Prosodic and Paralinguistic Speech Parameters for the Identification of Emotions and Stress

Johannes Jany-Luig

A thesis presented to the University of Music and Performing Arts, Graz in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Doctoral Committee:	**Prof. Robert Höldrich** *Institute of Electronic Music and Acoustics, KUG* **Prof. Gerhard Eckel** *Institute of Electronic Music and Acoustics, KUG* **Prof. Sonja A. Kotz** *Faculty of Psychology and Neuroscience, Maastricht University*

June, 2017

Abstract / Zusammenfassung

In addition to *what we say*, it is mainly *the way we say it* which reveals how we think about the information we are just transmitting. A human listener uses prosodic and paralinguistic cues to decode this information, probably supported by information from facial expressions and gestures. This thesis explores to which extent a computer can solve the task of emotion and stress recognition from the human voice alone. Linguistic and musical approaches to grasp "prosody" are brought together and implemented as algorithms.

Based on speech data both from an existing database of emotional speech and a self-created database of speech under stress, 27 different speech parameters are calculated from recorded speech signals without any meta-information. These parameters are then investigated regarding their ability to differentiate between different emotional states or different levels of cognitive stress.

Beim Sprechen vermitteln wir neben dem eigentlichen Inhalt durch unsere Sprechweise, wie wir zu dem Gesagten stehen. Ein menschlicher Hörer erkennt dies durch Interpretation prosodischer und paralinguistischer Merkmale der Sprache sowie an unserer Mimik und Gestik. In dieser Dissertation wird untersucht, inwieweit ein Computer in der Lage ist, Emotionen und Stress rein anhand der Sprechweise zu erkennen. Dazu wird *Prosodie* auf linguistische wie musikalische Art interpretiert und in Form von Computeralgorithmen implementiert.

Sprachdaten aus einer existierenden Datenbank mit emotionaler Sprache sowie aus einer selbst erstellten Datenbank mit Sprache unter Stresseinfluss dienen als Grundlage für die Berechnung von insgesamt 27 verschiedenen Sprachparametern, welche anschließend hinsichtlich ihrer Fähigkeit, zwischen verschiedenen emotionalen Zuständen oder Stress-Levels unterscheiden zu können, untersucht werden.

Acknowledgments

I would like to take this opportunity to express my gratitude to a few people who particularly contributed to the genesis of this thesis:

To my doctoral committee — **Robert Höldrich**, **Gerhard Eckel**, and **Sonja Kotz** —, for their willingness to supervise this thesis, for their interest in my work, and for the inspiration I took home from various discussions.

To the former head of the Institute of Electronic Music and Acoustics, **Alois Sontacchi**, for offering me a research position at the institute, and for being supportive in many ways over the years. In this respect, I also have to thank the **KUG Doctoral School** who granted me an interim two-year position as a university assistant and lecturer.

To **all my colleagues at IEM**, especially the former "first floor" which was then outsourced to P116, for creating a great atmosphere to work in.

To my manager at AVL, **Michael Kordon**, for granting me those three months of educational leave to finish this thesis. (And to my workmates from the methodology team who managed to do without me during this period! ☺)

To the many fellow researchers in the different fields I touched during my work on this thesis, for sharing their knowledge and for being open for collaboration: **Volker Dellwo**, **Nandu Goswami**, **K. Wolfgang Kallus**, **Roland Kehrein**, **Regine Porsch**, **Rudi Villing**, **Petra Wagner**, **Stephen Zahorian**, ... this is an alphabetic list, and I might have forgotten someone. Sorry!

A warm "thank you!" to **all my family and friends**, for making my life more fun and colorful.

Special thanks to my parents, **Renate and Klaus Luig**, who made clear from the very beginning that they would always support me in whatever I wanted to do. See what this can lead to!

Evelyne, **Simon**, and **David**: I love you. You inspire me every day, each of you in his and her own way. Thank you for sharing your life with me.

Credits

The flight simulator recordings took place on September 2nd and 3rd, 2010, at **Aviation Academy Austria** in Neusiedl/See, Austria. Thanks to the CEO, **Thomas Herrele**, for providing the simulator at cost price and for making contacts to the pilots and the supervisors.

The flight program has been designed and implemented by **Michael Kircher** and **Michael Kattner** who also supervised the flight simulator and played a few other roles throughout the recordings. Special thanks for your commitment! A big "thank you" also goes to the **eight pilots** (whose names will not be listed here for confidentiality reasons) for their interest and participation in the experiments.

Parts of the work reported in this thesis, especially the creation of the pilot speech database, were supported by **EUROCONTROL** under grant number Graz/08-120918-C. Thanks to **Chris Shaw** for the good collaboration. Those readers who are interested in using the database for their own research are encouraged to contact Chris at `chris.shaw@eurocontrol.int`.

Thanks to **Max Moser** from the **Human Research Institute** in Weiz, Austria, for providing the *ChronoCord* heart rate measurement devices. The data were post-processed and analyzed by **Matthias Frühwirth** who also explained the meaning of the several HRV parameters to me. Sorry that the diverse research projects we have applied for never were granted...

All the algorithms mentioned in this thesis have been written in *MATLAB* by myself, except for the following parts:

- The implementation of the *YAAPT* pitch tracking algorithm has been kindly provided by **Stephen Zahorian** and **Hongbing Hu**.
- **Rudi Villing** was so nice to share his implementation of the Pompino-Marschall model for p-center estimation with me.

Contents

1. Introduction

Speech originates from affect. This is why a speaker's emotional state and stress level are reflected in the prosodic and paralinguistic characteristics of the voice. This chapter presents an overview of relevant research on emotion and stress recognition from the speech signal, discusses various models of emotions and stress, and summarizes the aims and contributions of this thesis.

1.1. Setting the Scene

Imagine sitting in a coffee bar and overhearing a telephone conversation at the table behind you. The person does not speak your language, neither do you know who's talking at the other end of the line. But you will probably be able to assess the current mood of that person; independent of her age and sex. Where have you learned how to do that? Which acoustical cues indicate that the person just received some good news? The answer will probably be: "it's in the way (s)he talks".

In addition to *what we say*, it is mainly *the way we say it* which reveals how we think about the information we are just transmitting. According to Drach (1926), all speech originates from expressions of emotions which have been weakened and blurred through education and socialization. Any statement we make can be marked as important or dispensable, as serious or ridiculous, as substantiated or questionable, by sending meta information which are encoded in the acoustic properties of our speech. Similarly, one might be able to tell *from our voice* that we are unable to cope with a complex task we're given, since our main attention is currently centered on this task and not on our phrasing

While a human listener usually could tell the speaker's emotional state during a conversation without circumstances, this is a non-trivial task for a machine. Focusing on the voice implies that the computer will get no contextual information as facial expressions, gestures or any information on the spoken content. Furthermore, a machine by itself has no experience in differentiating between emotions; we have to introduce some a-priori knowledge. This raises the question what kind of experience a human listener utilizes when judging the emotional state of a speaker — not to mention that the notion of "emotions" is not universally agreed.

So, why should we make an effort in teaching a computer to recognize emotions and stress levels?

As a scientist, I have to respond: “because it’s interesting to see how far we can get in trying this!”, which should be the general **academic** mindset. The behavioral sciences might be interested in finding out which acoustic parameters are used to signal emotions or stress; if there are differences between the sexes or between cultures, and which of these parameters can be controlled consciously and which not. In applied computer science, people try to produce synthesized speech which sounds as natural as possible.

From a **clinical** point of view, the speech signal is a physiological marker which can be obtained in a non-invasive, non-intrusive and also non-expensive way. Mental disorders such as depression have complex clinical characterizations and symptoms which are not directly measurable; the patient’s voice is one of several indicators which is subjectively judged by the therapist. Here, an objective analysis result could facilitate more accurate diagnoses. Speech monitoring for the detection of voice activity or panic might be a potential feature of ambient assisted living systems which aim to support the everyday life of elderly and disadvantaged people in a non-intrusive way.

In many areas where **safety** is of the essence, human performance has become the limiting factor in a technologically advanced and highly automated environment. In aviation, aircraft pilots and air traffic controllers are responsible for logical reasoning and decision-making in short periods of time in a noisy high-stress environment, but their ability to respond is limited and dependent on a variety of environmental factors. As a reason, the majority of incidents and accidents can be traced back to human error rather than mechanical or technical failure. A speech monitoring system that evaluates indicators of fatigue and excessive demand could be easily implemented, since verbal communication is an integral part of the job and happens remotely; meaning that the voice is recorded anyway.

Possible **commercial** applications include human-computer interaction systems as in call centers or for ticket reservation applications, where annoyed or over-challenged customers could be automatically handed over to a human operator. Thought one step further, video game manufacturers could get reliable feedback from their users on the acceptance of certain features or if the level of difficulty has been chosen appropriate.

As evident from this multitude of potential applications, the question in which manner and to what extent emotions and stress are reflected in the human voice is investigated in different fields of research. Roughly grouping them by focus, we have the "basic research" group on the one hand, including phoneticians[1], psychologists, and medical researchers, who are concerned with the physical properties of speech sounds and typically analyze data under laboratory conditions to demonstrate fundamental concepts and relationships. On the other hand, we have the "applied research" group including engineers and computer scientists who are mainly interested in the feasibility of methods and algorithms. Research principles and standards differ greatly between these two groups; too often, basic research produces in-depth analyses of a few selected phrases only (which makes it hard to generalize the findings), while applied research produces results which allow conclusions to be drawn on the performance of an algorithm, but not on the fundamental problem.

The aim of this thesis is to contribute to a mutual understanding of both necessities and possibilities in this field. I want to introduce some general linguistic knowledge to automated speech analysis of large databases by considering not only basic acoustic properties of the speech signal, but *prosodic* variables which take human speech production and perception into account. At the same time, I hope to provide a tool for linguistically motivated research which facilitates the analysis of larger amounts of spoken utterances and even spontaneous speech by removing the need for metadata such as annotated syllable bounds or prominence levels.

[1] In simple terms, *phoneticians* are linguists who study the production and perception of human speech, while *phonologists* are concerned with the systematic organization of fundamental sounds and linguistic meaning.

1.2. Prosody and Paralinguistics

1.2.1. Terms and Definitions

Unfortunately, key terms for the description of *the way we say it* are sometimes used synonymously or even contradictory in the linguistic literature. What they all have in common, however, is their **suprasegmental** nature; they superpose the single speech segments [2], but are not temporally restricted to them.

I will use the following terms throughout this thesis:

Prosody (from ancient Greek: *proso-idía*, "song sung to music") represents the melodic, rhythmic, and dynamic properties of speech in general, of a certain language, or of a single utterance. The prosodic variables are intonation, duration, and prominence.

- **Intonation** is a term for all melodic aspects related to speech. This definition corresponds to what some publications refer to as *intonation in the narrower sense*, whereas *intonation in the wider sense* is equivalent to the definition of prosody as explained above.
- **Duration** is related to the rhythmic phenomena of an utterance. The basic unit for the impression of rhythm in speech is the syllable; however, we will see that the exact determination of rhythmic "events" is a non-trivial task.
- **Prominence** describes the extent to which a syllable or a word perceptually "stands out" of its environment (Terken, 1991). This term is equivalent to *linguistic stress* which can often be found in the literature as well.

In contrast to prosody, **paralinguistic** speech properties do not affect the linguistic identity or meaning of what is said (Schötz, 2003).

- **Timbre** is a term to describe those (mainly) spectral properties which make up the characteristic "tone" of a sound. Interestingly, the only technical description of timbre is a negative definition: "Timbre is that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar." (ANSI, 1960).

[2] A speech segment is defined as "any discrete unit that can be identified, either physically or auditorily, in the stream of speech" (Crystal, 2011). The basic phonetic unit is the *phone*, but to explore the musical aspects of speech, we will treat syllables as single segments.

- **Voice quality** coincides with timbre and is often used in a clinical sense; it generally refers to voicing characteristics which are associated with different vibratory patterns of the glottis.
- **Speaker-specific characteristics** as pitch range, average loudness or general speaking tempo emerge from the same acoustic and perceptual quantities as the prosodic variables, but are constant over a longer term and can thus be regarded as stationary characteristics in this context.

The following chart summarizes the main linguistic terms which will be used throughout this thesis.

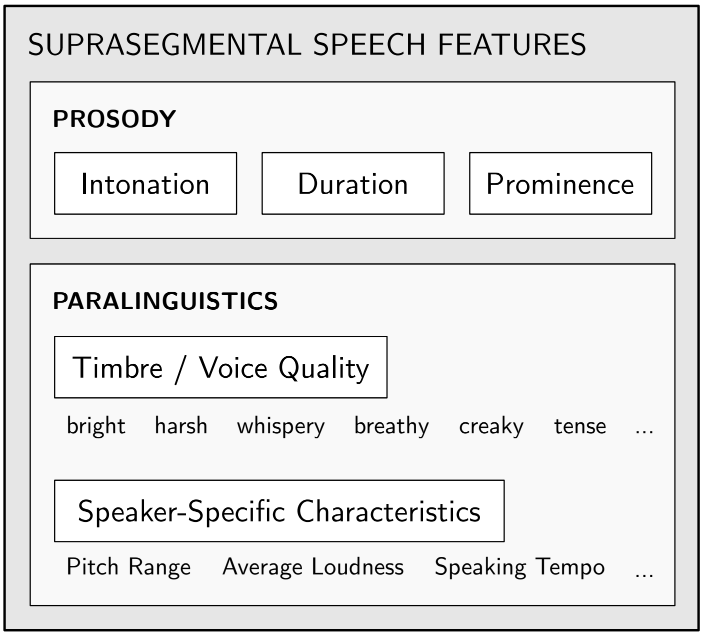


Figure 1.1.: Suprasegmental features.

1.2.2. The Role of Prosody in Speech

Prosody has primarily a **linguistic** function which corresponds to punctuation in written text and can be demonstrated by reading out the following three sentences:

1. *awomanwithouthermanisnothing*
2. *a woman without her man is nothing*
3. *A woman: without her, man is nothing.*

Speech itself is a continuous, complex flow of sounds. Without any prosodic cues, our brain has just as much difficulties to decode this flow into meaningful chunks of information as you have had to read the first sentence above. By bringing melody and rhythm into the language, prosody does not only facilitate intelligibility, but may also influence the meaning of a statement, as a comparison of the second and the third sentence shows.

Prosodic cues on the *sentence level* put the focus on a specific word to determine the meaning of the sentence; they also allow to differentiate between different sentence types by distinguishing a plain statement from a question or an exclamation. Speech pauses are indicated by periods, commas, colons and hyphens; question or exclamation marks are global signs for the melodic contour throughout the sentence or the emphasis given to what is said. Prosody on the *syllable level* realizes pronunciation rules by emphasizing single syllables while attenuating others. Interestingly, those pronunciation rules are usually not explicitly marked in normal orthography; except for some languages which use diacritics to mark the accented syllable within words (as, e.g., in the French word *café*).

In this thesis, however, I concentrate on the **paralinguistic** function of prosody. The statement “oh, that’s great!” can be uttered in different ways to demonstrate joy, frustration, or indifference, just by changing some melodic parameters. At this point, it is important to note that the prosodic parameters to be investigated will hopefully reflect the speaker’s emotional state or stress level, but also convey the realization of pronunciation rules on the syllable level (which are assumed not to be affected by emotions or stress). An attempt to separate the linguistic from the paralinguistic aspects will be the use of relative values instead of absolute values (see section 5.3.2).

As mentioned above, timbre has no linguistic function at all. It constitutes the identity of a speaker’s voice and may point to his or her emotional state as well, but the meaning of *what is said* will not be affected by timbre.

1.2.3. Acoustic, Auditive and Prosodic Variables

Kehrein (2002) created a prosody model for German as an “integrative proposal for the reassignment of aspects on which consensus might be obtained”. The model, shown in Fig. 1.2, differentiates between *form* (**acoustic** and **prosodic variables**) and *function* (discrete prosodic units). This differentiation is a controversial issue in the linguistic community (Mixdorff, 2002), but we can neglect this discussion, as we will anyway concentrate on the *form* component of this model here.

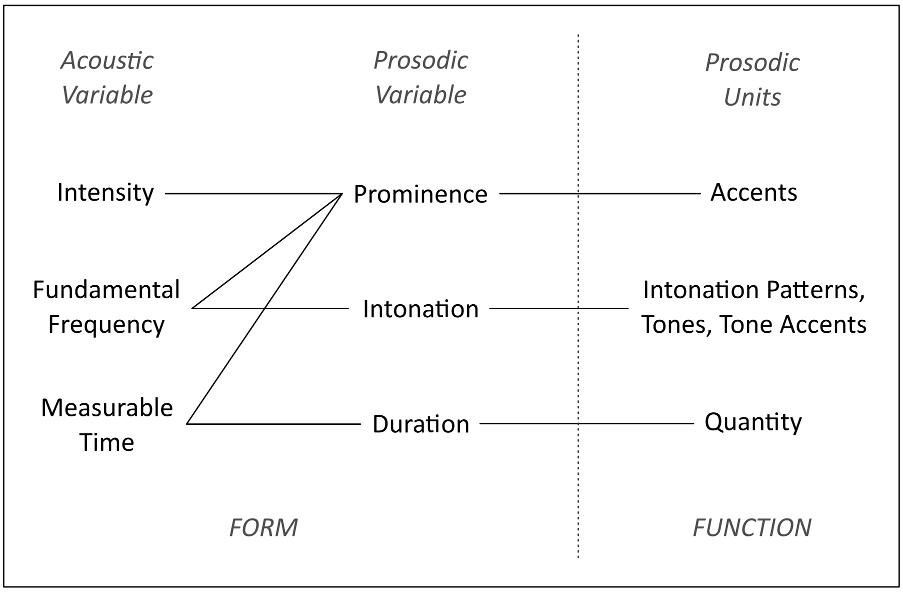


Figure 1.2.: Kehrein's prosody model (adapted from (Kehrein, 2002))

Although this model is missing an intermediate layer of auditive variables, it covers the important fact that the perception of prominence depends upon all three acoustic variables. Since at least Abercrombie (1967), the linguistic world is roughly divided into *syllable-timed* and *stress-timed* languages[3], with both English and German belonging to the latter class. This means that speakers of English or German make use of all three possible ways of accentuating a syllable, which includes the variation of intensity, fundamental frequency, and syllable length.

To account for the fact that there is no linear relationship of *what we can measure with a microphone* and *what we hear*, I will extend Kehrein's prosody

[3] Despite many a criticism in view of this hard categorization (and the fact that there exists at least a third rhythmic type of *mora-timed* languages), it is an indisputable fact that Romanic languages as French or Italian sound more "rhythmically even" than Germanic languages like English or German since there is less variation in syllable length.

model by inserting a set of **auditive variables** in between the acoustic and the prosodic variables, as depicted in Fig. 1.3. Auditive variables are calculated from acoustic variables using special formulas and filters which model the nonlinear characteristics of human auditory perception, such that a linear change in an auditive variable (e.g., "twice as much") leads to the same perceived change. Detailed explanations of the auditive variables *loudness*, *pitch*, and *perceived syllable length* as well as the underlying principles are given in section 3.2. The question to what extent these individual variables contribute to the prominence of a syllable is discussed in section 3.5, and may be language-specific as well as the result of an individual interpretation by the speaker.

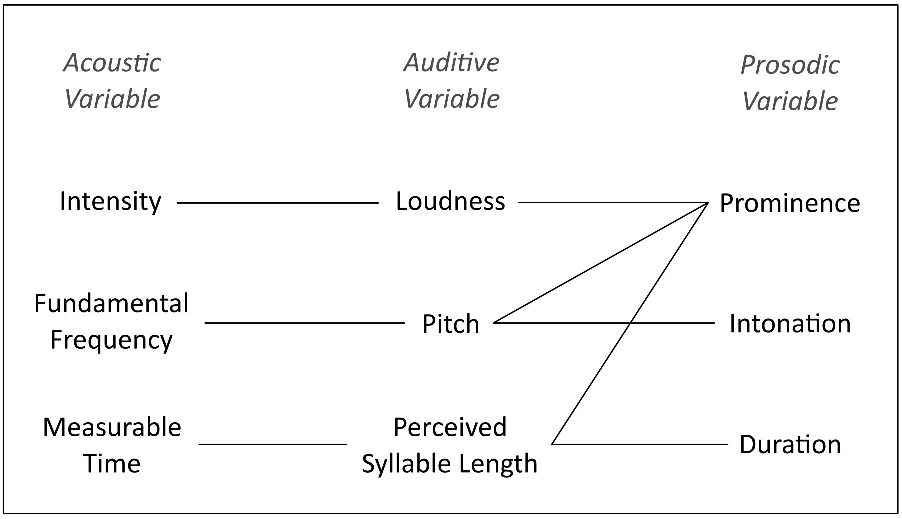
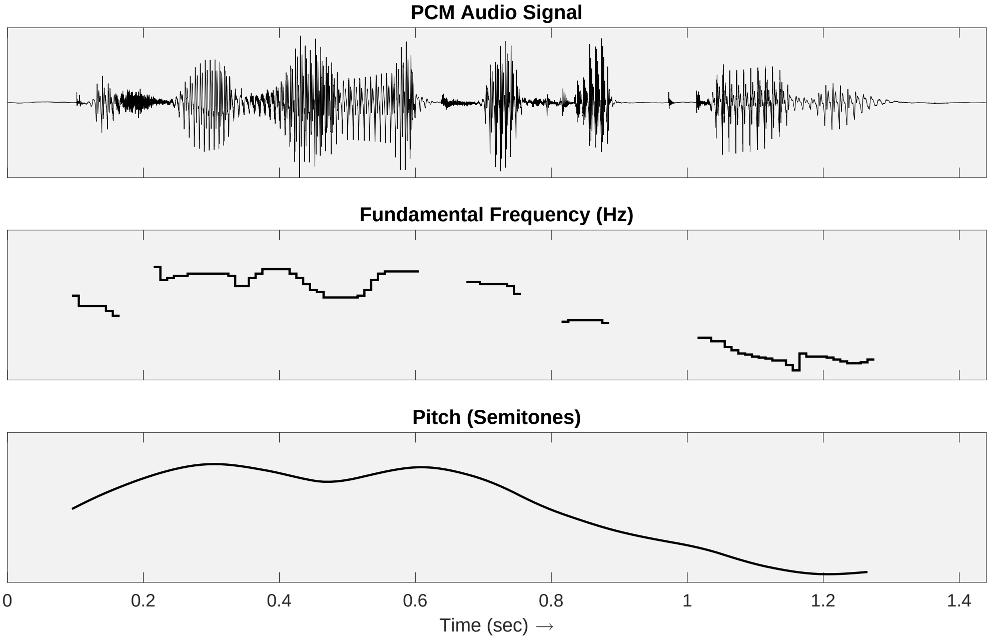


Figure 1.3.: Prosodic model as used in this thesis

To avoid confusion, the term **variable** will consistently be used for a general speech property on the several levels — acoustic, auditive or prosodic — which changes over time, whereas the term **parameter** implies its quantification[4]. Thus, the parameter "declination" quantifies a certain attribute of the auditive variable "pitch" which has for its part been derived from the acoustic variable "fundamental frequency", as visualized in Fig. 1.4. One can imagine that the continuous and relatively smooth pitch contour in the bottom plot matches our perception of *speech melody* much better than that bumpy and interrupted fundamental frequency contour in the plot above; even if the latter is closer to the acoustic "truth".

[4] In the literature, the term *speech feature* is also used frequently; it has the same meaning as *speech parameter* in our case.



