of emotional classes was, however, the fact that the intersection of emotional classes from the SUSAS and Emo-DB databases is confined to just these two classes.

For workload-induced stress analysis, classification results are significantly lower, but – with approximately 70% for the dual tracking task – still can be referred to as "fair". Intersecting best performing individual feature sets leads to a similar reduction in performance, which is, from a starting position of 64% for ATCOSIM data, no desirable result, as we slightly approach chance level.

As already expected to be a general restriction, quantitative labelled workload-induced stress data leads to diffuse classes for workload regions and therefore complicates classification attempts (cp. 1.3). In addition, the used method for assigning emotional classes to the ATCOSIM data may not have been very accurate, since classification rates are lower than for the dual tracking task; although the data set consists of comparatively long, continuous speech phrases which introduce a certain amount of context.

If the ATCOSIM corpus is still to be used for analysis purposes, it is suggested to combine both approaches mentioned in section 4.3.2 by multiplying the normalized "utterances per minute" with a (downscaled) ramp to take both the instantaneous workload and the growing fa-

tigue through a long-lasting task into account.

With glottal spectral slope and the first MFCC, the shared feature set for acted emotions contains features which both approximate the spectral shape. For workload-induced stress data, the set comprises spectral mean of the glottal excitation and mean intensity, which are also closely related to each other. This accounts for the theory that different stressors leave different traces in the speech signal.

Concerning the talking styles domain, SUSAS data significantly suffers from the absence of pitch as a feature (see table 5.2). One could conclude that, in general, f_0 plays an important role where context is missing; but surprisingly, it is not listed among the top ten features for the dual tracking task (cp. table 5.7), which consists of the same single-word vocabulary. On the other hand, classification of SUSAS' talking styles performs equally well with or without incorporation of intensity-related features.

The fact that the k-nearest neighbours algorithm is a rather simple classification scheme emphasizes the excellent results regarding classification of acted emotional speech. At the same time, it does not allow any excuse for the rather moderate classification rates for workload-induced stress, of course.

Results may be adversely effected by the fact that all data are treated equally, regardless from the speaker and thus ignoring certainly existent bias. Amir and Ron [2] state that "it is difficult to define an objective scale for subjective phenomena" and recommend speaker-dependent analysis. This can be implemented by mean and variance normalization before feature evaluation.

The implementation of the proposed analysis framework is, of course, not real-time capable. But the quintessence of this presented work in terms of a small number of features may be implemented as an online application in *C* code or using a graphical programming language like *Pure Data* (pd), when it comes to design a speech monitoring system as described in section 1.1. Besides, stress level analysis does not necessarily have to be performed completely delay-free; a value which is updated every few seconds would do for sure.

As a consequence of poor analysis results for workload-induced stress, it can be concluded that there is a need for a novel specific database of non-prompted speech produced during the performance of demanding tasks with calibrated taskload levels. A credible indicator of induced workload should be incorporated, which could be realized by means of, e.g., pulse measurements. Leaving the level of discrete vocabulary and isolated words also facilitates the investigation of new potential features for stress classification, including rhythmic and harmonic characteristics.

Appendix A.

Code Examples

A.1. Praat Feature Extraction Script

Due to different speech file formats and directory structures within the databases, an individual script had to be written for each of the three databases used.