Declination: -4.57 semitones/sec

Figure 1.4.: PCM audio signal (top plot), fundamental frequency contour (middle plot), pitch contour (bottom plot), and declination value (below).

1.3. Emotions and Stress

The main hypothesis of this thesis is that the emotional state or stress level of a speaker is reflected in "the way (s)he talks". In order to find descriptive speech parameters reflecting these influences, we must understand the commonalities and divergences between emotions and stress. The interdependence of both is beyond question; reasons for their co-existence as separate fields in the social sciences are mainly of historical nature. Lazarus (2006) even argues that emotions such as anger, jealousy, shame, or sadness should be called "stress emotions", since they arise from stressful conditions[5].

1.3.1. Emotions

From everyday experience, we can all tell that our emotions may be of varying intensity and quality. Furthermore, we often experience several emotions at once and may sometimes not be able to exactly describe "how we feel". As emotions play a major role in many applied fields from psychotherapy to advertising psychology, a variety of different emotion theories exists. Depending on the respective field of research, their focus is either on body responses to a stimulus (physiological approach), brain activity due to a stimulus (neurological approach), or the cognitive evaluation of the actual situation (cognitive approach).

Emotions as Categories

Many of these emotion theories have in common that emotions are considered *categorical*, meaning that any emotional state can be attributed to a set of *basic emotions*. This conjecture goes back to Darwin (1872), who postulated that emotions were biologically determined and thus universal across cultures. Ekman and Friesen (1971) tried to verify this theory by studying how people from different cultures assign facial expressions to a specific emotion. They were able to show that even subjects from a preliterate culture[6] in New Guinea provided results which agreed very well with those from college graduates from Brazil, the United States, or Japan. The six emotion expressions found to be universal included **anger**, **disgust**, **fear**, **happiness**, **sadness**, and **surprise**; they are sometimes referred to as the "big six" emotions.

[5] Lazarus goes even further by stating that even "positive" emotions such as happiness are related to stress in the way that we might fear that our happiness will eventually end. This is a rather philosophical point in my view, hence I will focus on the "negative" emotions when associating them with stress.

[6] A preliterate culture is a culture which does not have a written language.

Plutchik (1980) defined eight *primary emotions* in four pairs of opposites, namely **joy/sadness**, **trust/disgust**, **fear/anger**, and **surprise/anticipation**. By arranging them in a circumplex pattern such that the counterparts are placed vis-à-vis, Plutchik created a three-dimensional conical model, where the circular angle represents degrees of similarity among the emotions and the vertical dimension indicates the intensity. When mapping the cone onto a two-dimensional plane (as shown in Fig. 1.5), the blank spaces between the cone ends can be filled with *secondary emotions* as blends of two adjacent primary emotions.

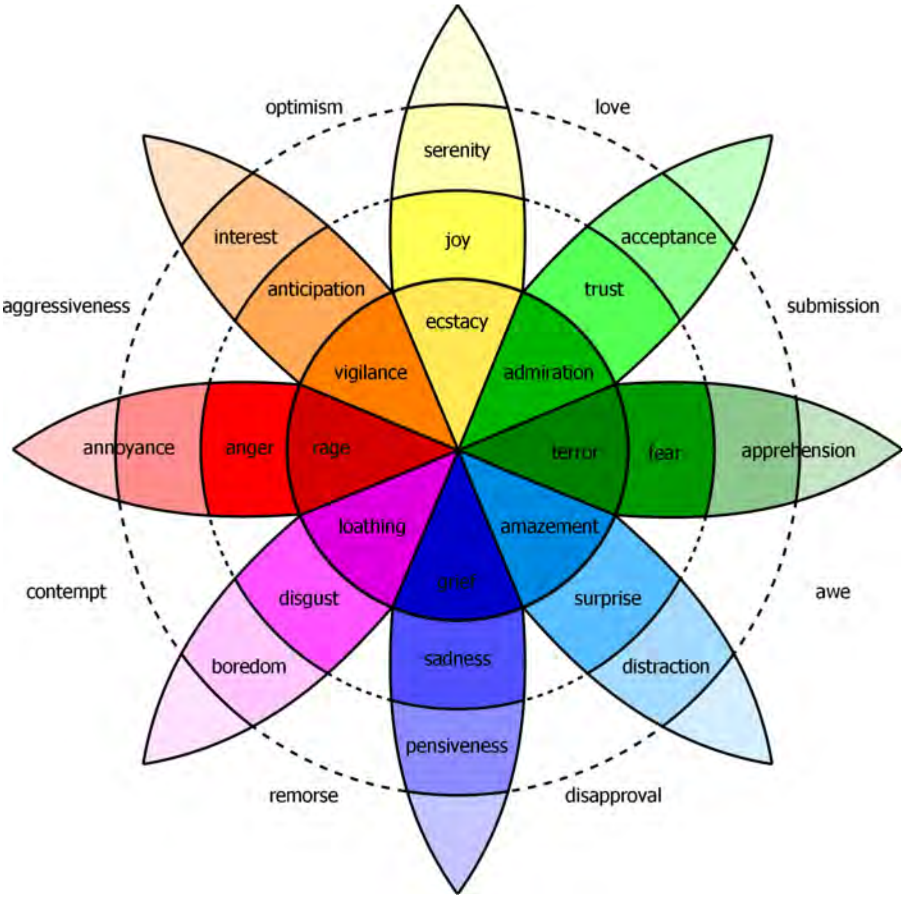


Figure 1.5.: Two-dimensional mapping of Plutchik's cone of emotions (from Plutchik (2001))

In the literature, many studies on emotions and speech use the following six basic emotions: **anger**, **anxiety**, **boredom**, **disgust**, **happiness**, and **sadness**. Compared to Ekman's "big six" emotions, *fear* has been replaced by *anxiety*[7], and instead of *surprise*, we now have *boredom* as a new category.

[7] Psychologists might disagree that these terms are fully interchangeable, but for these basic treatment of "emotions", I will assume it to be equivalent.

Emotions as Dimensional Variables

As an alternative to strict categories, emotions can also be described as points in a multidimensional emotion space. This *dimensional* interpretation has the advantage that the intensity of an emotion can be described as well as gradual transitions between emotions (Cowie and Cornelius, 2003). Virtually all researchers agree that there are at least two special characteristics of emotions, namely their **valence** (or *appraisal*, *evaluation*; positive vs. negative) and their **arousal** (or *activation*, *excitation*; active vs. passive). Many recent publications deal with an additional third dimension, **dominance** (or *potency*, *power*; strong vs. weak). Dominance allows to distinguish, e.g., anger from anxiety/fear, which are both "negative" and "active" emotions, but differ in the person's ability to cope with the situation (Grimm et al., 2007).

These three dimensions form an abstract emotional space in which the basic emotions can be roughly placed due to their attributes (Fig. 1.6). Except for *neutral* (which forms the center of this cube) and for *disgust* (which lies somewhere in-between), the basic emotions are commonly viewed as being extreme in any of the three dimensions and thus are sketched in the corners of that cube.

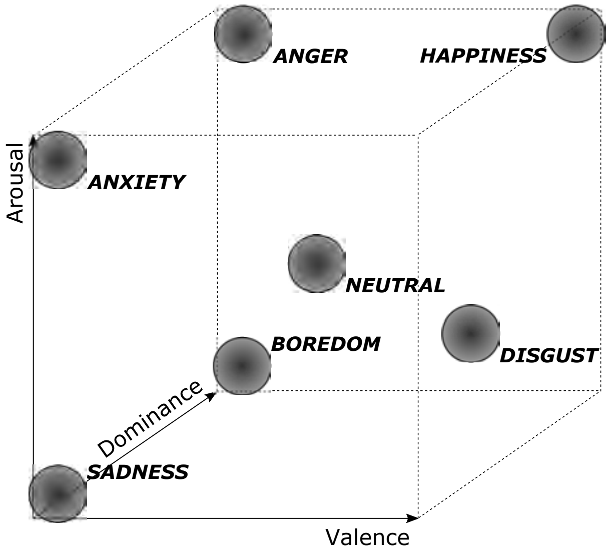


Figure 1.6.: Six basic emotions (plus "neutral") and their hypothesized positions in a three-dimensional emotional space.

Fontaine et al. (2007) applied Principal Component Analysis to a set of 144 emotional features — appraisals, bodily experiences, facial and vocal expressions, gestures, etc. — as evaluated by a large number of test subjects from three different European countries and languages, and found **unpredictability** as a potential fourth dimension.

Categorical? Dimensional?

The concepts of categorical and dimensional emotions are not mutually exclusive; lots of research has been done which aims at placing a set of emotional labels inside a two-, three-, or even four-dimensional space. Barrett (1998) reports that the extent to which we consciously experience the different dimensions of emotion might determine whether a categorical or a dimensional emotion model best captures how we label our affective states.

In this thesis, I will investigate the discriminative power of prosodic and paralinguistic parameters with regard to classification into discrete emotions only, because this coincides with the nature of the investigated speech data, which are sentences produced in different emotional categories.

1.3.2. Stress

Just as for emotions, the discussion on "stress" is a matter of controversy. Many theories exist which agree in some parts, but disagree in others. Stress is a subject of research in physiology, psychology and sociology – not to mention the origin of the term *stress*, which lies in the physical world, where it is understood as a force causing a material deformity that results in *strain.* The common denominator is, however, that stress is perceived as aversive and accompanied by negative emotions.

Based on pioneering work by the physiologists Bernard (19th century) and Cannon (1932), Selye (1950) defined stress as *the nonspecific response of the body to any demand* which triggers a "general adaptation syndrome". This concept was later refined by Lazarus (1998), who defined the following four elements which must be included in modern stress concepts:

1. A causal external or internal cause: the *stressor,*
2. An *evaluation* to distinguish "good" from "bad",
3. Coping processes to deal with stressful demands, and
4. A complex pattern of effects: the *stress reaction.*

With regard to speech production, **cognitive stress** is of special interest. The cognitive system is responsible for how we perceive things, how much attention we are able to pay on something, how we make decisions — and also which words we use and how we say them. Stress can be manifold and thus originate also from non-perceptual sources; we may, for example, be limited in our cognitive abilities today because we haven't got enough sleep last night, which in turn degrades our performance on a given task. Let's have a look on different categories of "stress".

A Taxonomy of Stressors

On a NATO workshop on speech under stress in 1995, a taxonomy of stressors has been worked out which was published in (Hansen et al., 2000). It differentiates between four degrees of "stress", as listed in Tab. 1.1, which are mutually independent[8]. Stressors from all categories can affect the speech production process on different levels, as we will see in section 3.1.1.

Order	Category	Examples
0	physical	Vibration, Acceleration (G-force), Personal Equipment, Pressure Breathing, Breathing Gas Mixture
1	physiological	Medicines, 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 1.1.: Taxonomy of stressors (from Hansen et al. (2000)).

In the course of this thesis, I have created a database of airline pilots' speech under stress (described in chapter 2). According to the test design, the test subjects have mainly experienced third-order stressors during the recordings. With regard to *cognitive stress*, however, all stressors from this list can affect the ability to keep one's head clear for a given task and it seems to be impossible to unravel the complex interaction of potential stressors.

The good news is that the identification of stressors being responsible for a stress reaction is not of the essence when we are only interested in the *intensity* of the reaction.

Demand and Response Capability

A suitable interpretation of "stress" for our needs can be found in a review article by Koolhaas et al. (2011):

> *We propose that the term "stress" should be restricted to conditions where an environmental demand exceeds the natural regulatory capacity of an organism, in particular situations that include unpredictability and uncontrollability.*

[8] We should be aware of the fact that the *North Atlantic Treaty Organization* is a military alliance which reflects in examples which are not commonplace, such as "pressure breathing". Nonetheless, I think that this taxonomy is generally valid.

This is identical to what Lazarus (2006) calls the *response approach*: a stress reaction is not only dependent on the stressor, but also (and to a considerable extent!) on the individual and the situation. Potential conditions include age, experience, training, personality, time-of-day and current mood, to name just a few.

So, we assume that different individuals will show different reactions when faced with the same situation or task. From an experimental point of view, this concept has two implications:

+ The charming aspect about it is the definition of "stress" as something *relational*, which makes it suitable for experiments with a **limited number of participants**, since individual differences in coping with stressful demands are already considered. If we took just the task complexity itself as a measure for the estimated workload, a large number of test persons would be necessary to assess trends of general behavior.

– In turn, the intensity of the **stress reaction has to be measurable** in some way, which means additional efforts to be made.

This drawback, however, carries only little weight because there are established physiological measures which have proven to indicate stress and which can be obtained in a rather uncomplicated way (see section 2.3).

1.4. State of the Art

1.4.1. Emotion Detection from Speech

The first empirical investigations on how emotions affect the human voice date back to the pre-computer era (Scripture, 1921; Skinner, 1935; Fairbanks and Pronovost, 1939; Fairbanks and Hoaglin, 1941), where the available technologies allowed for **qualitative analysis** of visualized waveforms only. This glass ceiling was broken through when the first digital computers entered the laboratories around the world in the 1960s. Psychiatrists started first attempts in diagnosing affective states through voice analysis, and linguists were more and more not only interested in automated speech recognition, but also in the electronic analysis of paralinguistic phenomena. It should take, however, another decade, until Williams and Stevens (1972) made the next step by performing **quantitative analysis** of speech parameters for a few sentences produced in four different emotions. They were able to produce some general statements with respect to basic parameters of the fundamental frequency, but had to conclude that “at present it is certainly not possible to specify any quantitative automatic procedures that reliably indicate the emotional state of a talker”.

That seemed to discourage potential imitators for a while, because — apart from several qualitative results from different studies which have been excellently summarized by (Cowie et al., 2001) — it was not until the mid-1990s[9], when Dellaert et al. (1996) renewed the venture towards automatic emotion recognition from speech by exploring the potential of several statistical **pattern recognition techniques**. A considerable number of related publications over the following years documents the grown interest into the topic; in the early years focusing on the **choice of suitable parameters**, which seemed to be found in the statistics of fundamental frequency and intensity as well as measures for speech rate (Banse and Scherer, 1996; Mozziconacci and Hermes, 1997; Amir and Ron, 1998; Polzin and Waibel, 2000; France et al., 2000).

In the “applied research group”, several studies then focused on alternative **classification techniques** to improve emotion recognition performance. Nicholson et al. (1999) was the first to test a neural network classifier for this purpose. Later, Hidden Markov Models (Schuller et al., 2003; Nwe et al., 2003; Lee et al., 2004) or Dynamic Bayesian Networks (Barra-Chicote et al., 2009)

[9] Ten years earlier, Van Bezooijen (1984) had demonstrated in her Ph.D. thesis that emotional speech can be classified quite well into 10 different categories by means of discriminant analysis, what (for whatever reason) has barely been acknowledged by the scientific community.

were used as classifiers to capture the temporal evolution of the acoustic variables over time. Other classification techniques have been investigated (Yang et al., 2009), but without remarkable success. Another observable trend was the increasing popularity of the ***bag-of-features*** **method** which calculates any possible feature from the speech signal and employs complex statistical methods to select the few meaningful descriptors from the vast number of candidates (Schuller et al., 2007; Vlasenko et al., 2007b; Wang et al., 2008). This method is quite successfully applied in image classification for object recognition (e.g., human faces) from pictures, but has not proven to outperform emotional speech classification with manually selected features so far.

The "basic research group", on the other hand, continued their in-depth investigations under laboratory conditions. Paeschke and Sendlmeier (2000) tested the just released Emo-DB database with prosodic parameters as pitch range or declination. Kehrein (2002) created his prosody model based on visual analysis of fundamental frequency, intensity, and syllable durations from manually extracted speech chunks. Alter et al. (2003) investigated several spectral features as potential correlates of "breathiness" and "roughness" on isolated vowels, and employed sentences with emotional content. By measuring event-related brain potentials, they were even able to track mismatches between emotional state and lexical content of a sentence, which was also reflected in one of their features. Liscombe (2007) compared emotions perceived by listeners to those intended by the speakers using rather simple f_0- and intensity-based speech parameters. Bulut and Narayanan (2008) went the other way by systematically changing f_0 characteristics of emotional speech and presenting these modified sentences to participants of a listening test in order to find out which modifications lead to changes in perceived emotion.

Voice quality as a potential descriptor of emotional state was first considered by Ishi and Campbell (2002, in terms of breathiness) and investigated in detail by Lugger and Yang (2007) who were looking for speech parameters which describe other emotional dimensions than *arousal*[10]. It was again (Yang and Lugger, 2010) who tried to break new ground by translating pitch values into musical tones in order to investigate the "harmony" of a sentence.

Some researchers bothered to record **large-scale corpora** of spontaneous speech with emotional content in order to provide a solid data foundation for substantiated analysis results. In turn, they had to focus on single emotions or mental states to prevent losing *scope* (see section 2.1). (Ang et al.,

[10] In their experiments, Lugger and Yang found out that the popular speech parameters based on fundamental frequency and energy mainly describe the *arousal* dimension (which is "active vs. passive") and thus other features had to be found.

2002) analyzed more than 20000 sentences from human-computer dialogs by people who made air travel arrangements over the telephone. Using mainly duration- and pause-based parameters, they were able to classify "neutral" vs. "annoyed/frustrated" equally well as human labelers. Karam et al. (2014) collected over 220 hours of daily telephone conversations from people suffering from bipolar disorder with their clinicians to come up with speech parameters which facilitate recognition of manic and depressive mood states.

In addition, there have been several **content-based approaches** involving automated speech recognition (ASR) which have been excellently summarized by (Batliner et al., 2011a); the idea is that emotions are also reflected in the usage of certain words or grammatical alterations. Some attempts were also made to recognize emotion-associated non-linguistic vocalizations without ASR, such as the automatic detection of cries (Pal et al., 2006) or laughter (Petridis and Pantic, 2008).

Finally, I would like to point the interested reader to the comprehensive review of Juslin and Laukka (2003) who compiled a list of over 100 studies on the expression of emotions in speech (and, interestingly, also in music) which had been published until then.

1.4.2. Stress Detection from Speech

Early contributions to the problem of stress detection in speech came from Stevens and Williams (1969) and Kuroda et al. (1976) who evaluated several characteristics of the fundamental frequency contour[11] of military pilot voice recordings. Although comparing "neutral" speech to speech in extremely stressful situations, large individual differences in the investigated parameters allowed for **qualitative statements** only. Streeter et al. (1983) analyzed recorded telephone conversations before and during the New York City blackout of 1977 between the responsible system operator of the involved energy company and his supervisor. Although there were only two different voices to be analyzed, they were not able to find one single reliable acoustic indicator of stress.

An **analysis-by-synthesis approach** was followed by Protopapas and Lieberman (1997) who created various versions of synthesized fundamental frequency contours based on real speech under neutral and high-stress conditions, respectively. A listening test revealed that maximum f_0 seemed to be the only reliable non-linguistic indicator for stress recognition.

[11] These characteristics also included short-term perturbations as jitter which is nowadays associated with voice quality rather than f_0.

After the ***SUSAS* database** (Hansen and Bou-Ghazale, 1997) had been released, various studies on speech under stress in terms of classification performance, suitable speech parameters and its impact on speech recognition accuracy were published which make use of this database (Yao et al., 2005; Schuller et al., 2006; Vlasenko et al., 2007a; Casale et al., 2008; Luig, 2009, in addition to many publications from Hansen's research group). A comprehensive overview on that database and the distributions of classic parameter values can be found in (Hansen and Patil, 2007). In the course of these *SUSAS*-based studies, an energy-based parameter called *Teager Energy Operator* (Hansen et al., 2003) gained growing popularity; however, this parameter was only rarely employed in other, non-*SUSAS*-based studies. An investigation on how well human listeners are able to distinguish between the single classes used in *SUSAS* was done by Bolia and Slyh (2003) who, however, just used different speaking styles rather than workload levels for their evaluation.

Fernandez and Picard (2003) were the first to classify **different levels of stress** rather than just "stress" vs. "non-stress" conditions. Drivers had to solve simple arithmetic tasks while driving at different speeds, their recorded answers formed the dataset under investigation. With different TEO-based parameters, they were able to produce results well above chance level, but still far from perfect recognition.

Since the acquisition of "actual stress" speech data under laboratory conditions seemed to be impossible, several researchers concentrated on the effects of ***workload*** on speech. Griffin and Williams (1987) report significant increases in f_0 and intensity as well as decreases in word duration as the effects of increasing cognitive task complexity. Lively et al. (1993) conducted a cognitive workload experiment and found intensity mean and variability to be significantly increased in speech produced during workload tasks. They concluded that, during multi-tasking, the speakers adapted their speech to maximize intelligibility. Baber et al. (1996) found that workload can significantly affect speech recognition performance. Scherer et al. (2002) performed a large-scale study with 100 participants speaking three different languages in which the subjects performed a logical reasoning test under two different conditions, the second condition assumed to induce psychological stress. Just as many of their predecessors, they had to report strong individual differences in their few basic acoustic parameters. The classic Stroop test (Stroop, 1935) was employed by Rothkrantz et al. (2004) to check the participant's voices for effects of workload, but the results — again — were very speaker-specific. Jameson et al. (2006) analyzed the possible effects of time pressure and cognitive load on mostly duration-based speech parameters, but found no significant relationships.

At the same time, researchers were engaged in finding methodologies for estimation of induced workload as a function of the given task (e.g., Averty et al. (2002) for the special case of air traffic control), which is a necessary step to move away from two-class scenarios which are rather abstract representations of the real world. In the meantime, also Hansen's group had developed another speech-under-stress database named *UT-Scope* which contains additional heart rate measurements as physiological markers of the actual stress level (Godin and Hansen, 2008) with the same objective.

Recent research is again aviation-related (Ruiz et al., 2010; Huttunen et al., 2011), but without surprising results or new insights, unfortunately. An interesting approach was presented by (Yao et al., 2015) who argue that a physical model of speech production could best describe the variations in airflow characteristics due to stress. Their model describes the airflow characteristics of the vocal folds, the vocal tract, and the laryngeal ventricle. Using isolated vowels from three different cognitive tasks, they achieve good classification results, though under very strict laboratory conditions.

1.4.3. Acoustic Correlates of Emotions and Stress

As a starting point for my research, I have been browsing the literature for acoustic, auditive and prosodic correlates of emotions and stress. The result is shown in Tab. 1.2 which compiles the results of several meta-studies on that topic (Scherer, 1986; Murray and Arnott, 1993; Cowie et al., 2001; Luig, 2009). This table contains qualitative statements on the overall behavior of commonly investigated parameters with regard to the "big six" emotional categories as used today, ranging from "strongly increased" (++) to "strongly decreased" (−−) in six degrees. Parameters on which the literature does not reach a consensus are marked with *n.c.*, for "no consensus".

A remarkable fact is that these "classic" acoustic parameters mainly seem to reflect the *arousal* dimension, meaning that higher values of parameters based on fundamental frequency and intensity indicate active emotions as *anger* and *happiness*, whereas lower values of these parameters point to *sadness*, for example[12] [13] What should also attract our attention is the sparse number of clear statements regarding *stress*. One reason for this is certainly that individual stress reactions are substantially different among people which makes it hard to generalize findings of any kind.

[12] The *arousal* dimension is the y-axis in Fig. 1.6.

[13] Cummins et al. (2015), by the way, published a similar collection for low and high levels of speaker depression; the results are similar to what is listed for *sadness* here.

Parameter	Emotion						Stress
	Anger	Anxiety	Boredom	Disgust	Happiness/Joy	Sadness	
f_0 mean	++	+	−	*n.c.*	+	*n.c.*	
f_0 range	++		−	[+]	+	[−]	
f_0 variability	++				+	−	−
f_0 pertubation (jitter)	+				+	+	
Intensity mean	++	−	*n.c.*	*n.c.*	+	−−	+
Intensity range	+				+	−	
Intensity variability	+				+	−	+
Speech rate	+		*n.c.*	++	n.c.	[−]	
High-frequency energy	+	+	*n.c.*	+	*n.c.*	*n.c.*	
VQ: tense	✓				✓		
VQ: breathy	✓		✓		✓	✓	
VQ: chest	✓			✓			
VQ: resonant						✓	
VQ: blaring	✓				✓		
VQ: lax						✓	
VQ: grumble				✓			
VQ: creaky			✓				

Table 1.2.: Commonly evaluated speech parameters and predicted effects by different emotions and stress when compared to some "neutral" reference: strongly increased (++), increased (+), slightly increased ([+]), slightly decreased ([−]), decreased (−), strongly decreased (−−), no consensus (*n.c.*). Compiled from (Scherer, 1986); (Murray and Arnott, 1993), (Cowie et al., 2001), and (Luig, 2009)

Fundamental Frequency and Pitch Parameters

Many studies — by far not just the “early” ones — tend to describe acoustic variables by their statistics (in terms of mean and standard deviation), which is especially true for the fundamental frequency; presumably due to the fact that statistic measures are a convenient way to capture the dynamic nature of *speech melody*. However, many authors seem to disregard the fact that f_0 values often have a non-normal distribution which, strictly speaking, makes these descriptors not applicable for this kind of data. But also nonlinear descriptors as the median and inter-quartile ranges do not lead to the desired result, because they also do not account for the fact that there is a linguistic motivation in, e.g., overemphasizing certain syllables to strengthen the meaning of what is said. Also for the purpose of speaker recognition, where pitch range is generally regarded as an important feature, these rather simple measures are not very successful (Ladd et al., 1985). Some studies employed logarithmized f_0 values to roughly approximate human pitch perception; sometimes temporal derivatives (Δf_0 and $\Delta\Delta f_0$) were used.

Even within linguists and phoneticians, there is no general agreement on the term “pitch range” and how this should be measured. Patterson and Ladd (1999) have come up with a proposal for linguistically motivated *level* and *range* measures which are based on initial peaks and final lows of a sentence, as well as the other accent peaks and valleys. In a large-scale study, they were able to show that their measures outperform the distribution-based features with regard to correlation with listener’s ratings for different emotions.

Initially planned as an original contribution of this thesis, but in the meantime also proposed by Yang and Lugger (2010), is the idea to describe the musical *harmony* of an utterance. Although the human voice is monophonic by nature and thus will not produce several tones at once, a harmonic impression can still be created by successive tones within short periods of time. Think of someone coming home and saying “hallo-oh!” to tell everyone he’s back: this might sound like a major triad, for example.

Some studies also include formant frequencies and bandwidths as potential indicators for emotion. Since formants are distinctive frequency peaks which determine the quality of a vowel (see section 3.1), they rapidly change with each syllable and thus are no *suprasegmental* characteristics at all. They might be applicable when analyzing isolated speech sounds, but not for fluent speech.

Intensity Parameters

Finding parameters for intensity seems to be a rewarding task, because its course over time is mainly determined by what is said, so the average intensity as well as intensity range and dynamics are the common parameters. Intensity is calculated as the average or the sum of absolute values of the PCM signal over short-term windows; later studies mostly calculate the signal energy instead of intensity, which is simply the squared intensity.

As for fundamental frequency, energy values are sometimes logarithmized to approximate human loudness perception. I have not found a psychoacoustically valid loudness calculation (→ 3.2.3) in any publication on emotional or stressed speech.

Duration-Based Parameters

When phoneme or syllable boundaries are known, duration parameters are easy to obtain. In such cases, average and maximum durations of these segments are commonly used measures. In all other cases, mainly two kinds of duration-related parameters are calculated:

- *Speaking rate* is commonly understood as the number of syllables per time unit; it can be approximated by counting the number of coherent voiced or unvoiced periods and normalizing the result by the length of the utterance.
- *Pause-related parameters* include the speech-to-pause-time ratio or the number of "long" pauses which exceed a predefined threshold duration. The separation of speech and pause segments requires a rather simple method for voice activity detection.

Voice Quality Parameters

In Tab. 1.2 on page 21, voice quality is described through keywords, but no acoustic correlates are given. This is due to the fact that these results have been produced by studies which judged speech parameters *qualitatively*.

Popular *quantitative* measures for voice quality are jitter and shimmer, which refer to microprosodic irregularities in the glottal excitation signal with respect to timing and amplitude, respectively. This glottal excitation signal can be calculated by applying an "inverse vocal tract filter" on the speech signal; earlier studies sometimes approximated jitter and shimmer by counting the changes in sign on temporal derivatives of f_0 and energy curves. Another parameter which can regularly be found is the harmonics-to-noise ratio of the speech signal (→ 4.4.3).

Recent studies which incorporate voice quality parameters mostly use the calculation of Stevens and Hanson (1995) which is explained in section 4.4.4.

Findings from *Bag-Of-Features*

Batliner et al. (2011b) report on a large-scale *bag of features* study incorporating more than 4000 different speech parameters which were evaluated by parameter selection algorithms to ultimately lead to a set of “most descriptive” parameters. Researchers from seven institutes in four countries participated in this study, which makes it — to my knowledge — the largest attempt in this direction so far. The aim was to find a set of 150 parameters and to investigate which acoustic variables participate to what extent in the final parameter set. Their results for a 4-class problem are given in Tab. 1.3:

	DUR	**ENG**	**PIT**	**SPEC**	**CEPS**	**VQ**	**WAV**	**All**
Share [%]	18.7	22.0	15.3	11.3	15.3	7.3	10.0	100.0
F measure	54.9	56.9	46.7	49.9	50.4	41.5	44.9	63.4

Table 1.3.: *Bag-of-features* results for emotion recognition (from Batliner et al. (2011b)). *Share* indicates the percentage of features from each category in the final feature set, *F measure* is the classification accuracy when using only the respective subset for classification. Parameter categories: DUR = duration, ENG = energy, PIT = pitch, SPEC = spectral, CEPS = cepstral, VQ = voice quality, WAV = wavelets, ALL = all features (complete set).

The *F measure* used by Batliner and his colleagues is the harmonic mean of *precision* (= percentage of correct classifications for class X) and *recall* (= covered percentage of true occurrences for class X) and is a valid measure for the classification accuracy obtained with the respective set of parameters. Keeping in mind that chance level for this 4-class problem is at 25%, the results are not convincing at all.

1.5. What to Analyze?

The ultimate **database of emotional speech** should contain spontaneously uttered sentences spoken in natural emotional states while covering the whole spectrum of possible emotions at the same time. Preferable, all speech data come with annotated phoneme and syllable boundaries as well as phoneme transcriptions to facilitate detailed analysis of the elementary parts of speech.

Unfortunately, no such database exists to date. There are several databases with more or less "spontaneous" speech, mostly consisting of recorded utterances from TV talkshows, radio programs, or call-center telephone conversations. These have the disadvantage that no transcription is available, and the quality as well as the intensity of the speakers' emotions have to be judged by oneself. Acted emotions, on the other hand, offer the advantage that all emotional states can be produced in equal parts and with unique class labels. Although it is commonly assumed that acted emotions tended to be exaggerated compared to natural emotions, there is evidence that acoustic correlates of natural and acted emotions are at least not contradictory, but point in the same direction (Williams and Stevens, 1972).

From a variety of publicly available databases of emotional speech (see, e.g., Anagnostopoulos et al. (2015)), I have chosen the *Berlin Database of Emotional Speech* (Burkhardt et al., 2005). The selection criteria were the following:

- The database should consist of English or German speech, which are the two languages I am fluent in; which is, in my opinion, a prerequisite for being able to judge the quality of the data.
- The database should comprise several emotional classes, ideally the "big six" formulated by Ekman and Friesen (1971). Acted emotions are acceptable for lack of alternatives.
- Preferably, the emotional classes should have been verified by human listeners to ensure that the perceived emotions match the intended emotions.
- There should be transcriptions of syllable boundaries to facilitate prominence estimation and the calculation of rhythmic parameters.

The last point is especially important, as no robust and reliable syllable segmentation algorithm exists to date which would allow to estimate syllable times and durations from spontaneous speech. I will develop such an algorithm during this thesis, but I will need reference data to assess its accuracy.

The number of available **databases of speech under cognitive stress** is more than limited. Although Ververidis and Kotropoulos (2006) list a total of 10 “emotional data collections” with cognitive stress subsets in their review, and although considerable research on the topic has been reported in various more recent studies (Hansen et al., 2005; Sigmund, 2006; Ikeno et al., 2007; Boril et al., 2011; Truong et al., 2015; Sabo et al., 2016), the 1997 *SUSAS* database (Hansen and Bou-Ghazale, 1997) remains the only publicly available speech corpus to date. All other experiments reported in the literature use self-recorded, “custom” speech data. The experimental design is not clearly stated in most cases which makes it impossible to reproduce or to compare the results. Fundamental questions are left open: which part of the data was used for training, which for testing? Which evaluation strategy was chosen? Which assumptions were made when defining the *no-stress* condition?

Although still the major source for researchers who do not record their own speech datasets[14], *SUSAS* suffers from a few, but significant limitations. First of all, it is restricted to single-word utterances from common aircraft communication vocabulary, which contradicts the idea of analyzing *suprasegmental* speech parameters which require a larger context. Second, two levels of cognitive stress are provoked by computer workload tasks of different complexity; so, although the demand level is set, the actual stress level (which depends on the subject’s response capability, → 1.3.2) is not assessed. There are additional recordings from amusement park rollercoasters and helicopter cockpits which comprise the “actual stress” domain; they consist of the same vocabulary, but have been produced by different speakers and can thus not be compared to any “neutral” reference from that database.

As a consequence, an essential component of this thesis is the conceptualization and realization of a(nother) database with speech under stress. This is presented in chapter 2.

[14] (Anagnostopoulos et al., 2015) argue that the small-scale collections of speech material created for a specific study which is not publicly available should be called a *dataset* instead of a *database*. I adopt their opinion.

1.6. Aim and Contributions of this Work

The aim of this thesis can be broken down into the following three tasks:

TASK 1: Calculate prosodic variables from the speech signal:
a) speech signal → acoustic variables
b) acoustic variables → auditive variables
c) auditive variables → prosodic variables
TASK 2: Find descriptive parameters for the prosodic variables.
TASK 3: Identify which parameters reflect emotions or stress.

I will present 27 linguistically and musically motivated speech parameters as potential descriptors of emotional state or stress level and describe their calculation as well as my motivation for selecting them. Some of these parameters are already common in this field of research, some are refined versions of familiar parameters, and a few parameters are based on completely novel ideas. The latter is especially true for the presented descriptors of *speech rhythm*, which is a complex matter by itself.

In accordance with the scientific parable of *standing on the shoulders of giants*, I have made use of a large number of ideas, concepts and algorithms of others during the accomplishment of the first and the third task listed above. This includes several established methods from the fields of statistics and machine learning, including statistical modeling as well as regression and classification methods. These methods will be briefly introduced in the respective chapters, but not be discussed in detail, as most of you will probably be already familiar with them anyway.

Due to the lack of some required resources or adequate methods, however, I had to develop a few things on my own:

- First of all, a suitable database with speech under cognitive stress had to be created. The creation of the **IEM Pilot Speech Database** is described in chapter 2; it included the recruitment of professional airline pilots and the design of a demanding 3.5 hours flight plan as well as the implementation of a multi-channel recording solution with automatic voice activity detection. The recorded data had to be post-processed, categorized, and synchronized with heart rate measurements which allow to assess the mental state of the speakers. For this purpose, several established parameters of *heart rate variability* were calculated (→ 2.3.2).

- As the several available algorithms for **blind syllable detection** turned out not to work satisfactorily, I had to develop an own solution (→ 3.4.1) which also includes the estimation of **perceived syllable durations**. This was necessary due to the fact that syllable durations are in many cases over-estimated when separating successive syllables by single markers only (→ 3.2.4).

- The calculation of melodic parameters requires a continuous and somewhat smooth pitch contour free from rapid fluctuations in frequency or unvoiced interrupts to match our impression of "speech melody". I have designed a method for the **creation of a continuous pitch contour** from a fragmented fundamental frequency track which includes interpolation, stylization, and smoothing (→ 3.3.2).

- For the calculation of *syllable prominence*, we need its loudness, its perceived length, and its pitch. Due to the dynamic nature of pitch, however, it is a non-trivial task to assign a single pitch value to a syllable in an appropriate way. During my work on this thesis, I came up with a novel method for the estimation of **perceived syllable pitch** depending on the course of the pitch contour during the syllable (→ 3.5.2).

- In this thesis, *rhythm* is interpreted in a musical sense. From the perceptual centers of the syllables and their corresponding prominence values, a **rhythmic grid** is created which forms the basis of several speech rhythm-related parameters (→ 4.3.1).

2. Creating a Database of Speech Under Stress

Unfortunately, no appropriate database of speech under stress is available which could be used for my investigations, so I had to create one. This chapter describes the motivation, the concept and the recording of the *IEM Pilot Speech Database*.

2.1. Fundamental Considerations

2.1.1. Theory ...

A high-quality database of speech under stress has to meet three main requirements: *scope*, *naturalness*, and *context* (Douglas-Cowie et al., 2003).

Scope refers to the number of speakers as well as to the number of different stress types, the number of tokens per stress type, and the gender of speakers. Many available speech databases aspire to the ideal of covering the whole domain of emotions and stress, rather than focusing on a specific sub-region. This leads to a natural complexity which is out of proportion to the number of speakers featured. **Naturalness** means that the recorded speech should be neither *prompted* nor *acted*, but that the speakers are free to choose their words and that they don't have to pretend being in a specific mood or situation. Although acted speech can be reliably classified by listeners, there are still systematic differences between acted and natural speech. Acted speech is often read instead of being spoken freely, which introduces distinctive "reading characteristics" (Johns-Lewis, 1986). In addition, inter-personal effects are not represented in acted speech, as it is often produced in non-interactive monologues (Douglas-Cowie et al., 2003).

Finally, **context** has several aspects. The *semantic* context of stress can be given by signal words indicating excessive demand; signs of stress can also be encoded in the *structural* context, in the sense of following or violating default patterns of intonation or accentuation. Furthermore, if not communicating remotely, we can see our communication partner and provide several kinds of stress-related information through our facial expression, gesture and posture (*intermodal* context). Although these three contextual aspects are not assessed in the later speech analysis if concentrating on prosody and paralinguistics only, they still contribute to naturalness. A fourth aspect of context is the time course of episodes (*temporal* context), since stress does not appear and vanish suddenly, but evolves over time. From a linguistic point of view, I would also highlight that — if a database is split into single chunks of speech — sample length is also a crucial parameter for the applicability of prosodic parameters, since *the way we say something* can hardly be accessed by analysis of single words.

2.1.2. ... and Practice

Having these theoretical aspects in mind and considering potential recording environments where people are used to talking under stressful conditions, the idea was born that airline pilots might be the ideal test subjects, for the following reasons:

- The job profile of an airline pilot is both comprehensive and challenging. Besides high demands on communication, comprehension and technical skills, it is accompanied by heavy responsibility. The recent "Most Stressful Jobs" list of the online job search portal *CareerCast* (CareerCast, 2017) lists airplane pilots on third position, after military personnel and firefighters[1].
- Communication is essential in aviation. The pilots gather information on weather conditions and on the route, they brief the cabin crew before the flight, they regularly communicate with air traffic controllers on the ground, and they process checklists throughout the flight to ensure that all systems are working properly. So, there will be a lot of talking in a regular flight scenario without the need for prompting speech at all.
- Most of the communication is done remotely. Pilots wear a headset consisting of headphones and microphone to keep their hands free for the various controls, so it is always possible to record their speech without adding an intrusive element which is not part of the daily job.

The cockpit of an airplane is a "realistic" environment, and at the same time, we are very close to laboratory conditions concerning the quality of the recorded speech.

[1] The amount of stress which is experienced in a certain job is calculated by rating emotional factors, physical demands, and hazards typically experienced in that occupation.

2.2. Database Creation

2.2.1. The Concept

The main hypothesis is that the presence of cognitive stress leaves traces in speech parameters, so we must design an experiment in which the participants are exposed to actual stress while speaking. As already discussed, the complexity of a task to be solved sets a specific demand level, but we can not draw concrete conclusions on the level of stress experienced by the participants. As a consequence, we need additional parameters which reliably reflect the stress response. Parameters of heart rate variability (HRV) have proven to be reliable indicators of cognitive stress during computer work tasks (Hjortskov et al., 2004) as well as arithmetical tasks (Vuksanović and Gal, 2007; Lackner et al., 2011; Traina et al., 2011; Visnovcova et al., 2014) and of mental stress during academic exams (Tharion et al., 2009; Papousek et al., 2010; Melillo et al., 2011). In addition, they have the advantage that they can be measured with small, portable devices, such that the degree of intrusiveness is minimal.

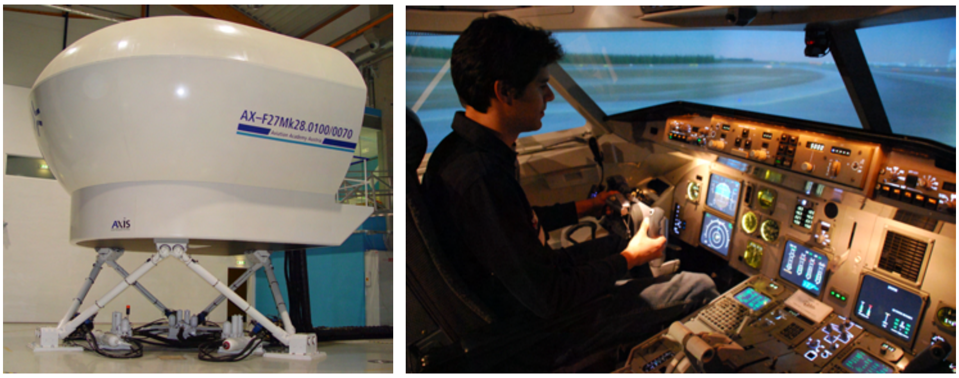
So, the experimental setup is aimed at exposing the pilots to a defined demand at a certain point in time, while recording both their heart rate and their speech simultaneously. The experiments take place in a level D full flight simulator, which is basically a real airplane cockpit mounted on a motion base, to simulate real flight characteristics. Regular airline pilots participate as volunteers in the experiments. A typical working day of a pilot is captured in a flight program of 3.5 hours in length which has been designed by professional flight instructors who graded the expected quality and level of stress by experience. A civil airplane is typically operated by two pilots; the *commander* has the overall responsibility for the safe operation of the airplane, while the *first officer* provides support in all tasks during the flight. The pilots usually take turns in flying to avoid fatigue; while the *pilot flying* operates the controls, the *pilot non-flying* takes care of most external communication tasks and checklist processing. The pilots experience a variety of unexpected events and technical malfunctions during the flights which require short-term decision making. They are free in the way they react to these situations to keep the degree of reality as high as possible.

Both speech recordings and heart rate data are afterwards synchronized, and the speech data are segmented into single utterances based on voice activity detection and manual post-processing. The totality of single speech files and corresponding heart rate data forms a database of pilots' speech under stress which is made publicly available as the **IEM Pilot Speech Database** (*IEM-PSD*) through EUROCONTROL, as mentioned on page v.

2.2.2. Test Setup and Design

Recording Environment

The experiments take place in a Fokker F70/100 series full flight simulator at *Aviation Academy Austria*[2]. The simulator is graded "Level D" which is the highest degree of realism available on ground; it is regularly used by the leading Austrian airline for biannual proficiency checks. All instruments and controls in the cockpit are original, and the pilots look through a glass window on a *collimated* display[3]. The simulator is equipped with a 3D sound system including infrasonic sound, and the motion base system simulates real flight characteristics. The audio channels (headsets and push-to-talk devices in the cockpit, flight instructor's voice) are digitally accessible in a server room outside the simulator, and can thus be captured in a completely "invisible" way.



(a) Exterior view.

(b) The author starting from Vienna Airport.

Figure 2.1.: The full flight simulator at Aviation Academy Austria.

The flight simulation is executed and controlled by a supervisor sitting in a visually separated area in the back of the cockpit who also acts as the air traffic controller communicating with the pilots via radio.

Test Subjects

Eight professional male Fokker F70/100 pilots participate as volunteers in the experiments. All pilots are full-time employees of a regional subsidiary of Austria's leading airline and familiar with the simulator, which is used for type rating tests and biannual proficiency checks. The native language is (Austrian)

[2] Aviation Academy Austria, Neusiedl/See, Austria, www.aviationacademy.at

[3] A collimated display makes sure that both pilots see the world outside the window without angular errors or distortions.

German in seven cases; 1 native Danish pilot is fluent in German. The pilots have been instructed to speak English during formal communication, while being allowed to switch to German during informal talk.

The pilots are grouped into 4 teams of one experienced pilot acting as the commander (*CMDR*) and one less experienced pilot as the first officer (*F/O*). The flight program has a total duration of approximately 3.5 hours and is executed once for each team over a period of 2 consecutive days. Personal statistics of the selected volunteers are given in Tab. 2.1; the *Pilot ID* specifies session number and crew role.

Pilot ID	**Age**	**Prof. Experience**		
		Pilot	F100	as CMDR
CMDR 1	31	11	11	6
F/O 1	35	3	3	—
CMDR 2	45	20	20	12
F/O 2	34	12	5	—
CMDR 3	48	22	5	5
F/O 3	29	10	5	—
CMDR 4	44	16	2.5	2.5
F/O 4	29	11	4	—

Table 2.1.: Personal statistics of pilots participating in the recording sessions (age and experience given in years, resp.)

Responsibilities in the Cockpit

The tasks and responsibilities for the pilots are manifold:

- Before the flight, they carry out several pre-flight checks to make sure that the navigation and operating systems work properly.
- They gather all information on the route and the weather, as well as on the distribution of passengers within the cabin and the total weight of the airplane.
- Based on this information, they create a flight plan including the route to be taken, altitudes during the flight, and the required amount of fuel.
- Before takeoff, they make sure all safety systems are working properly.

- Also during the flight, regular checks on the technical performance of the systems are carried out.
- The pilots regularly communicate with air traffic controllers on the ground regarding weather conditions and air traffic.
- In case of an emergency, the captain needs to decide quickly and appropriately which measures to take.

The Flight Program

The recording session consists of four different challenging scenarios (F1-F4) plus one "reference flight" (F0) at the very beginning. The latter is included for two reasons: first, from a psychological point of view, the pilots have a warm-up flight and can acclimatize to the simulator again. Second, they provide reference values for the three main flight phases of interest; (a) takeoff and initial climb, (b) en route flight, and (c) approach and landing. The whole F0 scenario takes about 20 minutes, and the test subjects know in advance that nothing exceptional will be happening during this warm-up flight.

A *strain trajectory*, graded by experience, has been sketched by the instructors for each of the four demanding scenarios (Fig. 2.2). It visualizes the presumed strain on the pilots as intended by the lesson plan; that is, the timetable of occurring events and malfunctions. The term "strain" is used in this context rather than task load, since the total demand is not only determined by task complexity, but also by mental stress and task-related anxiety.