The following Praat script performs feature extraction (pitch, intensity, formants, harmonicity, MFCCs and glottal source information) on the Emo-DB database and exports results as text files, which are then imported into Matlab for subsequent analysis.

```
# "QueryWaveFiles_EmoDB_win"
# Written by Johannes Luig, Inst. of Electronic Music & Acoustics (IEM)
# Last revision: 19-Mar-2009


# --- settings
frameLength = 0.025
hopSize = 0.01
fMin = 75
fMax_Pitch = 600

# --- specify directory
directory$ = "D:\DATA\UNI\DA\DATABASES\Emo-DB\wav"

# --- create new sub-directories
system_nocheck mkdir 'directory$'\praat

# --- clear object list
if numberOfSelected() > 0
  select all
  Remove
endif

# --- create file list
Create Strings as file list... AllWaveFiles 'directory$'\*.wav

# --- get number of file list entries
noFiles = Get number of strings
```

```
for jj to noFiles

  # --- select and read jj-th wave file
  select Strings AllWaveFiles
  fileName$ = Get string... jj
  nameLength = length(fileName$)
  shortFileName$ = left$(fileName$, nameLength-4)
  Read from file... 'directory$'\'fileName$'

  # --- extract pitch info, export to text file
  To Pitch... hopSize fMin fMax_Pitch
  Down to PitchTier
  Write to text file... 'directory$'\praat\'shortFileName$'_pitch.txt

  # --- extract intensity info, export to text file
  select Sound 'shortFileName$'
  To Intensity... fMin hopSize yes
  Down to IntensityTier
  Write to text file... 'directory$'\praat\'shortFileName$'_intensity.txt

  # --- extract formants info, export to text file
  select Sound 'shortFileName$'
  To Formant (burg)... hopSize 2 4000 frameLength 50
  Down to FormantGrid
  Write to text file... 'directory$'\praat\'shortFileName$'_formants.txt

  # --- extract harmonicity, export to text file
  select Sound 'shortFileName$'
  To Harmonicity (cc)... hopSize fMin 0.1 1
  Write to text file... 'directory$'\praat\'shortFileName$'_harmonicity.txt

  # --- extract MFCCs, export to text file
  select Sound 'shortFileName$'
  To MFCC... 12 frameLength hopSize 100 100 0
  Write to text file... 'directory$'\praat\'shortFileName$'_mfccs.txt

  # --- extract glottal source info
  select Sound 'shortFileName$'
  To Manipulation... hopSize fMin fMax_Pitch
  Extract pulses
  nP = Get number of points
  if nP > 1
    jitter$ = Get jitter (local)... 0 0 0.0001 0.02 1.3
    plus Sound 'shortFileName$'
    shimmer$ = Get shimmer (local)... 0 0 0.0001 0.02 1.3 1.6
    To Ltas... 4000 50 0.0001 0.02 1.3
    specmean$ = Get mean... 0 0 energy
    specslope$ = Get slope... 0 1000 1000 4000 energy
  else
    jitter$ = "NaN"
    shimmer$ = "NaN"
    specmean$ = "NaN"
    specslope$ = "NaN"
  endif
```

```
  # --- write glottal source info to text file
  jitter$ = "Local Jitter:"+tab$+jitter$+newline$
  shimmer$ = "Local Shimmer:"+tab$+shimmer$+newline$
  specmean$ = "Spectral Mean:"+tab$+specmean$+newline$
  specslope$ = "Spectral Slope:"+tab$+specslope$+newline$
  infofile$ = "'directory$'\praat\'shortFileName$'_glottal.txt"
  jitter$ > 'infofile$'
  shimmer$ >> 'infofile$'
  specmean$ >> 'infofile$'
  specslope$ >> 'infofile$'

  # --- remove objects from list
  select all
  minus Strings AllWaveFiles
  Remove

endfor


# --- delete file name list
select Strings AllWaveFiles
Remove
```

Colophon

All parts of this work, including this documentation and the implementation of the presented analysis framework in terms of MATLAB m-files and Praat scripts, were autonomously written by the author, except where explicitely otherwise stated.

MATLAB is a registered trademark of The MathWorks, Inc. All other product names are trademarks or brand names of their respective owners.

This document was written in LaTeX using the MikTeX distribution and the TeXnicCenter editor. The RTF version has been converted using the free latex2rtf tool.

My diploma thesis was supported in part by Eurocontrol under Research Grant Scheme - Graz, (08-120918-C). Special thanks in this regard to Horst Hering for his helpful remarks.

Did I forget something? Add your comment here:

_____

_____

_____

Bibliography

[1] K Alter, E Rank, SA Kotz, U Toepel, M Besson, A Schirmer, and AD Friederici. Affective encoding in the speech signal and in event-related brain potentials. *Speech Communication*, 40(1-2):61–70, 2003.