These four scenarios are implemented as a *Line-Oriented Flight Training* (LOFT), which is a "full mission" simulation of scheduled flights. This includes a full cockpit preparation during the first scenario ("first flight of the day"), as well as intensive air traffic control and cabin crew communication during the flight. A 5-minute rest period after the reference flight shall ensure that the test subjects will have had a sufficient amount of time to fully recover. The following four flights are then simulated in one stretch,

(a) from a temporal point of view; i.e., there are no pauses in between (except for a short break to fill out the mood questionnaires); and
(b) from a local point of view; meaning that a scenario starts at the same airport at which the previous one has ended.

The *pilot flying* role is assigned to the CMDRs during F1 and F2 (the more demanding flights), and to the first officers during F3 and F4. In addition, we provoke verbal communication in an extreme situation during takeoff abortion in F3, as the commander has to tell the first officer verbally to immediately abort takeoff.

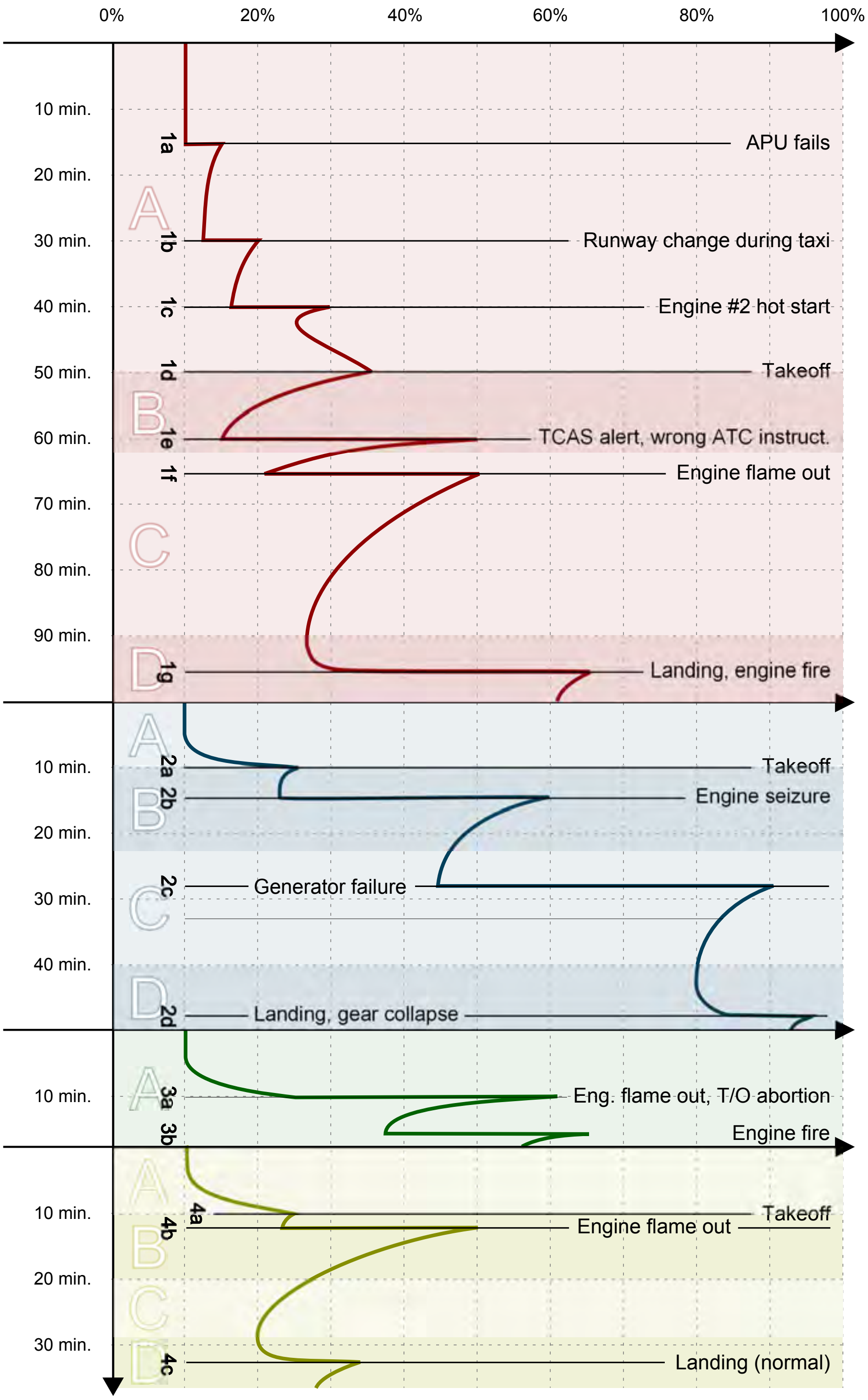


Figure 2.2.: Strain trajectories and events/malfunctions for each of the four demanding scenarios, F1-F4. Marked flight phases are: [A] cockpit preparation, [B] takeoff and initial climb, [C] en route flight, [D] approach and landing.

Audio Recording Setup and Quality

The pilots are asked to communicate via headset only and not to press the “push-to-talk” button when using the radio transceiver or the line to the cabin. All speech signals are pre-amplified and dynamically compressed before being multiplexed into *ADAT* format[4]. Usage of a state-of-the-art professional audio interface (*RME Multiface*) allows simultaneous capturing of all audio channels while still providing the non-modified ADAT signal at the output (“feed-through”). This means that the audible signal which is fed back into the cockpit is not altered in any way. Single channels are recorded in the standard, uncompressed *WAV* format with a sampling rate of 44.1kHz and a resolution of 16bit.

A patch written in the graphical programming language *PureData (pd)* is used to record all relevant channels to hard disk and to create speech activity information data at the same time[5]. The raw speech data furthermore have to be split into single files at a maximum of 100 minutes in length, since the WAV file format does not allow file sizes greater than 2GB. For further details, the interested reader is referred to the *IEM-PSD* technical report (Luig, 2011).

Although the speech recording quality conforms to compact disk (CD) standards, the “perceived quality” of the speech recordings may be lowered by the fact that the pilots speak through standard headset microphones which they are allowed to adjust at will. Common artifacts thus include sound level variability and distortion of consonants, especially plosives or fricatives. There is audible crosstalk (speaker B recorded through speaker A’s microphone), but at a sufficiently lower level compared to the primary source (speaker A). This is considered to be tolerable, since the main objective for creating this database is the maximum degree of reality rather than making high-fidelity speech recordings.

Communication and Speech Data

All external dialog partners are simulated by the instructor sitting in the back of the simulator, invisible to the pilots. Also speaking through a headset microphone, his voice reaches the pilots’ ears through a band-limited communication channel. The instructor was asked to keep an unagitated tone when acting as an air traffic controller, as a technician or member of the cabin crew.

[4] The ADAT signal is a multichannel optical signal transmitted via glass fibre.

[5] This perfectly synchronized speech activity log considerably facilitates subsequent data segmentation into single utterances.

My personal impression is that, depending on the actual context, the recorded speech data show considerable variations in terms of melody, rhythm, expressiveness and utterance length: Air traffic control (ATC) communication shows reduced prosodic intensity, as its focus is on clarity and transmission of information; whereas a discussion on what further action is required may be held in a more “emotional” way. To facilitate any kind of further analysis, each single speech file is thus labeled with one of the following categories:

ATC Communication	radio communication with ground control, radar etc.
Cabin Address	captain’s address to passengers
Checking Procedures	one- or two-word communication (*“eighty knots”, “my controls”*)
Cockpit Communication	instructions, questions, discussions etc.
Free Speech	all conversation not related to controlling the aircraft or handling a situation (jokes, everyday talk)

Table 2.2.: Speech file categories for *IEM-PSD* speech data.

Time Grids

To allow for the highest possible degree of realism, the pilots are completely free in their course of action. As a consequence, the unexpected events and malfunctions occur at different points in time (relative to the session start time). The Traffic Alert and Collision Avoidance System (TCAS), for example, should give the alarm eight minutes after takeoff, during the initial climb phase. Such events are triggered by the flight instructor from the back of the cockpit.

Since these times at which the events occur are of great importance for event-based speech analysis, a detailed event log is created manually, and the database documentation contains a list with all times of relevant events as marked in Fig. 2.2.

2.3. Measures of Heart Rate Variability

2.3.1. Heart Rate and the Autonomic Nervous System

Our autonomic nervous system (ANS) regulates essential bodily functions as heart rate, respiration or digestion, to name just a few. Heart rate is thus a valid quantitative marker of autonomic activity, and it is popular in stress research due to because it can be measured in a non-invasive way.

Figure 2.3 on page 39 displays the heart rate of all four commanders over time during flight 2, which is considered the most demanding flight in this LOFT scenario. The instantaneous heart rate, sampled at f_s = 4Hz, is represented by the fine gray line; a smoothed version (50-samples moving average filter) is displayed as a dark solid line for clarity. Four remarkable events are indicated be vertical lines, which are takeoff (*T/O*), engine seizure (*ENG*), generator failure (*GEN*), and landing with gear collapse (*LAND*). The time scale has been normalized, since the length of the flights varies as the crew is relatively free in their actions.

The immediate reactions to the single events are clearly visible in the heart rate, as well as "anticipatory increases" before takeoff and landing. However, once the necessary measures for coping with the changed situation have been taken, a significant decrease in heart rate indicates that the subjects have settled down again. Over the whole flight of approximately 50 minutes, the heart rate recordings show a consistent positive trend for all four commanders which range between [14..32] bpm/hour.

But recordings of heart beats are just the basic signal for physiological analyses, comparable with the acoustic pressure fluctuations recorded by a microphone. To quantify what we perceive as qualities of a sound, we extract a small chunk of audio and perform Fourier analysis to estimate which frequencies are present in the signal, or we average over all values within the analysis window to estimate the intensity of the sound. In a similar manner, there are several well-established measures in cardiovascular physiology to describe the function of the autonomic nervous system. The ANS consists of two complementary divisions, the *sympathetic* and the *parasympathetic* nervous system. While the sympathetic nervous system primarily controls the bodily reactions to all kinds of stressors ("fight-or-flight response"), the parasympathetic nervous system stimulates so-called "rest-and-digest" activities when the body is at rest. Both divisions are always active, but at different levels (McCorry, 2007) which can, for example, be assessed using frequency-domain methods.

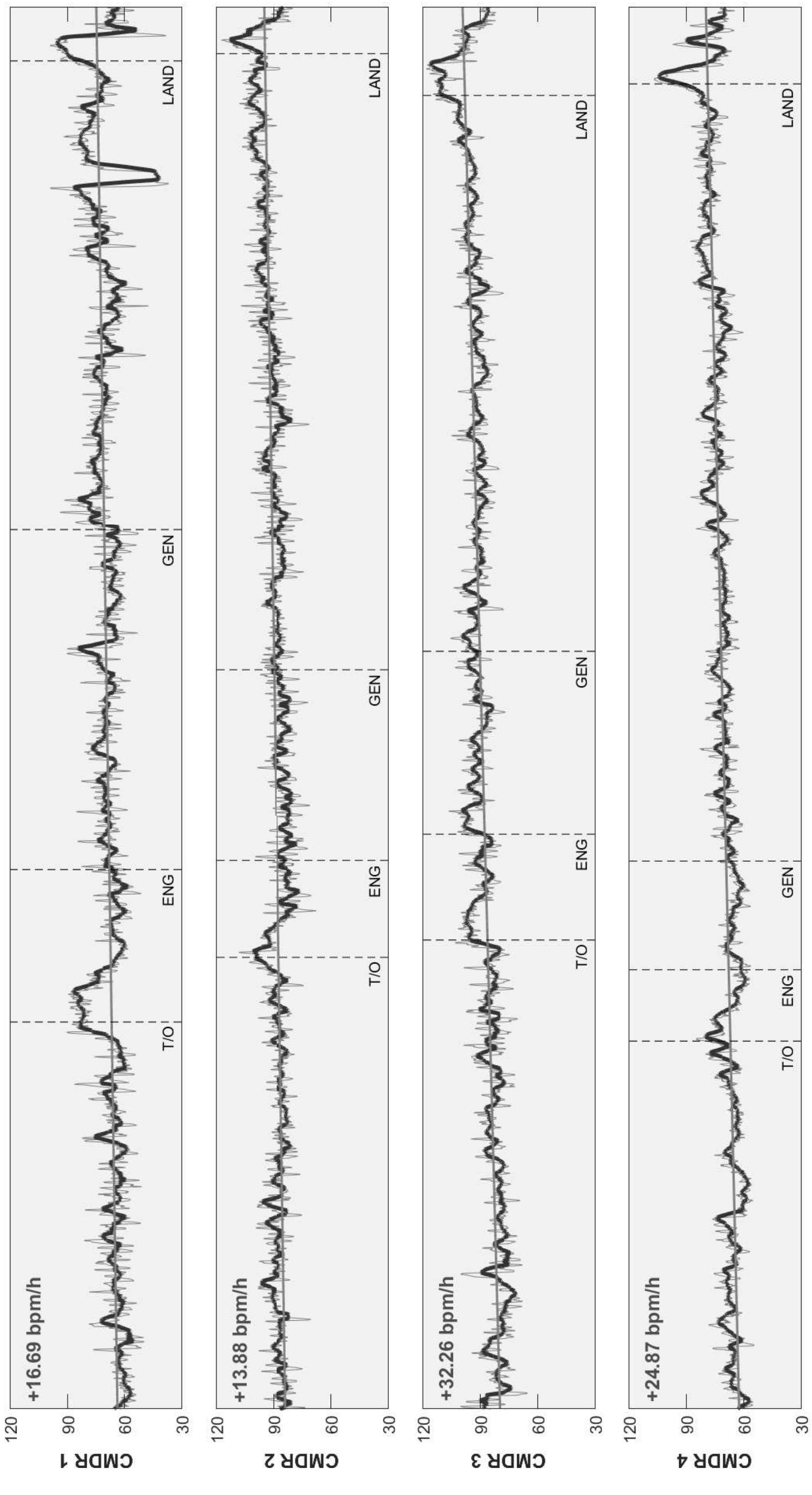


Figure 2.3.: Heart rate recordings of the four commanders during flight 2. Marked events are takeoff (T/O), engine seizure (ENG), generator failure (GEN), and landing with gear collapse (LAND).

2.3.2. Heart Rate Variability

All these heart rate parameters are based on measurements of the interval between consecutive heart beats and are generally referred to as *heart rate variability* (HRV), although not all of them are measures of variance.

The characteristic shape of a electrocardiogram (ECG) is determined by different overlaid waves which are denoted by the letters *P*, *Q*, *R*, *S*, and *T*. The remarkable peak in the waveform is caused by the *R* wave. Since it is easy to detect, the interval between successive *R* peaks ("RR interval") is the common measure for a heart period.

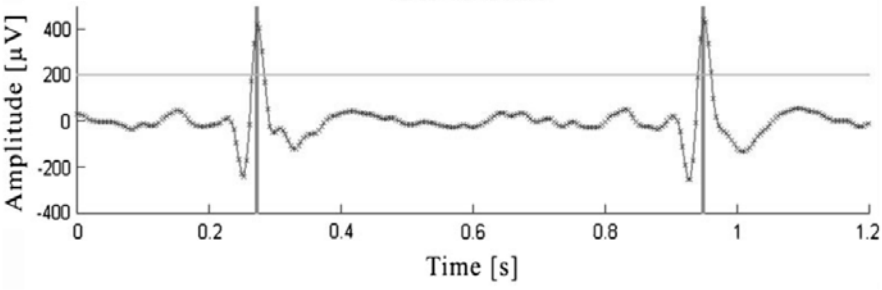


Figure 2.4.: RR interval detection (from (Kaufmann et al., 2011)): the *R* peaks are prominent enough to be robustly detected with a global threshold.

Changes in heart rate variability are generally an indicator that more energy is needed to prepare an appropriate response to a stimulus. To arrive at something that could be called *physiologic state*, it is common to analyze HRV parameters over an observation period of 5 minutes (Malik et al., 1996). HRV parameters thus do not serve as indicators of stress reactions, but rather describe a steady state of autonomic balance.

In the following, I will briefly present the HRV parameters which are calculated from the pilots' heart rate measurements as a reference for the actual stress level.

- As a basic parameter, the **average heart rate** in beats per minute is calculated.
- The simplest, but yet popular parameter is the **standard deviation of RR intervals** over time windows of 5 minutes. Castaldo et al. (2015) reviewed existing studies examining correlations between HRV parameters and mental stress and found that in the vast majority of studies, the *SDRR* parameter[6] was decreased under stress.

[6] The RR interval is synonymously called as the *NN interval* (for "normal-to-normal" heart beats), such that this parameter is also known as *SDNN*.

- The phenomenon that heart rate increases during inspiration and decreases during expiration is called *respiratory sinus arrhythmia* (RSA); our heart rate is modulated in frequency by respiration. This facilitates the recording of the **respiratory rate** from the HRV signal, as well as the calculation of other physiologically meaningful parameters such as the **degree of modulation**, and the **pulse/respiration ratio** which is the quotient of heart rate and respiration rate.

- The balance between sympathetic and parasympathetic nervous system is measured by analyzing the *power spectral density* in two characteristic frequency bands (LF: [0.04..0.15]Hz, HF: [0.15..0.4]Hz) and calculating the **LF/HF ratio** as a measure of vegetative activation level Malik et al. (1996).

- Following an approach by Bettermann et al. (1999), the algebraic sign of the HRV derivative (that is, $sign(d\text{HRV}/dt)$) is encoded binary, leading to a pattern of zeros and ones which can be interpreted as a rhythmic pattern. A distinctive predominance of certain pattern classes corresponds to a small number of different **rhythmical HRV patterns** and thus serves as an indicator of cardiac regularity.

2.3.3. Measurement Device and Data Analysis

The participants' beat-to-beat heart rate signal is recorded with a high-precision mobile measurement device, the ChronoCord (8kHz sampling rate, 16bit resolution). All data are stored on a flash card and analyzed offline. The ChronoCord offers the opportunity to set marker flags at the push of a button, so that heart rate and speech data can easily be synchronized afterwards by setting an "acoustical marker" while pushing the button at the same time. Such a marker was set by the pilots on request at the beginning and at the end of each of the four flights.

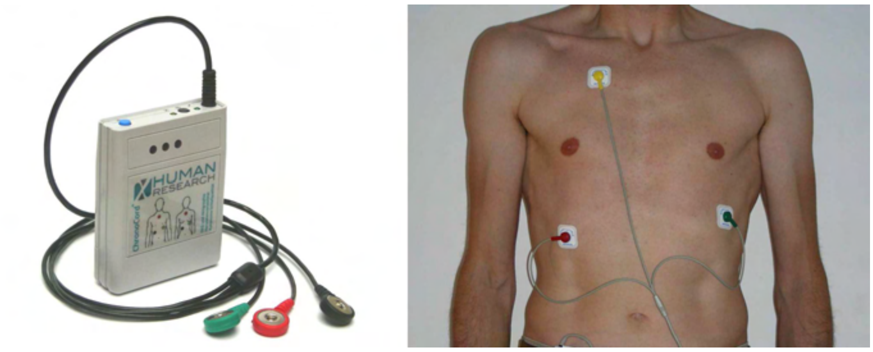


Figure 2.5.: The ChronoCord device (left), electrode placement on a male body (right).

The raw heart rate data are further processed and analyzed manually by professional physiologists[7], including outlier detection and correction. Finally, the data are resampled on a regular time grid with $f_s = 4\text{Hz}$.

[7] HUMAN RESEARCH Institute, Weiz, Austria, www.humanresearch.at.

3. Quantifying Prosodic Variables

This chapter deals with the task to derive prosodic variables from the acoustic signal. For this purpose, its fundamental frequency and intensity are extracted. These acoustic variables serve as a basis for the calculation of auditive variables using models of human sound perception. Finally, we apply linguistic knowledge to derive prosodic variables.
Applying these techniques and models requires profound knowledge of the fundamental processes of speech production and perception as well as essential signal processing methods. Both are given at the outset of this chapter.

3.1. Speech Production

The process of speech production uses more motor fibers than any other human mechanical activity does (Kent, 2000). The resulting speech signal is thus a very complex sound mixture which carries a lot of information. To understand what we might be able to find in the signal, we should have a look at how speech sounds are produced and how emotions or stress could affect the speech production process.

3.1.1. The Human Speech Production System

The human speech production system, as sketched in Fig. 3.1, consists of two major parts which have complementary functions. The organs of **phonation** include the *lungs* and the *larynx*. On the larynx, we have two folds of skin, the *vocal cords*, which blow apart and come together as we force air through the *glottis* between them. This oscillation, driven by the sub-glottal air pressure produced by the lungs, is the basic sound signal. The air pressure and the oscillation rate determine the intensity and the fundamental frequency of the speech sound. The organs of **articulation** include the cavities of both *oral* and *nasal tract* as well as the *lips*, the *tongue*, the *jaw* and the *velum*. They add

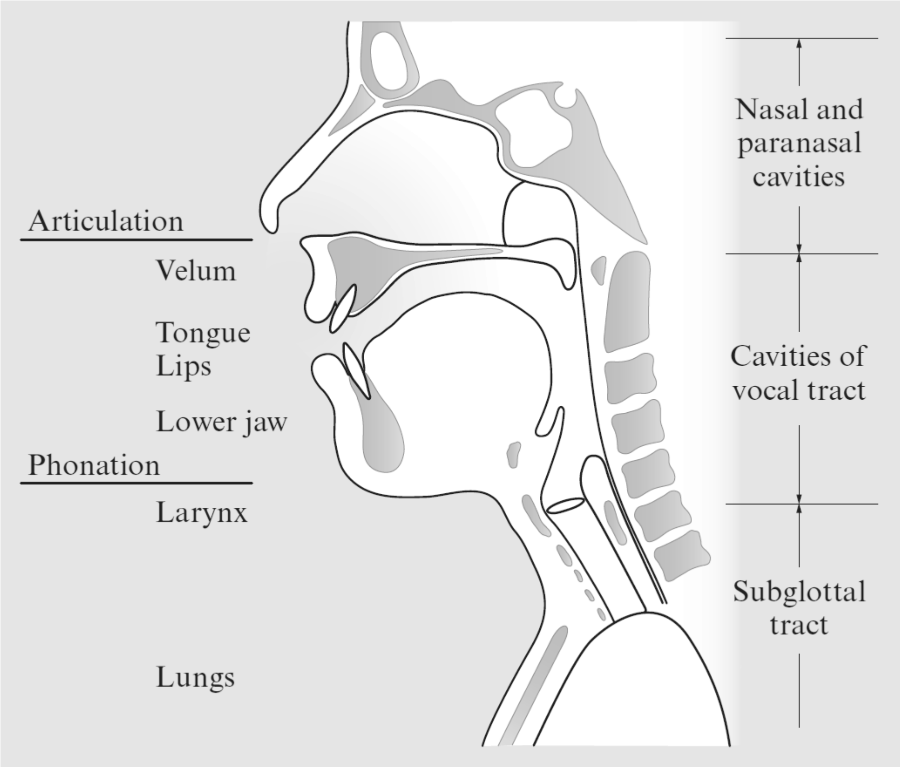


Figure 3.1.: Sketch of the human speech production system (from Honda (2008))

resonances or modulations to the basic sound signal and produce additional sounds for some consonants (cp. Honda (2008)).

From an acoustic point of view, the oral tract is an irregular tube between larynx and lips, whose volume and cross-sectional area can be varied by the muscles controlling lips, tongue, jaw and velum. The velum works like a movable flap which controls the acoustic coupling between the oral tract and the nasal tract, which on the other hand is an acoustic tube of fixed volume and length. These tubular shapes produce characteristic resonances which we call **formants**. Formants determine the identity of different vowels and diphthongs; in other words, the difference between an “aaah” and an “oooh” sound with the same pitch is due to the different formant structures.