[2] N Amir and S Ron. Towards an automatic classification of emotions in speech. *Fifth International Conference on Spoken Language Processing*, Jan 1998.

[3] G Aversano, A Esposito, and M Marinaro. A new text-independent method for phoneme segmentation. *Circuits and Systems*, Jan 2001.

[4] JP Bello. Music information retrieval (lecture notes), chapter 9. `http://www.nyu.edu/classes/bello/MIR_files/9_classification.pdf`, retrieved December 08, 2007.

[5] P Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17:97–110, 1993.

[6] P Boersma and G Kovacic. Spectral characteristics of three styles of croatian folk singing. *The Journal of the Acoustical Society of America*, 119:1805, 2006.

[7] P Boersma and D Weenink. Praat: doing phonetics by computer, retrieved May 05, 2009.

[8] P Boersma and D Weenink. Praat manual (for version 5.1.05), retrieved May 05, 2009.

[9] SE Bou-Ghazale and JHL Hansen. A comparative study of traditional and newly proposed features forrecognition of speech under stress. *IEEE Transactions on speech and audio processing*, 8(4):429–442, 2000.

[10] M Brenner, HH Branscomb, and GE Schwarz. Psychological stress evaluator – two tests of a vocal measure. *Psychophysiology 16*, 1979.

[11] F Burkhardt, A Paeschke, M Rolfes, and W Sendlmeier. A database of german emotional speech. *Ninth European Conference on Speech Communication and ...*, Jan 2005.

[12] F Burkhardt and W Sendlmeier. Verification of acoustical correlates of emotional speech using formant-synthesis. *ISCA Tutorial and Research Workshop (ITRW) on Speech and* ..., Jan 2000.

[13] GJ Clary and JHL Hansen. A novel speech recognizer for keyword spotting. *Proceedings ICSLP-1992*, 1992.

[14] R Cowie and RR Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32, 2003.

[15] S Davis and P Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[16] D de Waard. The measurement of drivers' mental workload. *PhD Thesis*, pages 1–135, Sep 1996.

[17] E Douglas-Cowie, R Cowie, and M Schröder. A new emotion database: Considerations, sources and scope. *ISCA Tutorial and Research Workshop (ITRW) on Speech and* ..., Jan 2000.

[18] R Fernandez and R Picard. Modeling drivers' speech under stress. *Speech Communication*, Jan 2003.

[19] M Hagmüller, E Rank, and G Kubin. Evaluation of the human voice for indications of workload induced stress in the aviation environment. *eurocontrol.be*, 2006.

[20] JA Hanley and BJ McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29, 1982.

[21] J Hansen and S Bou-Ghazale. Getting started with susas: A speech under simulated and actual stress database. *Fifth European Conference on Speech Communication and* ..., Jan 1997.

[22] J Hansen and S Patil. Speech under stress: Analysis, modeling and recognition. *LECTURE NOTES IN COMPUTER SCIENCE*, Jan 2007.

[23] J Hansen, M Rahurkar, E Ruzanski, J Meyerhoff, G Saviolakis, and M Koenig. Robust emotional stressed speech detection using weighted frequency subbands. *IEEE Transactions on speech and audio processing*, 2003.

[24] J Hansen, C Swail, A South, and R Moore. The impact of speech under 'stress' on military speech technology. *published by NATO Research Technology Organization RTO-TR-10* ..., Jan 2000.

[25] JHL Hansen. Evaluation of acoustic correlates of speech under stress for robust speech recognition. *Bioengineering Conference, 1989., Proceedings of the 1989 Fifteenth Annual Northeast*, pages 31–32, 1989.

[26] JHL Hansen. Morphological constrained feature enhancement with adaptive cepstral compensation (mce-acc) for speech recognition in noise and lombard effect. *IEEE Transactions on speech and audio processing*, 2(4):598–614, 1994.

[27] K Hofbauer, S Petrik, and H Hering. The atcosim corpus of non-prompted clean air traffic control speech. *spsc.tugraz.at*, 2008.

[28] J Smith III and J Abel. Bark and erb bilinear transforms. *Speech and Audio Processing*, Jan 1999.

[29] F Jabloun, AE Cetin, and E Erzin. Teager energy based feature parameters for speech recognition incar noise. *IEEE Signal Processing Letters*, 6(10):259–261, 1999.

[30] JC Junqua. The lombard reflex and its role on human listeners and automatic speech recognizers. *The Journal of the Acoustical Society of America*, 93:510, 1993.

[31] JF Kaiser. Some useful properties of teager's energy operators. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993. ICASSP-93.*, 3, 1993.

[32] D Klakow. Digital signal processing (lecture notes), chapter 8. `http://www.lsv.uni-saarland.de/Vorlesung/Digital_Signal_Processing/Summer09/DSP_09_Chap8.pdf`, retrieved May 05, 2009.

[33] S Lively, D Pisoni, W Van Summers, and R Bernacki. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America*, Jan 1993.

[34] Aleix Martinez. Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–6, May 2001.

[35] I.R Murray, C Baber, and A South. Towards a definition and working model of stress and its effects on speech. *Speech Communication*, page 10, Jan 1996.

[36] Encyclopædia Britannica Online. "phoneme". `http://www.britannica.com/EBchecked/topic/457241/phoneme`, retrieved May 05, 2009.

[37] P Price, WM Fisher, J Bernstein, and DS Pallett. The darpa 1000-word resource management database for continuousspeech recognition. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 651–654, 1988.

[38] LR Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[39] MA Rahurkar and JHL Hansen. Frequency distribution based weighted sub-band approach for classification of emotional/stressful content in speech. *Eighth European Conference on Speech Communication and Technology*, 2003.

[40] MA Rahurkar, JHL Hansen, J Meyerhoff, G Saviolakis, and M Koenig. Frequency band analysis for stress detection using a teager energy operator based feature. *Seventh International Conference on Spoken Language Processing*, 2002.

[41] P Rajasekaran, G, and J Picone. Recognition of speech under stress and in noise. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, pages 1–4, May 1986.

[42] E Ruzanski, J Hansen, J Meyerhoff, and G Saviolakis. Effects of phoneme characteristics on teo feature-based automatic stress detection in speech. *Acoustics*, Jan 2005.

[43] J Siegemund. Hauptkomponentenanalyse, proseminar "robuste signalidentifikation", ws 2003/2004. `http://www-mmdb.iai.uni-bonn.de/lehre/proprak0304/siegemund.pdf`, retrieved May 05, 2009.

[44] Malcolm Slaney. An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer Technical Report #35*, pages 1–42, 1993.

[45] H Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(5):599–601, 1980.

[46] `http://en.wikipedia.org/wiki/File:Illu01_head_neck.jpg`, retrieved May 05, 2009.

[47] P Vary and M Rainer. *Digital Speech Transmission*. John Wiley & Sons, 2006.

[48] D Ververidis and C Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9):1162–1181, Sep 2006.

[49] WA Wahyudi and S Mohamed. Intelligent voice-based door access control system using adaptive-network-based fuzzy inference systems (anfis) for building security. *Journal of Computer Science*, 3(5):274–280, 2007.

[50] B Womack and J Hansen. N-channel hidden markov models for combined stressed speech-classification and recognition. *Speech and Audio Processing*, Jan 1999.

[51] S Young, G Evermann, and M Gales. The htk book (for htk version 3.4), 2006.

[52] G Zhou, J Hansen, and J Kaiser. Classification of speech under stress based on features derived from the nonlinear teager energy .... *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL* ..., Jan 1998.

[53] G Zhou, J Hansen, and J Kaiser. Linear and nonlinear speech feature analysis for stress classification. *Fifth International Conference on Spoken Language Processing*, Jan 1998.

[54] G Zhou, J Hansen, and J Kaiser. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing*, Jan 2001.