The phonation process as described so far is valid for **voiced sounds** (such as [a] or [m]) only. When the glottis is opened, but the vocal cords do not vibrate, we are producing **unvoiced sounds**. Unvoiced sounds can be either *aspirated* or *fricative*:

- Aspirated sounds (such as the [h] in “house”) are caused by airflow turbulences at the partly opened glottis. These turbulences produce a random noise sound which is modulated by the articulatory system.

- Fricative sounds (such as the [s] in “sea” or the [ʃ] in “dish”) are caused by turbulences in the vocal tract due to a constriction of the tube close to the mouth.

Finally, we are able to produce **plosive sounds** (such as the [p] in “pin”) by closing the vocal tract completely for a small amount of time, allowing air pressure to build up before suddenly releasing it by opening the mouth.

The Source-Filter Model

A common, simplified model of speech production is the *source-filter model* which assumes the glottal excitation source to be linearly separable from the transmission characteristics of the vocal tract. Depending on the kind of sound produced (voiced or unvoiced), the excitation signal is modeled either by a sequence of equidistant pulses, a random-noise generator, or a mixture of both[1].

As shown in Fig. 3.2, this excitation signal is fed through a resonance filter which models the vocal tract by emphasizing certain frequencies corresponding to the formants. Finally, radiation characteristics from the mouth are modeled by applying a frequency-dependent gain of +6dB per octave[2].

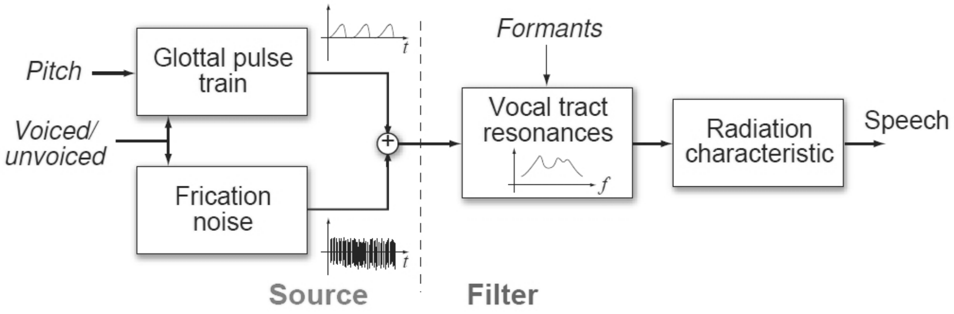


Figure 3.2.: The source-filter model (from Ellis (2006))

This model forms the theoretical basis for the diverse speech analysis techniques, of which the most important are presented in section 3.1.2.

[1] For voiced fricatives, such as the [ʤ] sound in “jealous”, we produce a mixed excitation signal.

[2] The harmonic spectrum of the excitation source (which consists of the fundamental frequency and integer multiples of f_0) decreases in amplitude with increasing frequency at a rate of around -12dB/octave, such that voiced speech sounds produced by the source-filter model will show a spectral slope of -6dB/octave.

How Can Emotions or Stress Affect Speech Production?

The speech production process as described above covers the physical domain only. The whole picture, however, includes a cognitive component as well (this is the upper area in Fig. 3.3): speech production starts with an idea what to say, followed by the creation of an appropriate sentence in terms of the sequence of sounds to be uttered and the prosodic realization. This *language code* is then translated into appropriate neuromuscular commands which are further transmitted to the muscles which control the respiratory system and the vocal tract.

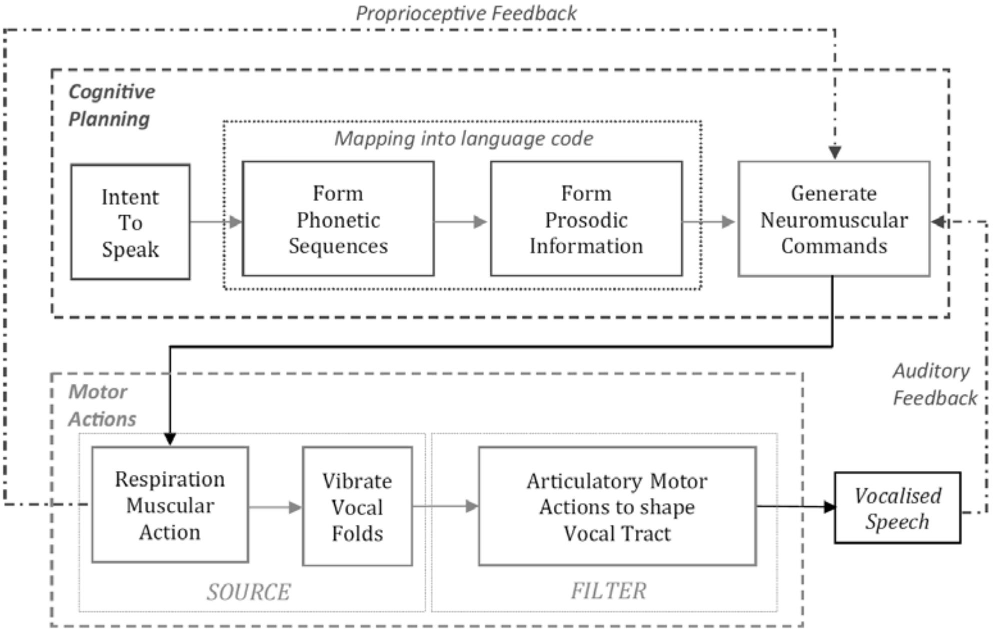

Proprioceptive Feedback
Cognitive Planning
Mapping into language code
Intent To Speak
Form Phonetic Sequences
Form Prosodic Information
Generate Neuromuscular Commands
Auditory Feedback
Motor Actions
Respiration Muscular Action
Vibrate Vocal Folds
Articulatory Motor Actions to shape Vocal Tract
Vocalised Speech
SOURCE
FILTER

Figure 3.3.: Schematic diagram of speech production (from Cummins et al. (2015))

The British psychologists Baddeley and Hitch (1974) developed a model of human *working memory* which describes the system that enables us to keep things in mind while performing complex tasks like reasoning or comprehension. Their model consists of a central attentional control system which is supported by two short-term storage systems: one for visual material, which is called the *visuo-spatial sketchpad*, and one for acoustic and verbal material, which has been named the **phonological loop** (Baddeley, 2003). The phonological loop helps to control the articulatory system and to store verbal information for a few seconds. Reduced cognitive capacity due to external influences does not only affect ideation, but also impacts the generation of neuromuscular commands as well as the proprioceptive feedback loop (Krajewski et al., 2012).

These external influences can be stressors of second or third order (cp. Tab. 1.1 on page 14 — note that, in this taxonomy, emotions act as third-order stressors).

- **Third-order stressors** affect the speech production process at its highest level. Excessive demand due to high workload or highly negative emotional states like anxiety or fear may affect the ideation process (Levelt, 1999).

- **Second-order stressors** have impact on the conversion of language code into neuromuscular commands, with noise being the most prominent stressor. The term *perceptual stressor* indicates that there is some kind of conscious interpretation of the stressor (Murray et al., 1996), but without involving higher-level emotions.

- **First-order stressors**, on the contrary, step in at the link between cognitive planning and motor actions by unconsciously modifying the neuromuscular signal transduction process and thus provoking changes in articulator movements; 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.

In stressful situations, we tend to increase our respiration rate which causes an increase in subglottal pressure during speech (Rajasekaran et al., 1986). This will, on the one hand, lead to an increased pitch; on the other hand, it will also affect the rhythmic pattern when the same amount of words is to be produced within shorter time windows between consecutive breaths. For decision-making under time pressure as well as in noisy environments, we tend to raise our vocal effort to make ourselves heard (Hansen and Patil, 2007). This will presumably not only be reflected in average intensity over the utterance, but also in particular emphasis of the more important syllables. Since the capacity of the lungs is limited, other syllables will have to be de-emphasized in compensation; in other words, the prominence pattern of the utterance would gain in dynamics.

Summing up, there is considerable evidence that emotions and stress affect the speech signal in many ways. In the next section, we will see in which ways useful information can be extracted from the recorded speech signal as a basis for the calculation of prosodic variables.

3.1.2. Analyzing Speech

Speech is recorded with microphones. A microphone is a transducer which converts the air pressure variations of a sound wave into an electrical signal. This signal is nowadays commonly *digitized*, resulting in a sequence of numbers which represent the air pressure in equidistant time steps. This representation is referred to as the **time domain** which contains valuable information on the temporal occurrence of sound events and their intensity. To determine which frequencies are present in a sound sample, the signal has to be transformed into the **frequency domain** using the *Fourier transform*, yielding the *spectrum* of the sound. Fig. 3.4 exemplarily shows time- and frequency-domain representations of different phonemes[3].

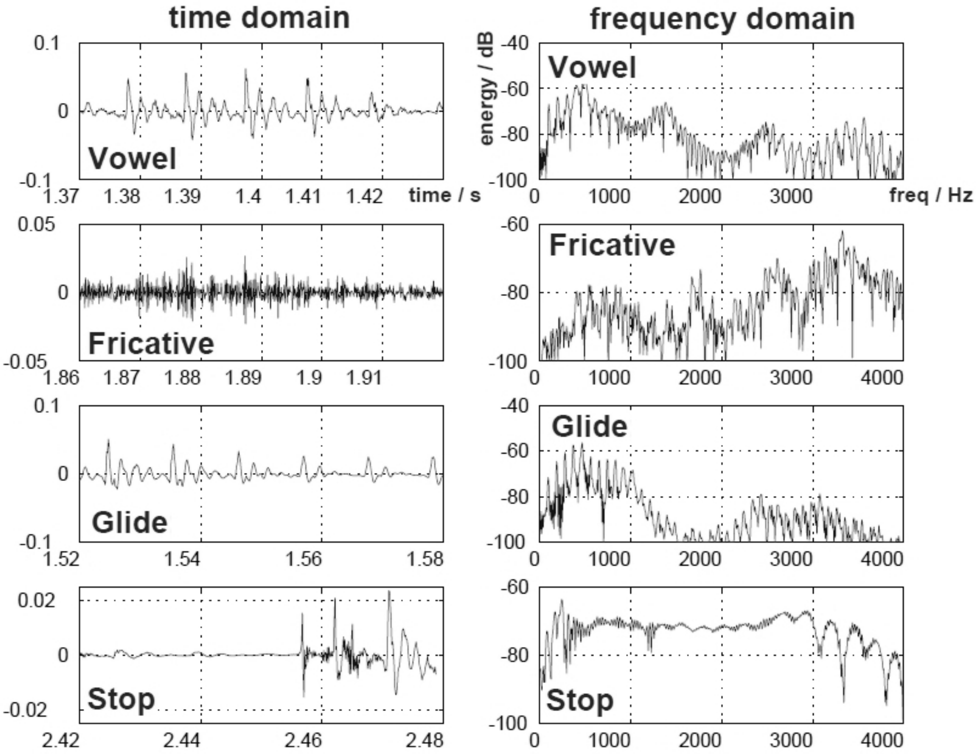


Figure 3.4.: Different speech sounds in time (left) and frequency domain (right). (From Ellis (2006), slightly modified)

As we can see here, the different elementary sounds of speech look completely different in both time and frequency domain.

[3] The Fourier transform results in a complex spectrum which contains both magnitude and phase information. However, for our needs, the *magnitude spectrum* contains all necessary information; so whenever you read the term "spectrum" without further specification in this thesis, I am talking about the magnitude spectrum.

- **Vowels** are characterized by a regular, periodic course over time, as visible in the topmost left plot. The period is approximately 0.1 seconds, which corresponds to an f_0 of about 100Hz. In the spectral representation on the right side, the first three formants are visible. With $F_1 \approx$ 500Hz and $F_2 \approx$ 1500Hz, this vowel is likely to be a German "ö" sound ([ø] in IPA[4] notation).

- The second row shows both time- and frequency-domain examples for a **fricative** sound as the [ʃ] in "dish". There is apparently no periodicity visible in the time-domain plot, and the spectrum is dominated by high-frequency parts, contributing to the *sharpness* of this sound[5].

- A **glide**, as visualized in the third row, is a voiced sound in which the airflow is *gliding* over the tongue before exiting the lips; think of the [w] in "why". Glides are also called *semi-vowels*, because they act as consonants before or after vowels; they also show a certain degree of periodicity, but are "transitional" in nature.

- In the bottom row, the nature of a **plosive** (sometimes also called *stop*) becomes obvious — this could be a [p] as in "pin", for example. The time-domain representation reveals the three stages of a plosive sound, which are catch–hold–release.

The examples in Fig. 3.4 have been manually extracted, so the temporal and spectral representations are only valid for these short extractions of a speech signal. If we applied the Fourier transform to longer speech signals, we would get a smeared spectrum which contains overlaid information from many different speech sounds. In order to track spectral information over time, we can instead apply the *Short-Time Fourier Transform* (STFT) which effectively calculates individual spectra for successive short excerpts of the signal. Its visualization, the *spectrogram*, shows the spectral amplitude both over frequency (ordinate axis) and time (abscissa axis), as shown in Fig. 3.5. Amplitude values are coded as colors or gray-scale values; in the example shown here, high amplitude values are white while low values are black. The fundamental frequency as well as the formants (if present) are vaguely visible as bright horizontal lines during voiced parts.

The single speech chunks are extracted from the speech signal using a *window function* which is zero outside a chosen interval and non-zero inside. Depending on the length of the extracted speech chunks, either temporal or

[4] IPA is short for *International Phonetic Alphabete*.

[5] More on sharpness can be found in section 4.4.2. Have you noticed that "sharpness" itself starts with a *sharp* sound?

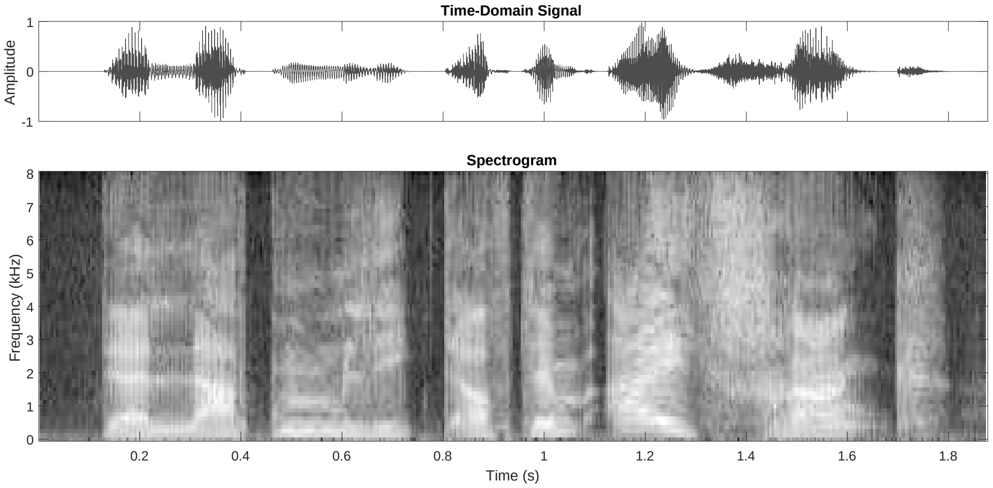


Figure 3.5.: Time-domain representation of speech signal (top), and corresponding spectrogram (bottom). High amplitude values are marked white.

spectral signal characteristics are emphasized: the longer the analysis window, the greater the number of spectral bins[6], and the lower the lowest frequency which can be captured by the Fourier transform. On the other hand, the shorter the analysis window, the greater the *temporal resolution*; i.e., the more accurate the information in the time domain. Depending on the purpose of analysis, different window lengths are reasonable; in general, a window length of about 20ms has established itself as a good choice for speech analysis since speech sounds can be assumed as wide-sense stationary over this period. For f_0 tracking, on the other hand, the window must be long enough to capture frequencies down to 50Hz for male voices, and if a non-rectangular window is used, at least 3 fundamental periods should fit into the window as a rule-of-thumb. One can achieve an adequate step size nonetheless by shifting the analysis window just by, e.g., 10ms in each time step while maintaining a window length of, e.g., 60ms[7].

[6] The spectral sampling points are called *bins*. The discrete-frequency spectral energy, ranging from $0Hz$ to $\frac{f_s}{2}Hz$, is quantized equally into N bins. The number of bins is often referred to as the *spectral resolution*.

[7] The step size is the “update rate” of the spectral information, while the temporal resolution is still determined by the window length.

3.2. Speech Perception

3.2.1. The Human Ear

The prosodic model presented in section 1.2.3 includes acoustic, auditive and prosodic variables. *Auditive* variables are the perceived equivalents of their acoustic counterparts which are calculated based on models of human sound perception. To understand the nonlinear relationship between the acoustic and the auditive variables, let's have a look at Fig. 3.6 and follow a sound through the ear.

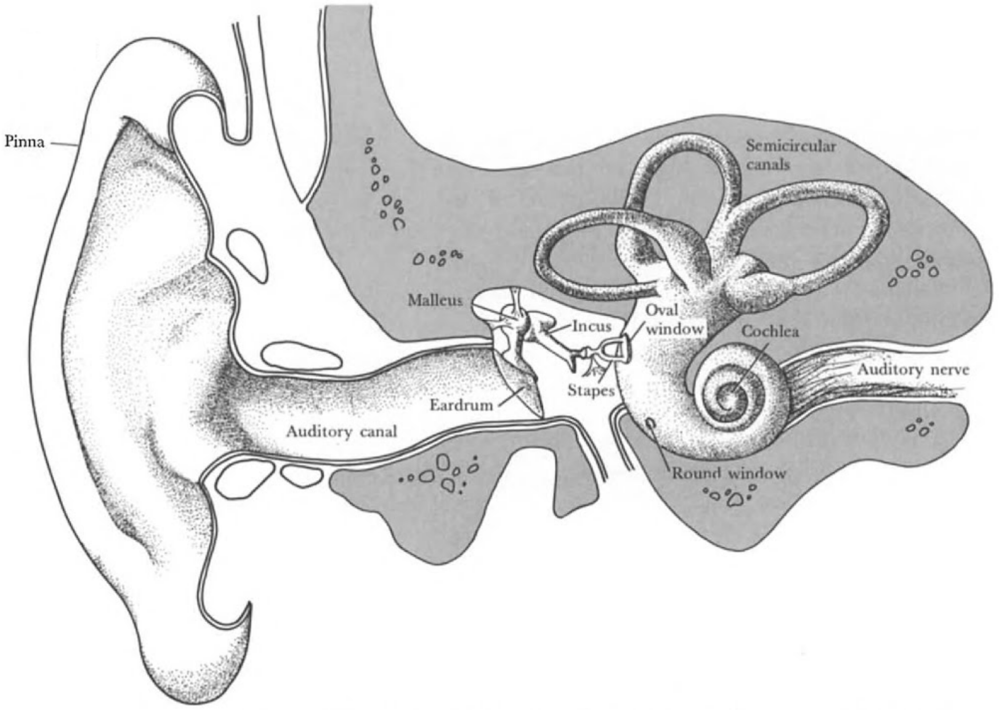


Figure 3.6.: Physiology of the human ear (from Lindsay and Norman (1977))

What we perceive as "sound" starts with the propagation of a pressure wave through the air. Having been collected by the pinna and having traveled through the auditory canal, this pressure wave causes the eardrum to vibrate. These vibrations are transported by the bones of the *middle ear* to the oval window, which marks the transition to the *inner ear*. The vibrating oval window brings the fluids in the cochlea into motion, which in turn produces vibrations on the basilar membrane which is located inside the winding part of the cochlea. The basilar membrane is covered with tiny hair cells which are each connected to an auditory nerve fiber. Depending on the frequency of the incoming sound wave, a certain region of the basilar membrane is stimulated.

The frequency resolution of our auditory system is limited with regard to its discrimination ability between different sounds. Fletcher (1940) discovered that the cochlea behaves like a bank of overlapping bandpass filters where frequency interaction phenomena within one bandpass filter are evaluated differently than when they exceed the bandwidth of such an auditory filter. These so-called **critical bands** play a major role in sound perception, as they influence the sensing of loudness (→ 3.2.3), roughness (→ 4.4.1), and sharpness (→ 4.4.2).

So, the process of hearing involves sound propagation in the air as well as in solid bodies and in liquids; and involves a frequency-to-location transformation on the basilar membrane. The nonlinear effects emerging along this transmission chain can be captured by small microphones placed inside the auditor canal, through the calculation of transfer functions between outer and inner ear, or by adequately designed listening tests.

In the following two sections, I will summarize the relevant findings from the literature concerning pitch and loudness perception. A third section is concerned with the perception of syllable timing and syllable length, which builds not only on the physiology of the human ear, but also on the cerebral evaluation of the neural stimuli.

3.2.2. Pitch

The harmonic structure of voiced speech signals is produced by a periodic opening and closing of the vocal folds. The inverse of the period is the *fundamental frequency* of speech, often abbreviated as f_0. There is a nonlinear relationship between this acoustic variable and its auditive equivalent which we call *pitch*.

The first attempt to construct a subjective scale for the measurement of pitch was made by Stevens et al. (1937) who asked test persons to determine the "half-value" of pitches at various frequencies, aiming towards a measure which doubles in values when a sound is perceived as "twice as high". The resulting **mel scale** (taken from the root of the word *melody*) approximately shows a linear behavior up to 500 Hz and a logarithmic behavior above 500 Hz. This means that, e.g., the musical tone a‴ at 1720 Hz — two octaves above a′ at 440 Hz, which is a quadruplication in frequency — is perceived only 2.5 times as high as a′.

More extensive experiments on the subject resulted in alternative calculation formulas for subjective pitch, including those of Fant (1968), Makhoul and

Cosell (1976), and Lindsay and Norman (1977). A comparison of these notations, shown in Fig. 3.7, reveals that they are very similar in their global behavior; especially in the interesting frequency range for human speech.

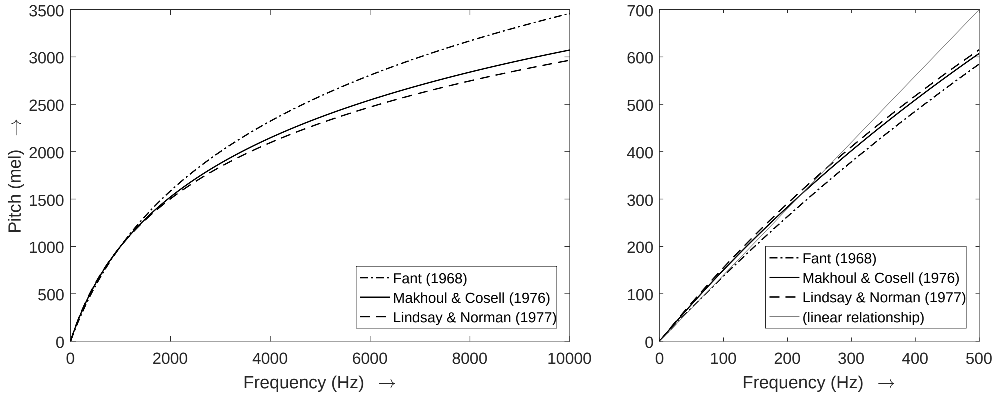


Figure 3.7.: Comparison of several pitch scales as a function of fundamental frequency: wide-band frequency range (left) and interesting frequency range for f_0 in human speech (right).

All pitch values used for intonation and prominence calculations in this thesis employ the widely-used formula published by Makhoul and Cosell (1976):

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) . \tag{3.2.1}$$

3.2.3. Loudness

Loudness is a subjective quantity corresponding to the perceived sound intensity. Loudness perception is by far not just a function of sound intensity, but also depends on frequency, bandwidth, duration, input direction (frontal vs. lateral), and possible temporal masking effects (Zwicker and Fastl, 1999). In this section, I try to summarize the most relevant aspects of sound intensity perception.

Sound Intensity and Sound Pressure

Sound intensity corresponds to the energy carried by a sound wave which depends both on the sound pressure p and the particle velocity $\mathbf{v}$:

$$\mathbf{I} = p_{eff} \cdot \mathbf{v_{eff}} . \tag{3.2.2}$$

For plane waves, sound pressure and particle velocity are related via the specific acoustic impedance ($z = p/v \Leftrightarrow v = p/z$), such that sound intensity is proportional to the square of the sound pressure ($I \propto p^2$). The human hearing system is able to detect pressure variations of just a few micropascals, while the threshold of pain for very loud sounds is approximately 100Pa. This enormous range can best be captured by a logarithmic measure, so the **sound pressure level** L_p has been defined as

$$L_p = 10 \log_{10} \left(\frac{p^2}{{p_0}^2} \right) = 10 \log_{10} \left(\frac{p}{p_0} \right)^2 = 20 \log_{10} \left(\frac{p}{p_0} \right) \quad \ldots \text{[dB SPL]} \qquad (3.2.3)$$

where the reference pressure $p_0 = 2 \cdot 10^{-5}$Pa approximates the threshold of audibility. In the same manner, a reference sound intensity has been defined as $I_0 = 1 \cdot 10^{-12} \frac{\mathrm{W}}{\mathrm{m}^2}$, such that

$$L_I = 10 \log_{10} \left(\frac{I}{I_0} \right) \equiv L_p \; . \qquad (3.2.4)$$

In acoustics, the usage of sound pressure level is common; presumably due to the fact that microphones capture variations in pressure rather than in intensity. The unit *dB SPL* indicates that the value is an absolute value with respect to the reference pressure p_0.

Loudness Level

The auditory system does not have a flat frequency response, which means that some frequencies are more attenuated than others. The results of several psychoacoustic experiments have led to the formulation of *equal-loudness contours* (Fig. 3.8) which demonstrate the effect of frequency on loudness perception[8].

Loudness level is measured in *phons*. By definition, 40 phons are the loudness level of a 1 kHz pure steady tone at 40dB SPL, and the corresponding 40-phons curve indicates sound pressure levels leading to the same loudness level as a function of frequency. The threshold of audibility also has an equal-loudness contour which is the 3-phons contour by definition[9]. All equal-loudness contours clearly show that our ears are most sensitive in the area between 2kHz and 5kHz. Towards both ends of the audible frequency range, the sound pressure levels needed to maintain a certain loudness level increase exponentially.

[8] Loudness level is also a function of direction. These curves in Fig. 3.8 are valid for a plane sound field; Zwicker and Fastl (1999) have calculated a frequency-dependent correction curve for diffuse sound fields; but this fact is not of relevance for speech analysis.

[9] The threshold in quiet at 1kHz is 3dB SPL and not 0dB SPL, so the threshold of audibility is indicated by the 3-phons curve for the sake of consistency.

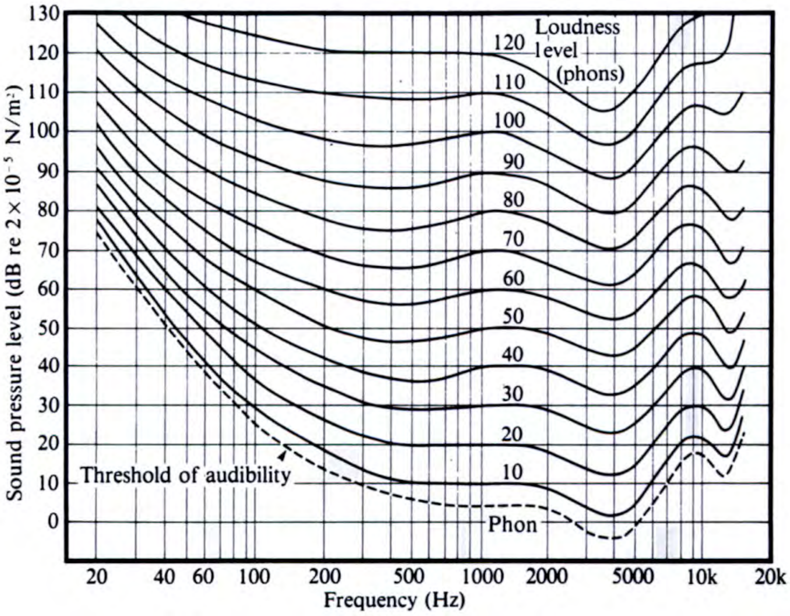


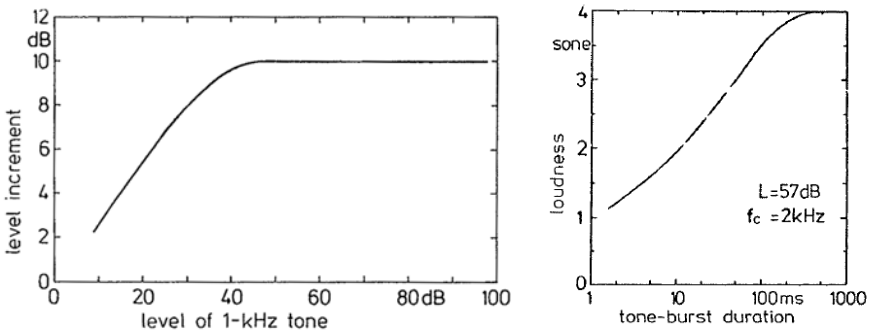
Figure 3.8.: Equal loudness contours (from Hartmann (2004))

Loudness

Loudness is a *ratio quantity*, meaning that that one sound can be, for example, "twice as loud" as another. Stevens (1936) introduced an exponential function to calculate the respective loudness N (in *sone*) for a certain loudness level L_N (in dB SPL), where a halving or doubling in sones corresponds to a halving or doubling in perceived loudness and 1sone was defined as the loudness of a 40-phons sound. The quintessence of many experiments is that the level of a 1-kHz tone in free field has to be incremented by 10dB SPL to achieve a doubling of perceived loudness, and vice versa, that a decrement by 10dB SPL results in a halving of the loudness impression. This is especially true for values larger than 40dB SPL; below that sound pressure level, the level difference which is necessary to achieve a loudness doubling or halving becomes smaller (Zwicker and Fastl, 1999). This relationship is sketched in Fig. 3.9 a).

The impression of loudness is also dependent on the number of critical bands covered by the sound. When we hear a narrow-band noise signal and increase its bandwidth step by step while maintaining the overall energy, the loudness impression remains constant up to a certain point, which is the *critical bandwidth* (→ 3.2.1). From this point, the loudness increases as a function of its bandwidth. In other words, our hearing system evaluates the contributions of a sound within the critical band differently than it does outside the critical band.

If we are talking about transient (i.e., not steady-state) sounds, the duration of the sound also contributes to loudness perception; is it constant for durations greater than 100ms and decreases for shorter durations, as displayed in Fig. 3.9 b).



(a) Level increment needed for doubling of perceived loudness as a function of level.

(b) Loudness as a function of sound duration.

Figure 3.9.: Further sound properties which affect loudness perception (from Zwicker and Fastl (1999))

Another important contribution to the perception of loudness are both temporal and spectral masking effects, which are, however, not relevant when considering a single speaker in a quiet environment.

(Zwicker, 1982) formulated an algorithm for loudness calculation which has been standardized in an ISO norm based on values of **specific loudness** in single critical bands. In a first step, the excitation of the basilar membrane is calculated from the intensity of the acoustic signal by considering the response characteristics of the outer and middle ear as well as of the cochlea. The specific loudness N' is then calculated as a function of the excitation E with (Zwicker and Fastl, 1999):

$$N' = 0.08\left(\frac{E_{TQ}}{E_0}\right)^{0.23}\left[\left(0.5 + 0.5\frac{E}{E_{TQ}}\right)^{0.23} - 1\right] \quad \ldots[\text{sone/Bark}] \qquad (3.2.5)$$

where E_{TQ} is the excitation at the threshold in quiet and E_0 is a "reference excitation" which corresponds to the reference sound intensity I_0. The critical bands are commonly approximated by a third-octave band filterbank. The overall loudness N is finally calculated by integrating over all specific loudnesses:

$$N = \int_{z=0}^{24Bark} N' dz \qquad (3.2.6)$$

The difference of a signal's short-time energy and loudness is impressively shown in Fig. 3.15 on page 72.

3.2.4. Syllable Timing and Syllable Length

The syllable is the basic rhythmic unit of speech. Before the first single letters were invented, it was syllabic writing which replaced the pictographs around 2000 BC (Fischer, 2001). Today, it is beyond controversy that timing happens on the syllabic level; however, it is still pretty unclear what exactly constitutes a *rhythmic event*.

From a phonologic point of view, a syllable is made up of a vowel (or a vowel-like sound) which can be preceded or followed by one or more consonants. Preceding consonants form the syllable's *onset*, following consonants are called the *coda*; the central vowel is also known as the *nucleus*. Linguists classify syllabic structures as shown in Tab. 3.1, marking vowel-like phonemes with a *V* and "consonantic" phonemes with a *C*.

Orthography	**Phonemes (IPA)**	**C-V Structure**
"a"	/æ/	V
"do"	/duː/	CV
"at"	/æt/	VC
"cat"	/kæt/	CVC
"scratched"	/skrætʃd/	CCCVCCC

Table 3.1.: The relationship of consonants and vowels to syllable structures (examples taken from Villing (2010))

A syllable thus has a characteristic structure with both a defined starting and ending point. From this definition, we can determine its *length*, but it is not possible to derive a precise point in time to be the *syllable event time*. If you had to clap your hands synchronously while pronouncing the monosyllablic words from Tab. 3.1, you probably would clap at the very beginning of "at", but more towards the end of "do" — this means you have a certain feeling of that syllable event time.

To answer the question where the rhythmic beats of speech are located, Allen (1972) conducted a series of listening tests and concluded that, for English, the *rhythmic stress beat* is perceived "around the release of the initial consonants and the onset of the nuclear vowel". Morton et al. (1976) introduced the term **perceptual center** (p-center) — initially with respect to whole words — for the specific moment at which a brief event is perceived to occur. Several mathematical models for the determination of p-centers for isolated syllables have been formulated since then (Marcus, 1981; Gordon, 1987; Howell, 1988; Pompino-Marschall, 1989; Scott, 1993; Harsin, 1997), each based on different acoustic properties of the speech signal. A comparative study has been conducted by Villing (2010), who tested both consistency and accuracy of predicted p-centers for speech, musical and synthetic sounds.

To arrive at something one could call *speech rhythm*, timing information alone is not sufficient; a rhythmic impression arises with the interaction of "strong" and "weak" elements, that is, less and more accentuated (or *prominent*) syllables. As we have learned, prominence is a function of all three auditive variables (see Fig. 1.3 on page 8), so **syllable length** also contributes significantly to the rhythmic impression of an utterance.

But what constitutes the perceived length of a syllable? A simple musical example shall form the basis of discussion:

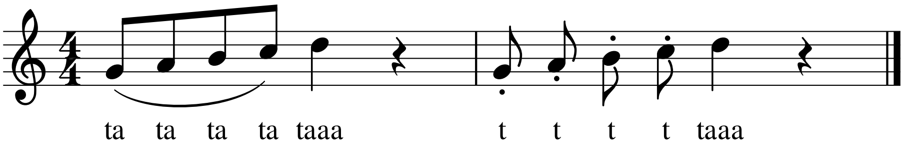


Figure 3.10.: A sequence of musical tones, expressed *legato* (left) and *staccato* (right).

What we see is a repeated sequence of musical tones consisting of four eighth notes, one quarter note, and one quarter rest. In the first bar, a slur indicates that the eight notes are to be played *legato*, which is the Italian word for "bound". In the second bar, the dots indicate that the eight notes are to be played *staccato*, which means "detached". In both cases, the (temporal) intervals between the note onsets are identical[10]; the perceived note length is determined by how long the notes are held.

It is thus important to note that the length of a syllable does not necessarily extend to the beginning of the subsequent syllable. A method to estimate the perceived syllable length is presented in section 3.4.1.

[10] In the case of, e.g., a piano sound, its onset can be equated with its perceptual center.

3.3. Intonation

3.3.1. Tracking the Fundamental Frequency

The acoustic basis of all melodic phenomena in speech is the fundamental frequency (f_0) contour of a speech sound. The term *fundamental frequency* refers to the lowest frequency in a mixture of harmonic waveforms. Strictly speaking, *periodic* implies “repeated exactly over time”, which is not the truth even for very short extractions of a speech signal. We can, however, assume *wide-sense stationarity* if we cut the signal into very short parts for analysis purposes, such that the signal statistics are approximately constant over that short time, and estimate the fundamental frequency for this short *frame*.

Fundamental frequency tracking is a common task in the world of audio signal processing, so there are several established methods to choose from; most of them even available as open-source programming code. Each of these methods, however, has its own advantages and disadvantages. Most of them can be reduced to three alternative basic principles for f_0 estimation which will be presented in the following, before discussing the algorithm of choice for this thesis.

The Auto-Correlation Method

A straight-forward approach for f_0 tracking is periodicity detection using the *auto-correlation function* (ACF). The ACF is the integral over the product of a windowed signal excerpt with a shifted version of itself:

$$r_x(\tau) = \int x(t)x(t-\tau)dt\,, \tag{3.3.7}$$

where the *lag* τ indicates the relative time shift. This function has a global maximum for $\tau = 0$. If there is any periodicity in the signal, we will observe local maxima for $\tau > 0$. Assuming the fundamental frequency to be the most prominent frequency in the complex signal mixture[11], the fundamental period T_0 can be determined by picking the maximum peak of the ACF at $\tau = \tau_{max}$. The fundamental frequency then is given by the reciprocal of the period,

$$f_0 = \frac{1}{T_0}\,. \tag{3.3.8}$$

Since the peaks of the ACF are periodically repeated, the auto-correlation method is prone to *octave errors*, meaning that it might erroneously pick a

[11] This assumption does not generally hold for all kinds of sounds, but is valid for a monophonic speech signal.

peak in the ACF which corresponds to half or double the fundamental frequency. Therefore, the range of lags has to be restricted to reasonable values, $\tau_{min} <= \tau <= \tau_{max}$.

When calculating the *normalized auto-correlation*,

$$r'_x(\tau) = \frac{r_x(\tau)}{r_x(0)}, \tag{3.3.9}$$

we get a measure for the *harmonic strength* of the signal with $R_0 = r'_x(\tau_{max})$, $R_0 = [0 \dots 1]$ which can be used to decide if a small chunk of speech is voiced or unvoiced.

An excellent, in-depth description of the auto-correlation method has been published by Boersma (1993).

Spectral Peak-Picking

The fundamental frequency of a sound can obviously also be determined from the magnitude spectrum in the frequency domain by selecting the most prominent peak within a reasonable range, $f_{min} <= f_0 <= f_{max}$.

Though simple in nature, this approach has the drawback of limited frequency resolution for short time windows as a consequence from the time-frequency uncertainty principle[12]. The frequency resolution of a given spectrum is simply the sampling frequency divided by the window length,

$$\Delta f = \frac{f_s}{N}, \tag{3.3.10}$$

which results in at least $15Hz$ for a comparably long analysis window[13].

On the contrary, the spectral approach allows for polyphonic detection, which is not possible with time-domain methods. Techniques such as *time-frequency reassignment* use local estimates of instantaneous frequency and group delay to enhance the resolution both in time and frequency (Auger et al., 2013).

[12] Küpfmüller's uncertainty principle says that it is impossible to sharply localize a signal in both time domain and frequency domain at the same time; when using a long window, we achieve good frequency resolution at the cost of temporal resolution, and vice versa.

[13] As an example, consider male speech recorded at $f_s = 16kHz$, which is a typical sampling rate for speech. Setting the minimum frequency to be detected to $f_{min} = 50Hz$ demands a window length of $60ms$ which equals 960 samples. To employ the fast Fourier transform (FFT), we need a window length which is a power of two, so we will chose $N = 1024$.

The Cepstral Approach

Another Fourier transform of the logarithmized magnitude spectrum yields the so-called *cepstrum* of a signal:

$$C_y(t) = \mathcal{F}\left\{\log|Y(f)|\right\} \ . \tag{3.3.11}$$

The cepstral domain is of interest, because it allows the separation of "source" and "filter" (see the simplified model of speech production in section 3.1.1) using a mathematical trick: the Fourier transform turns a convolution of two signals into a multiplication of their spectra, and at the same time, the logarithm turns a multiplication into an addition (Oppenheim and Schafer, 2004). In mathematical terms,

$$y(t) = x(t) * h(t) \tag{3.3.12}$$

$$Y(f) = X(f) \cdot H(f) \tag{3.3.13}$$

$$\log|Y(f)| = \log|X(f)| + \log|H(f)| \tag{3.3.14}$$

So, peak-picking in the cepstral domain theoretically allows to track the fundamental frequency of the glottal source alone, without disturbing influences from the vocal tract. Still, other algorithms using temporal and/or spectral methods have to be proven to be equally precise and sometimes even more robust.

Approach Used in this Thesis

Zahorian and Hu (2008) have developed *yet another algorithm for pitch tracking* (*YAAPT*) which combines time-domain and frequency-domain processing. I have chosen this algorithm for two reasons: first, it has been designed and tested to be used with speech signals, and second, it has proven to be robust also for lower signal-to-noise ratios.

One common artifact of f_0 tracking algorithms are octave errors. To overcome this phenomenon, the *YAAPT* algorithm includes a special refinement procedure of the estimated frequency track over time.

The algorithm works as follows:

1. **Preprocessing**: The algorithm considers both the original input signal as well as a squared copy of it. Doing so, sum and difference frequencies are created, which can be used to partially restore a missing fundamental in case of band-limited signals.

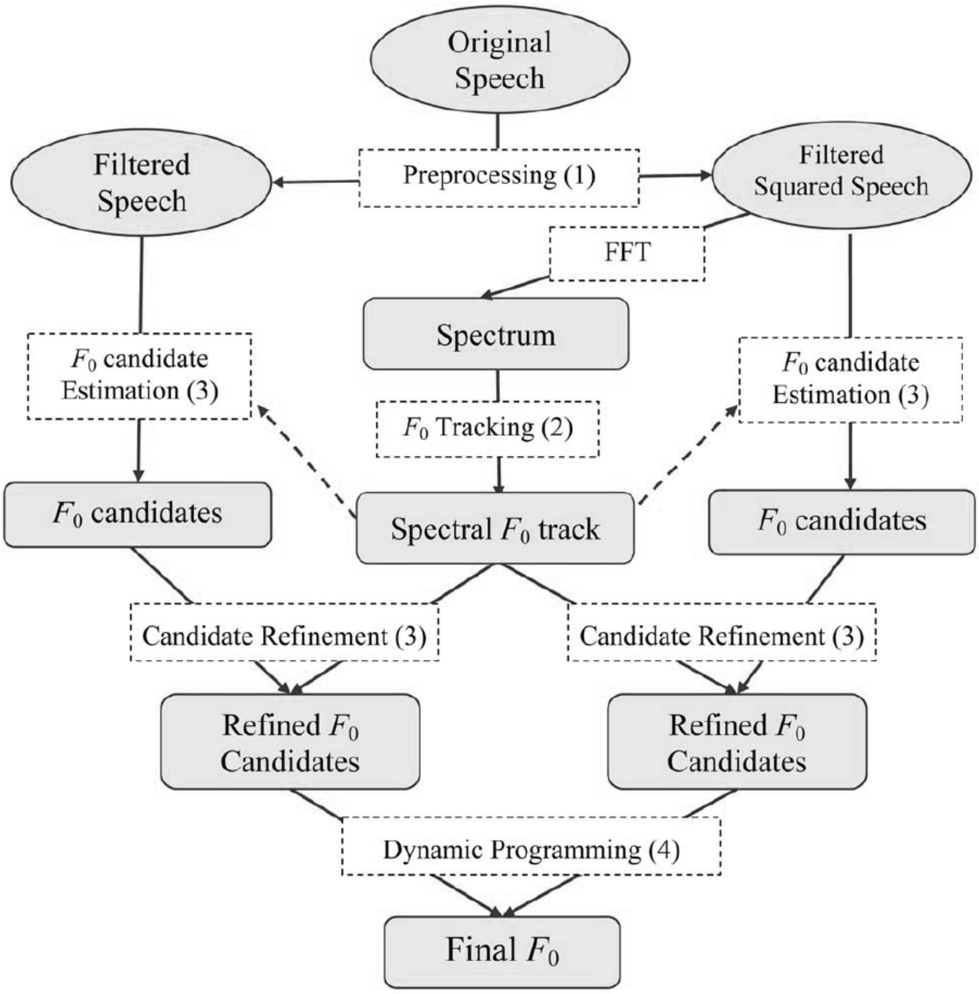


Figure 3.11.: Working principle of the YAAPT algorithm (from Zahorian and Hu (2008))

2. **Spectral f_0 tracking**: a spectrogram of the squared signal is created, and a series of f_0 candidates is estimated using a *spectral harmonics correlation* (SHC). In addition, a voiced/unvoiced decision is made by calculating the normalized low-frequency spectral energy for each frame and comparing it with a predefined threshold value.

3. **Temporal f_0 tracking**: another series of f_0 candidates is estimated using the *Normalized Cross-Correlation Function* (NCCF) between the non-modified and the squared input signal. Compared to the standard approach using the auto-correlation function, the NCCF results in more prominent peaks which are also more robust versus amplitude fluctuations. In a subsequent step, these f_0 candidates are further refined based on the spectral f_0 track estimated in step 2.

4. **Combination**: The "most likely" sequence of f_0 values is determined using Viterbi decoding[14], based on the assumption that the frame-to-frame variation of the fundamental frequency track should be minimal over voiced regions. Based on the voiced/unvoiced information from step 2, the final f_0 track is then interrupted during unvoiced regions.

For further details on the algorithm, the interested reader is referred to the original publication (Zahorian and Hu, 2008). Fig. 3.11 visualizes the steps described above; the numbers in the graphic correspond to the numbers in the text.

Fundamental frequencies are estimated in steps of $10ms$ using Hann-windowed segments of $60ms$, which is three times the period of a $50Hz$ signal, thus ensuring that at least one "full" period of the lowest frequency to be found would fit into the tapered analysis window.

3.3.2. Creating a Continuous Pitch Contour

Talkin (1995) has compiled a comprehensive list of natural speech phenomena which make f_0 estimation a challenging task for any algorithm. This is a summary of the most relevant points:

- Many parts of the speech signal are not "purely voiced" or "purely unvoiced", but mixed-excitation. When tracking fundamental frequency in speech, we make the simplifying assumption that only these two extremes exist.
- The fundamental frequency of speech is likely to change rapidly with time, so the assumption of wide-sense stationarity over the complete analysis frame of $60ms$ may not hold in some cases.
- A multiple of the true fundamental frequency can be emphasized by vocal-tract resonances and transmission-channel filtering, thus appearing as the most prominent harmonic in the complex sound mixture. It is furthermore common that sub-harmonics appear at integer fractions of the true f_0.
- The listener might also perceive a distinct pitch during unvoiced excitation due to narrow-band filtering by certain vocal-tract configurations.