List of Figures

1.1. Taxonomy of stress measurement methods. 1
1.2. Workload and performance as a function of the demand (from [16]). 3

2.1. Stress as one of several influences on the speech production process (from [19]). 6
2.2. The human vocal tract (from [46], adapted). 6
2.3. An extended model of speech production and the influence of external stimuli (stressors) of various orders (from [24]). 7

3.1. The feature extraction and evaluation framework. 10
3.2. Pitch histogram plots for talking styles angry vs. neutral (Emo-DB database) . . 14
3.3. Intensity histogram plots for talking styles angry vs. neutral (Emo-DB database) 15
3.4. Duration histogram plots for talking styles angry vs. neutral (Emo-DB database) 16
3.5. Histogram plots of jitter, shimmer, spectral mean, and spectral slope for talking styles angry vs. neutral (Emo-DB database) 17
3.6. F1 histogram plots for talking styles angry vs. neutral (Emo-DB database) . . . 19
3.7. F2 histogram plots for talking styles angry vs. neutral (Emo-DB database) . . . 19
3.8. The nonlinear wave propagation model after Teager (from [54]). 20
3.9. Extraction of the TEO-CB-AutoEnv feature (from [54]). 21
3.10. TEO-derived feature characteristic histogram plots for talking styles angry vs. neutral (Emo-DB database) 22
3.11. MFCC histogram plots for talking styles angry vs. neutral (Emo-DB database) . 24
3.12. Harmonicity histogram plots for talking styles angry vs. neutral (Emo-DB database) 24
3.13. Histogram plots of zero-crossing rates for talking styles angry vs. neutral (Emo-DB database) 25
3.14. PCM waveforms and a variety of extracted features for the word *enter* (from the SUSAS database) spoken under three different stress conditions. 27
3.15. A sequence $x_i(n)$ and the corresponding jump function $J_i(n)$, computed with a span of ± 5 frames (from [3]). 29
3.16. A typical accumulation function (from [3]). 30
3.17. A reciever operating characteristics curve (from [20]). 32
3.18. Feature space visualization as a part of the MATLAB tool. 33

3.19. Consecutive steps of PCA (from [43]). 33
3.20. PCA vs. LDA: two extreme cases (from [32]). 34
3.21. The basic principle of k-nearest neighbours classification. 35
3.22. The basic principle of n-fold cross validation. 35

4.1. Flow chart of the feature extraction stage. 36
4.2. MATLAB structure containing features extracted on frame level. 38
4.3. Phoneme boundary detection using the text-independent approach. 39
4.4. MATLAB structure containing features extracted on phoneme level. 40
4.5. MATLAB structure containing feature characteristics. 41
4.6. Flow chart of the feature evaluation stage. 41
4.7. Normalized histogram over phoneme durations (SUSAS and Emo-DB) from 10 to $600ms$. 43
4.8. Example of bad SUSAS label file. 45
4.9. Extract of a feature matrix. The gray area indicates mutually exclusive feature characteristics. 45

List of Tables

2.1. Taxonomy of Stressors (from [24]). 5

3.1. Feature characteristics taken for evaluation. 26

4.1. Correct detection rates (CDR) of correctly identified phoneme boundaries within specified uncertainty intervals Δ for both detection methods. 39

5.1. The 10 best performing features (SUSAS Talking Styles). 48
5.2. Percentages of correct classification rate (CCR) for varying numbers of features (SUSAS Talking Styles). 48
5.3. The 10 best performing features (Emo-DB). 49
5.4. Percentages of correct classification rate (CCR) for varying numbers of features (Emo-DB). 49
5.5. Correct classification rates for talking styles angry vs. neutral; individual feature sets. 50
5.6. Correct classification rates for talking styles angry vs. neutral; shared feature set. 50
5.7. The 10 best performing features (SUSAS Dual Tracking). 51
5.8. Percentages of correct classification rate (CCR) for varying numbers of features (SUSAS Dual Tracking). 51
5.9. The 10 best performing features (ATCOSIM). 52
5.10. Percentages of correct classification rate (CCR) for varying numbers of features (ATCOSIM). 52
5.11. Correct classification rates for neutral/low stress vs. high stress; individual feature sets. 53
5.12. Correct classification rates for neutral/low stress vs. high stress; shared feature set. 53