These inaccuracies in the measured f_0 track are inevitable due to the nature of speech signals, and they remain after having calculated the corresponding pitch values using formula 3.2.1 (because we just apply a nonlinear scaling

[14] *Viterbi decoding* is a method to find the most probable sequence of values from the several possibilities with regard to a certain criterion using dynamic programming.

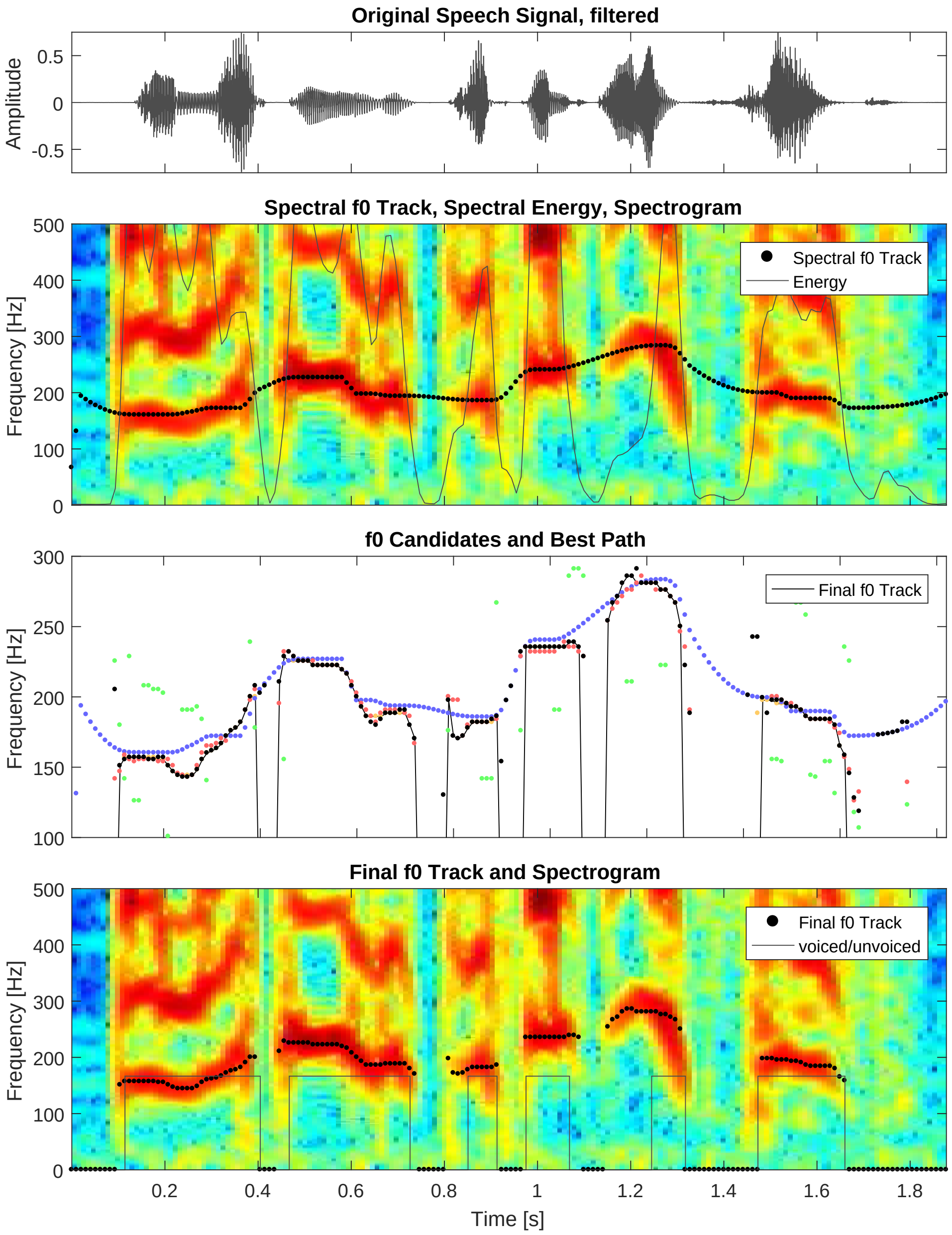


Figure 3.12.: Illustration of the main steps of the YAAPT algorithm. From top to bottom:
(a) original speech signal in the time domain;
(b) spectrogram of the original signal, low-frequency energy and approximate f_0 track from SHC (step 2);
(c) the various f_0 candidates from the NCCF, refined using spectral information (step 3, colored) as well as the final f_0 track found through Viterbi decoding (step 4, black);
(d) spectrogram of the original signal, voiced/unvoiced bit signal and final f_0 track.

of the f_0 values). However, our impression of "speech melody" does neither involve such rapid fluctuations as can be observed in the micro structure of the pitch signal, nor do we consciously perceive all these bursts and stops which happen during non-vocalic phonation as significant interruptions of the melodic flow. This is presumably due to the fact that a human listener usually concentrates on *what is said* and recognizes prosodic cues only in the background. Beyond, as t Hart (1981) found out, tonal differences of less than 2-3 semitones are commonly imperceptible in communicative situations[15]. This *just noticable difference* (JND) is remarkably higher than those reported for pure tones (Rakowski, 1971, amongst others), as it accounts for the highly dynamic nature of speech sounds.

In addition to that, we should keep in mind that many of the prosodic parameters mentioned throughout the literature consider timing and level of peaks in the pitch track. In other words: also from the analysis perspective, there emerges the need for a continuous and reasonably "smooth" curve to be described by parameters and statistics. This opinion is generally agreed by other phoneticians; a smoothed f_0 track is the basis of several existing tools for intonation transcription (Maghbouleh, 1998; Mixdorff, 2000; Hirst et al., 2000).

In the following, I will describe the **algorithm for pitch curve interpolation and smoothing** which creates the basis for all melodic parameter calculations (see section 4.2). As depicted in Fig. 3.13, it consists of three main steps:

1. **Interpolation** — In-between voiced segments, the missing pitch values are estimated using spline interpolation[16]. The pitch contour starts with the first and ends with the last voiced segment; there is no extrapolation.

2. **Stylization** — Inspired by the MOMEL algorithm (Hirst and Espesser, 1993), the interpolated signal is cut into overlapping frames of 300ms in length (50% overlap). The pitch course in these frames is then one-by-one approximated by a second-order polynomial, before a stylized pitch track is synthesized by combining the single polynomials using an overlap-and-add method.

3. **Smoothing** — The stylized curve is smoothed by a moving-average filter which length has been experimentally determined to 70ms (= 7 samples at f_s = 10Hz).

[15] Three semitones are the average threshold for untrained subjects; trained subjects performed only slightly better.

[16] Splines are piecewise cubic polynomials which consider the further course of the preceding and following values.

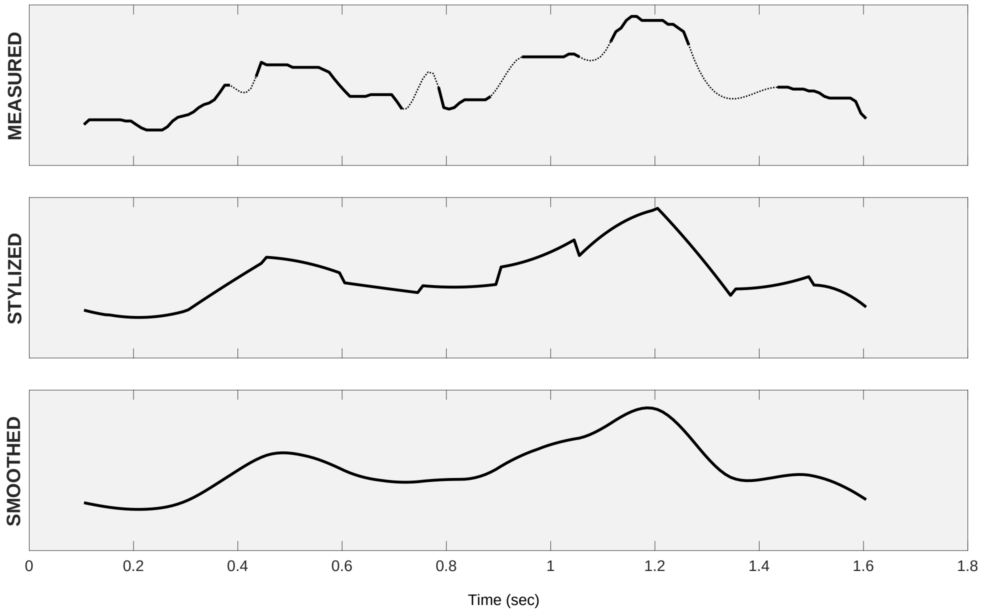


Figure 3.13.: Pitch contour stylization. Top: pitch values available from f_0 measurements (solid) with spline approximations during unvoiced parts (dotted). Middle: pitch contour stylization using piecewise parabolic fits (50% overlap). Bottom: averaged version of stylized curve.

3.4. Duration

As discussed in section 3.2.4, we need to know both the perceptual center and the duration of a syllable as a basis for the calculation of rhythmic speech parameters. As all of the p-center models evaluated by Villing (2010) expect an isolated syllable as input anyway, it is an obvious move to detect syllable boundaries first.

Segmenting a continuous flow of speech sounds into single syllables is a nontrivial task even for humans. Although listeners widely agree on the number of syllables (Villing et al., 2006), they show a considerable amount of inconsistency when asked to assign syllable boundaries to recorded speech. Especially a consonant between two vowels ("VCV") seems to be ambiguous between the end of the first syllable and the beginning of the second syllable (Goslin and Frauenfelder, 1999); phonologists call this phenomenon *ambisyllabicity* (Kahn, 2015).

3.4.1. Blind Estimation of Syllable Boundaries

Within the scope of my work, I want to facilitate analysis of spontaneous speech; a syllable boundary detection algorithm thus has to work rule-based and data-independent.

Several approaches for *blind syllable boundary estimation* have been published over the years, starting with Mermelstein (1975) who used the difference between the signal intensity envelope and its convex hull for the identification of potential boundaries. If a certain threshold was exceeded, the segment was divided into sub-segments which were then recursively evaluated. Extensions of Mermelstein's technique were proposed by Wu et al. (1998) and Meinedo et al. (1999), but as well as neural-net based approaches for syllable segmentation (Noetzel, 1991; Shastri et al., 1999), none of them prevailed as the tool-of-choice in the linguistic research community; presumably due to the fact that they generally perform well with clearly articulated syllables, but not with short, unstressed syllables (Villing et al., 2004).

In general, envelope-based syllable segmentation works quite well if the algorithm parameters are set suitably. Depending on the speaker, however, important prosodic parameters such as speech tempo may vary and thus might seriously affect the performance of the syllable detection algorithm (Villing et al., 2006). Simply speaking, it is possible to tune any algorithm in such a way that it detects a lot of potential syllable boundaries — we might end up with 99% of the syllable boundaries identified correctly, but at the expense of an enormous number of misdetections ("false positives").

A Self-Tuning Algorithm

So, what we desire is some kind of self-tuning algorithm which autonomously determines the number of syllables before setting the syllable boundaries.

Since we may assume that every syllable consists of a vowel or a voiced consonant which is optionally surrounded by consonants (see section 3.2.4), an approximate measure for the number of syllables in an utterance is the number of voiced segments in the speech signal; that is, homogeneous blocks for which the fundamental frequency tracker has found reasonable f_0 values. It may, however, be the case that a syllable ends with a voiced region and the following syllable starts with one — think of "the end" —, so adjacent syllables may share one single voiced segment.

Based on this assumption, I have designed an **algorithm for blind syllable segmentation** which creates a suitable decision function based on signal loudness and refines both number and duration of the detected syllables in subsequent steps:

1. **Boundary Candidates** — The algorithm creates an extremely smooth version of the loudness contour using moving-average filtering (Fig. 3.14, top). This smooth version is then subtracted from the original loudness contour, resulting in something which could be interpreted as "relative local loudness" (Fig. 3.14, middle). This curve serves as a the decision function, from which those local minima are selected as candidate boundaries which fulfill the following two criteria:

   a) Their value must be below zero (i.e. lower than the average loudness in this point).
   b) They must be at least 100ms apart from each other (= minimum syllable length).

2. **Boundary Verification** — For each potential syllable between two boundary candidates, the corresponding pitch values are analyzed. Depending on the number and the length of the voiced segments within the boundary candidates, one of the following actions is taken:

   a) One voiced segment of sufficient length: no action. The boundary candidates are marked as verified boundaries.
   b) No (or too short) voiced segment: the left candidate boundary is deleted, and the region is attached to the previous syllable.
   c) More than one voiced segment: the region is split into *n* syllables (with *n* being the number of voiced segments) by inserting additional boundaries centrally between two voiced segments.

3. **Perceived Syllable Lengths** — Within each verified syllable, the perceived start of the syllable is set to that point where the first rising slope reaches 15% of its total height, and the perceived end of the syllable is accordingly set to that point where the last falling slope falls down to 15% of its total height.

This procedure is illustrated in Fig. 3.14, using an arbitrary sentence from *Emo-DB* ("Der Lappen liegt auf dem Eisschrank"). The first plot shows the signal loudness curve and its smoothed version. The curve has been normalized such that the range of values covers $[0\ldots1]$. The second plot shows the decision function and the voiced segments as well as the syllable boundary candidates. Obviously, the first two syllables share the same voiced segment: the coda of /deːɐ/ and the onset of /lap/ are both voiced. A look at the last two syllables demonstrates the working principle of the boundary verification step: the potential boundary around 1.3s is deleted, because the part of the voiced segment in the following syllable is too short; the potential boundary around 1.65s is deleted, because the following syllable has no voiced content at all. The third plot illustrates how the criteria for the perceived syllable length calculation are applied; interestingly, the start and end times of the perceived syllables match quite well with the impression of the human annotator, whose labeled boundaries are depicted in the fourth plot.

I clearly have to admit that, in some cases, the algorithm fails to meet the correct number of syllables in the utterance. In this example, the human annotator has placed a syllable boundary at approximately 0.6s, whereas the algorithm identified two boundaries around that point, thereby introducing an additional syllable. This is due to the fact that the algorithm does not have any linguistic information on *what is said*, but only places the boundaries according to a signal characteristic by looking for minima. Still, this algorithm performance seems to be the best compromise we can reach for spontaneous speech where no linguistic meta-information is available.

The algorithm performance is controlled by three parameters: the assumed minimum length of the voiced syllable nucleus (set to 50ms), the assumed minimum syllable length (set to 100ms), and the window width of the moving average filter (set to 200ms). These values originate from a full-factorial parameter variation study[17] on Emo-DB data and have been found optimal with regard to the number of syllables detected. Since these parameters are of linguistic nature and not speaker-specific, I assume that this algorithm should work equally well on other kinds of speech data.

[17] *Full-factorial* parameter variation means that the syllable estimation algorithm has been tested with any possible combination of different values for these parameters. Details on this study can be found in appendix A.4.

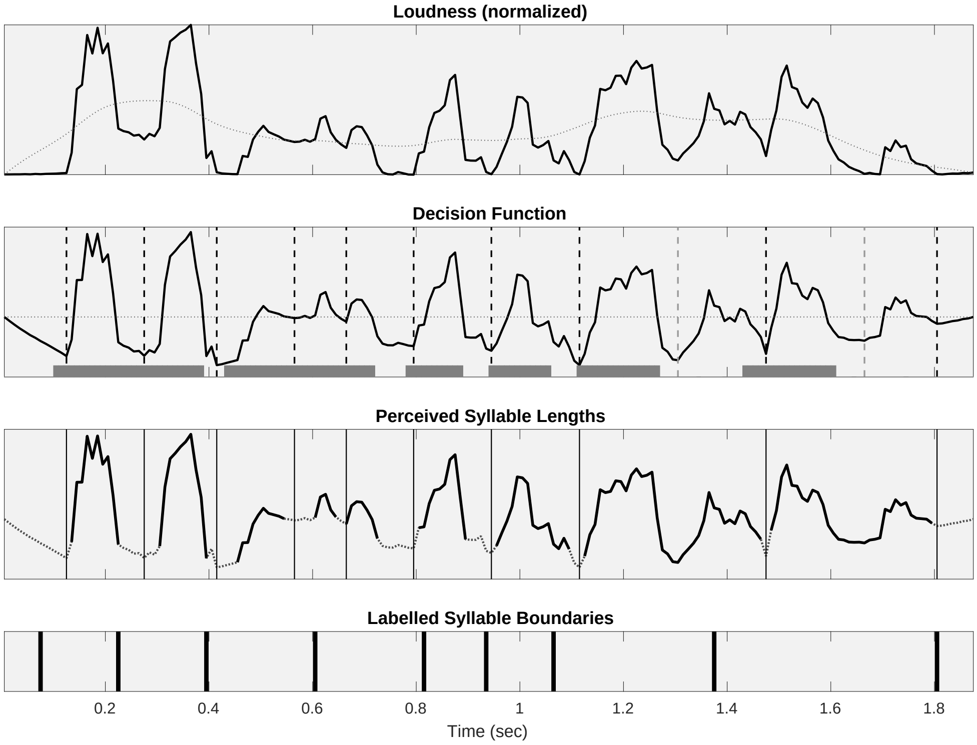


Figure 3.14.: First plot: loudness curve (solid) and its averaged version (dotted). Second plot: relative local loudness curve (solid), voiced segments (gray blocks on the bottom), and boundary candidates (dashed vertical lines). Third plot: decision function as above (dotted) and during perceived syllable duration (solid). Fourth plot: syllable boundaries as labeled by a human annotator.

3.4.2. P-Center Estimation

Thankfully, Villing (2010) has conducted a comparative study between several models for p-center (PC) estimation which includes the comparison of prediction accuracy on speech samples. Out of 8 models under test, the approach proposed by Pompino-Marschall (1989) significantly outperformed its competitors in terms of root mean square error, maximum absolute error, and percentage of *noticeable* erroneous p-center predictions[18].

The Pompino-Marschall model is based on the specific loudness in critical bands, calculated by Zwicker's method (Paulus and Zwicker, 1972). Having detected the syllable boundaries of an utterance as described above, the PC of an isolated syllable is calculated in the following way[19]:

[18] The last term indicates that, just as for pitch, there is a *just noticeable difference* for timing as well; it has been determined as 6ms for inter-onset-intervals (IOIs) which are shorter then 240ms, and 2.5% of the IOI for longer intervals (Friberg and Sundberg, 1995). Villing uses 5% of the IOI as a threshold value.

[19] A more detailed description of this PC estimation method can be found in (Villing, 2010).

- Within each critical band, rising and falling edges are detected and marked as *partial events* if they are sufficiently steep.
- Several coherent rising or falling partial events are individually weighted in a non-uniform way (depending on their distance to the peak between “rising” and “falling”) before being integrated into a single *peak onset event* or a *peak offset event*, respectively.
- Matching peak onset and offset events are integrated to form *peak events* which are calculated as the center of gravity of its corresponding onset and offset events.
- Finally, all peak events from all critical bands are integrated to form the center of gravity of the syllable, which is equivalent to its perceptual center.

A comparison of annotated and calculated syllable boundaries including their respective perceptual centers is shown in Fig. 3.15.

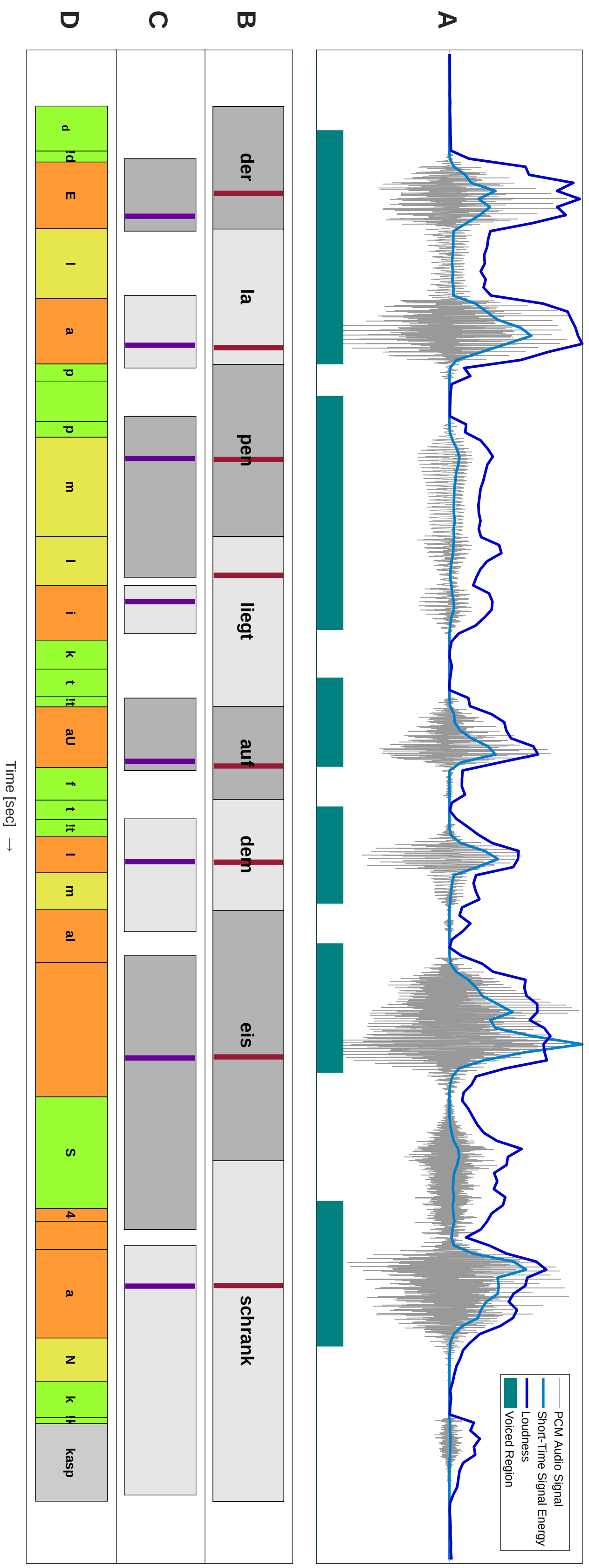


Figure 3.15.: [A] PCM audio signal (gray), energy (light blue), loudness (dark blue), and voiced regions (cyan). [B] Syllables as labeled by a human annotator and calculated p-centers. [C] Perceived syllable lengths as determined by the algorithm and calculated p-centers. [D] Single phonemes as labeled by a human annotator.

3.5. Prominence

From the literature, we know that all three acoustic variables (intensity, fundamental frequency, and time) contribute to the perceived prominence of a syllable (compare the sketch of the prosody model in Fig. 1.2 on page 7). These three variables are connected through articulatory relationships, so they are usually not consciously controlled independently of each other in spontaneous speech (Kehrein, 2002).

Prominence is thus not only a perceptual quantity; it is a weighted superposition of perceptual quantities. As explained in the previous sections and depicted in Fig. 3.16, loudness, pitch and perceived syllable lengths can be calculated from measurable variables using the corresponding models of sound perception, M_1, M_2, and M_3. The question is to what extent these auditive variables each contribute to prominence perception. Their individual contributions are denoted by the weighting factors w_1, w_2, and w_3. The literature provides mostly qualitative statements, e.g., that pitch and duration contribute more to prominence impression than intensity (Fry, 1958). A well-working concept for automatic prominence detection in German based on acoustic parameters has been presented by Tamburini and Wagner (2007), but the authors note that their weighting factors might be language-specific and thus not universally valid.

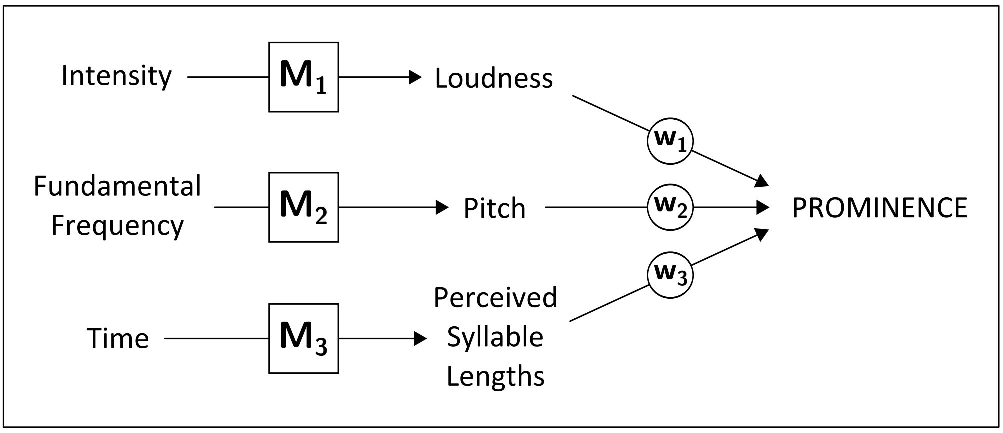


Figure 3.16.: Prominence as a weighted superposition of perceptual quantities.

The weighting factors w_1, w_2, w_3 can be estimated using *regression analysis*. Regression is a statistical approach to model the relationship between an independent *predictor variable* x and a dependent *response variable* y. If several predictor variables are observed, we are talking of *multiple regression*. Assuming that the response variable can be expressed by a linear combination of the prediction variables, a multiple linear regression model writes to

$$\boldsymbol{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon} = \sum_{i=0}^{p} \boldsymbol{x_i} \beta_i + \epsilon_i \,, \tag{3.5.15}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ x_{2,1} & \dots & x_{2,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}^T \text{and } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

This means that the p prediction variables, $\{\mathbf{x_1}, \dots, \mathbf{x_p}\}$ — which are in our case the auditive variables loudness, pitch, and perceived syllable length —, are each weighted by the corresponding *regression coefficients*, $\{\beta_1, \dots, \beta_p\}$, to predict the perceived prominence y for each of the n syllables in the utterance. The additional *error term*, ϵ, represents the variance which can not by explained by the model. The goal is to find regression coefficients which minimize the error, i.e., which lead to the most accurate model. The prediction variables may be non-linear, as long as the overall model remains linear in the parameter vector β. This allows combinations of predictors, e.g., $\mathbf{x_1} \cdot \mathbf{x_2}$, as well as higher-order predictors, e.g., $\mathbf{x_1}^2$. However, linear regression requires the prediction variables to be normally distributed and to have equal variance within each group.

Before we try to estimate the optimum regression coefficients for syllable prominence, I would like to discuss some general phenomena of prominence rating and perception.

3.5.1. Linguistic Function of Prominence

Prominence: A Continuous Variable?

The annotation system used for *Emo-DB* prominence annotation allows assigning one of 31 different levels of prominence to a syllable, which can be considered as a quasi-continuous rating scale. This scale has been introduced by (Fant and Kruckenberg, 1989); and it may prevent the annotator from feeling limited in his freedom to quantify the perceived prominence. It also allows studies on the relationships between measurable acoustic properties and prominence perception. However, the *just noticeable difference* (JND) between successive levels of prominence perceived by humans is considerably larger than 1/31 of the full scale. In other words: the annotators are given a tool which allows more accurate adjustments than they are able to do. Linguists widely agree on three or four distinguishable degrees of prominence at the most, especially in fluent speech (Kehrein, 2002); the *International Phonetic Alphabet* (IPA) allows the notation of three different levels of prominence.

There has been some research on how individuals subjectively rate prominence. Fant and Kruckenberg (1989) found relationships between test persons' prominence ratings and the self-evaluation of their own "inner voice" while reading, which suggests that not only acoustic or auditive cues are involved in prominence ratings, but also some sort of reading experience. These findings are supported by (Streefkerk, 2002), who states that human annotators have a certain "expectation of prominence" based on their linguistic knowledge which is combined with cues from the speech signal itself. Wagner (2005) also found a considerable influence of the raters' introspection in situations where the acoustic and auditive cues were difficult to interpret; unsurprisingly, especially amongst native speakers of that language. So, there is a wide consensus in the linguistic community that subjective prominence ratings are not solely based on acoustic or auditive variables, but that every annotator introduces some sort of "noise"; which in turn means that it will not be possible to perfectly predict subjective prominence ratings based on acoustic analysis of the speech signal.

To investigate if the *Emo-DB* annotator implicitly used "prominence categories" or not, I performed *agglomerative hierarchical clustering* on the union set of all syllable prominence values for the entire database. This means that the clustering algorithm starts with single elements and merges them into small clusters, before merging these small clusters into larger clusters[20]. Figure 3.17 displays a histogram of all syllable prominences in the upper plot. The bar colors indicate the cluster affiliations; prominence values ranging from [0..2] form the *low-prominence* cluster, values from [4..11] are considered *medium prominent*, and all values above that are *highly prominent*[21]. These results have been achieved with the aim to find exactly three clusters within the data; one could even argue to further subdivide the *medium* cluster into two if four degrees of prominence were to be found. To compensate for possible effects due to *final lengthening*[22], the same cluster analysis has

[20] I implemented an agglomerative hierarchical clustering approach using Ward's minimum variance method (Ward Jr, 1963) as the linkage criterion which minimizes the total within-cluster variance. This seemed to be the most appropriate method for my needs.

[21] The category labels are, of course arbitrary; linguistic equivalents could, e.g., be "unaccented", "medium accent", and "strong accent".

[22] Final lengthening is the linguistic term for the effect that a speaker unconsciously extends the final syllable of a sentence or a phrase in length to mark a textual boundary (Beckman and Edwards, 1990).

been performed for all syllables except of the final ones in each utterance. The differences between these two histograms, shown in the undermost plot, reveal that the final syllables are substantially given low prominence.

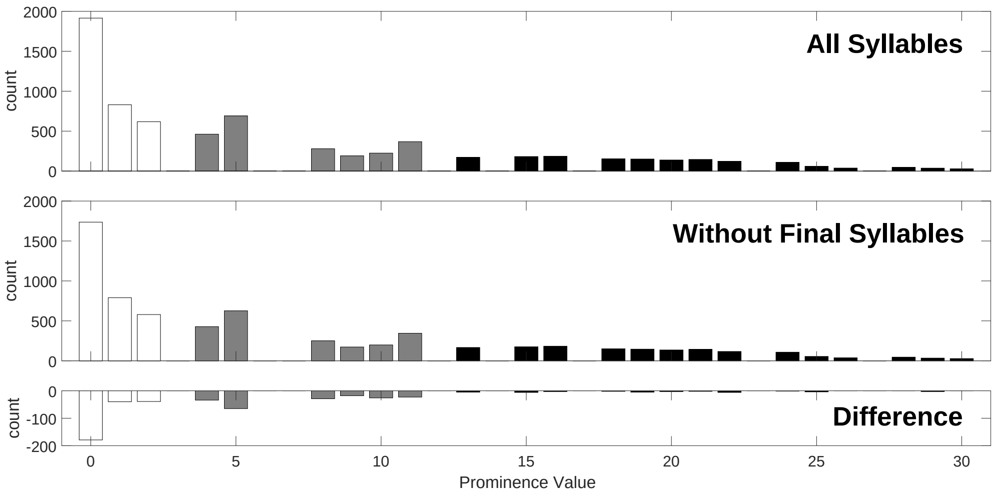


Figure 3.17.: Histograms and hierarchical clustering results of annotated Emo-DB prominence values. Top: all syllables; clusters indicated by colors (white = low prominence, gray = medium prominence, black = high prominence). Middle: All syllables but the final syllable of each utterance; colors as above. Below: difference between both histograms.

When looking at Fig. 3.17, we realize that the clustering algorithm surprisingly assigns more than (the upper) half of the scale to the category of highest prominence. This, however, reflects the natural behavior of the speakers. To illustrate this, let's take a look at the distribution of labeled prominence values per syllable for a specific sentence from *Emo-DB*, subsuming data from all speakers and emotions (Fig. 3.18). This sentence comprises 8 syllables, the second and the second last clearly being strongly accentuated. However, depending on speaking style and acted emotion type, their values vary in the range of [15..30], without losing their salient nature.

Prominence in the Context of Speech Rhythm

At this point, we should recall that the prosodic variable prominence is, together with duration, merely an intermediate on our way to create a rhythmic pattern for an utterance[23]. Since rhythm as such is already a complex matter, I have decided to keep it as simple as possible by only differentiating

[23] There is, in fact, one single rhythmic parameter which relies on estimated prominence values directly (*prominence dynamics*, → 4.3.5). All others describe properties of a previously calculated *rhythmic pattern* (→ 4.3.1).

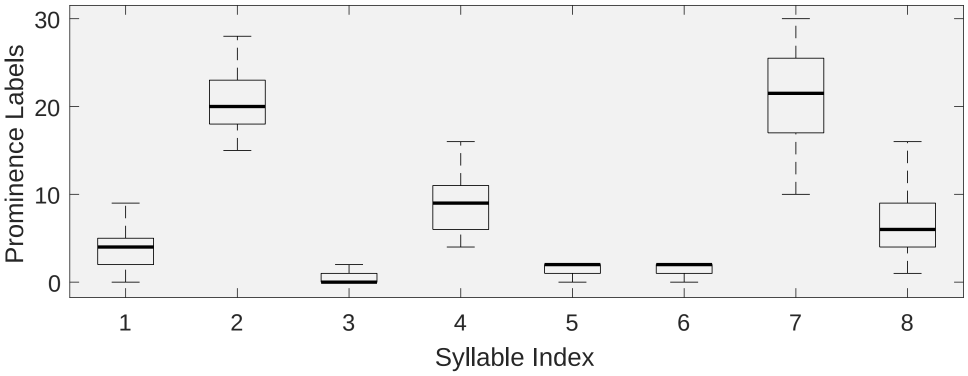


Figure 3.18.: Labelled prominence values from *Emo-DB* for the sentence "Der Lappen liegt auf dem Eisschrank", shown as Box-Whisker plots: the bold line shows the median, the box represents the range between the 25% and the 75% quartile, and the whiskers indicate the complete range of the data without outliers.

between *prominent* and *non-prominent* syllables, which reduces the number of prominence categories to two. This is a classification task which requires a certain value to be defined as the *split point* between the two prominence classes. The histograms in Fig. 3.17 as well as the syllable prominence distribution plots for the single *Emo-DB* sentences (which are all depicted in section A.1 in the appendix) provide evidence that a value of 12.5 would be a suitable threshold value for this purpose.

So, we would like to create a model for prominence prediction based on loudness, pitch, and syllable length information. We will use multiple linear regression to estimate a set of regression coefficients, $\{\beta_1, \ldots, \beta_p\}$, which constitute this model, and we will determine the prediction accuracy of this model by comparing predicted to actual prominence values. Since we are not mainly interested in the exact value which has been predicted, but rather if the correct category has been chosen – namely, "prominent" or "non-prominent" –, each prominence value in the range of [0..30] is assigned a *class probability.* This probability takes values between 0 and 1 and linearly levels off when approaching the split point, as shown in Fig. 3.19.

Why are we doing this? If we made a binary assignment of "prominent" or "non-prominent", any misclassification would lead to equal high error values. A prediction of 12 ($\Rightarrow C_1$) when the annotator noted 13 ($\Rightarrow C_2$) would be judged identically as a prediction of 28 when the annotator noted 3. The fuzzy approach introduces a region of uncertainty in the range of [10..15], accounting for the fact that deviations in this area fall below the JND for fluent speech.

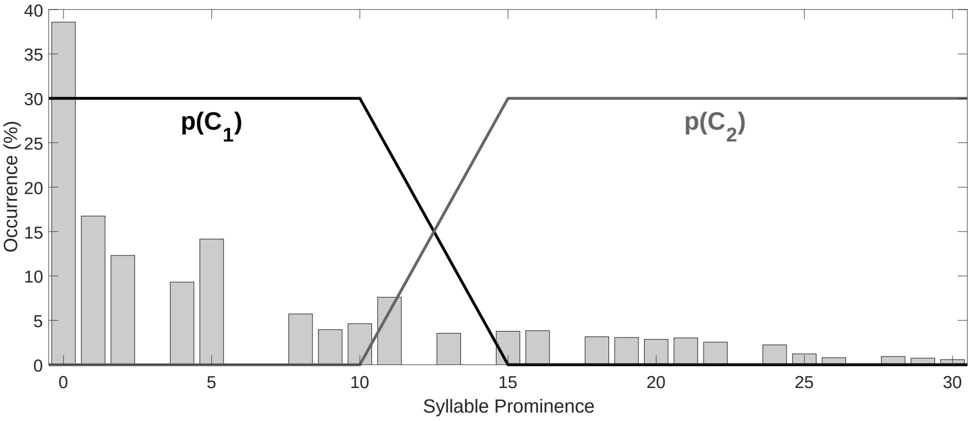


Figure 3.19.: Prominence values histogram and assigned class probabilities for "non-prominent" (C_1) and "prominent" (C_2) values.

Absolute or Normalized Values?

A closer look at labeled prominence values for an arbitrary sentence from the *Emo-DB* database reveals that the available spectrum of prominence values is not fully exploited in most cases. This seems to be reasonable if we assume that a speaker will usually not make use of his or her full range of tones and loudnesses during one and the same utterance. On the other hand, we may hypothesize that it is the relative prominence of a syllable compared to the rest which makes it *accentuated* or not. This last hypothesis demands for normalization of all prominence values within an utterance to the full range:

$$P_{norm} = \frac{30 \cdot (P - min(P))}{max(P)} \,. \tag{3.5.16}$$

In this case, also the predictor variables will have to be normalized in a similar way (without the factor of 30) for regression analysis. I will consider both possibilities and evaluate if one method is superior to the other or not.

3.5.2. Syllable Pitch

Before discussing how the optimum regression coefficients for loudness, pitch, and perceived syllable length can be found, we should give some thought to the way we can assign values from continuous variables — as loudness and pitch are — to a syllable. For loudness, it seems reasonable to simply calculate the average over the syllable duration, but it is highly questionable if this is also valid for pitch.

I agreement with other authors (Clark, 1999), I hypothesize that, if the pitch contour shows a peak or valley during a syllable, the pitch value at this turning point will be perceptually most important. If the pitch contour doesn't show a turning point during that syllable, its perceived pitch can be estimated by calculating the average pitch between the syllable boundaries. In mathematical terms,

$$p_{syl} = \begin{cases} f_{mel}[n_{tp}] & \text{if turning points detected} \\ \frac{1}{L}\sum_t f_{mel}[n] & \text{otherwise} \end{cases} . \tag{3.5.17}$$

In the formula above, $f_{mel}[n]$ is the discrete-time pitch contour, and n_{tp} denotes the time index of the turning point.

We thus need an analysis algorithm which is able to detect if an arbitrary extraction of the continuous pitch contour contains a distinctive turning point, meaning that an audible change in pitch direction occurs. One could intuitively think of fitting first- and second-order polynomials to the pitch contour extractions and to calculate the mean squared error as a measure for the *goodness-of-fit* (GoF). If the parabola showed the better GoF compared to the straight line, one could argue that this extraction of the frequency contour is "rather parabola-like" and thus might show a distinctive turning point. This approach will, unfortunately, not work in this case due to the fact that the stylized pitch contour has itself been created based on second-order polynomials, such that an algorithm would always prefer second-order to first-order fits.

After some guesswork, I finally found a robust classification procedure which works subject to the following two criteria which must be fulfilled to classify a pitch contour extraction as *having a turning point*:

- The extraction must have an extreme value (minimum or maximum) inside a certain temporal margin around the syllable center. This margin is defined using Eq. 3.5.18.
- If an extreme value lies within the margin, it must also be sufficiently distinctive. This is tested by calculating the differences to both the first and the last pitch value of the extraction and comparing the normalized distance to a threshold value (Eq. 3.5.19).

To fulfill the first criterion, the relative frame index of the extreme value must lie within the margin

$$m(L) = \left[\frac{L}{2} - (\alpha \cdot \frac{L}{2}) \,..\, \frac{L}{2} + (\alpha \cdot \frac{L}{2})\right] , \tag{3.5.18}$$

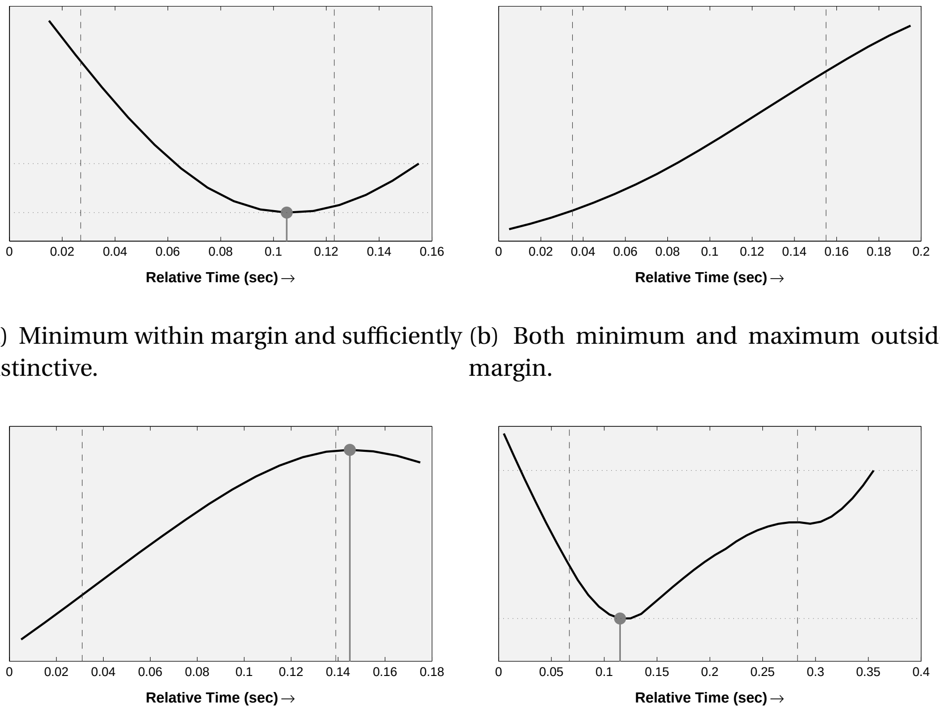
where L is the length of the syllable in frames and α is a scaling factor, which has experimentally found to be optimal as 0.6, such that the margin covers 60% of the syllable duration.

The second criterion is fulfilled if the inequality

$$\frac{\min\left(\left|f_{mel}[1]-f_{mel}[n_{tp}]\right|,\ \left|f_{mel}[L]-f_{mel}[n_{tp}]\right|\right)}{\max(f_{mel})-\min(f_{mel})} \geq \beta \qquad (3.5.19)$$

holds. Again, L is the length of the syllable in frames; $f_{mel}[n_{tp}]$ is the extreme value under test and β is a threshold value experimentally set to 0.2. This means that the difference must exceed 20% of the total dynamics of the syllable's pitch contour.

The working principle of the algorithm is visually summarized in Fig. 3.20, where the dashed, vertical lines represent the temporal margin and the dotted, horizontal lines mark the distance between the extreme value and the "closest" first/last value.



(a) Minimum within margin and sufficiently distinctive.

(b) Both minimum and maximum outside margin.

(c) Maximum outside margin, turning point not distinctive.

(d) Minimum within margin and sufficiently distinctive.

Figure 3.20.: Working principle of the turning point detection algorithm.

3.5.3. Retrieving Optimum Regression Coefficients

Recalling that we have defined *prominence* as the extent to which a syllable perceptually "stands out" of its environment, it seems reasonable to not only include the absolute values of the three auditive variables for each syllable in our regression analysis, but also their differences and quotients with regard to their surrounding syllables. So, if we regard any of the three auditive variables as our predictor variable y, we can define four additional predictors for each of them:

$$y_{\Delta,l} \equiv \left[\begin{array}{ccccc} 0, & y(2)-y(1), & y(3)-y(2), & \ldots, & y(n)-y(n-1) \end{array} \right] \tag{3.5.20}$$

$$y_{\Delta,r} \equiv \left[\begin{array}{ccccc} y(1)-y(2), & y(2)-y(3), & \ldots, & y(n-1)-y(n), & 0 \end{array} \right] \tag{3.5.21}$$

$$y_{\div,l} \equiv \left[\begin{array}{ccccc} 1, & \frac{y(2)}{y(1)}, & \frac{y(3)}{y(2)}, & \ldots, & \frac{y(n)}{y(n-1)} \end{array} \right] \tag{3.5.22}$$

$$y_{\div,r} \equiv \left[\begin{array}{ccccc} \frac{y(1)}{y(2)}, & \frac{y(2)}{y(3)}, & \ldots, & \frac{y(n-1)}{y(n)}, & 1 \end{array} \right] . \tag{3.5.23}$$

Doing so, we have determined the number of potential predictors, p, to 15. Depending on the number of syllables in the utterance, n, the set of linear equations formulated in Eq. 3.5.15 will in most cases either describe an *overdetermined* or an *underdetermined system.* In an **overdetermined system**, there are more equations than unknowns ($n > p$). Provided that there are no (or not enough) linear dependencies between the equations, such a system has no exact solution. In an **underdetermined system**, there are fewer equations than unknowns ($n < p$), and it has either no solution or infinitely many solutions. The good news is that, in both cases, an approximate solution can be found using the *ordinary least squares* method; and in the case of an underdetermined system, this is the only relevant solution. The regression model is fitted to the labeled prominence values by minimizing the sum of squared differences between the labeled values and their corresponding modeled values.

This results in a closed-form expression[24] for the estimated values of the unknown parameters β_p:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \left(\sum \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\left(\sum \boldsymbol{x}_i y_i\right) \tag{3.5.24}$$

For both absolute and normalized variables, a full-factorial regression analysis has been performed. This means that optimum regression coefficients β were calculated for all possible parameter sets emerging from the 15 predictor variables ($N = 2^{15} - 1 = 32767$) both for each speaker individually (*local* models) and over all speakers (*global* models). The expected results are:

- **A set of relevant regressors**. Which perceptual quantities contribute most to prominence perception? Are these regressors the same for all speakers, or are there individual differences?

[24] A closed-form expression is an expression which can be evaluated in a finite number of operations.

- **A model for prominence prediction**. We want to be able to predict syllable prominences in spontaneous speech, and thus we need the relevant regressors and their corresponding coefficients $\{\beta_1, \ldots, \beta_p\}$.
- **Effects of normalization**. Will normalized auditive parameter values (in the range [0..1]) predict normalized prominence values (using the full scale from [0..30]) better than their absolute counterparts?
- **The generalization loss**. To which extent are speaker-specific prominence models more accurate than a global model?

As a measure of model accuracy, both the predicted and the actual prominence values are translated into their corresponding class probabilities for the classes C_1 and C_2. Using the following notations

$p(\hat{\mathbf{y}} = C_1)$	probability of predicted PV belonging to C_1
$p(\mathbf{y} = C_1)$	probability of actual PV belonging to C_1
$p(\hat{\mathbf{y}} = C_2)$	probability of predicted PV belonging to C_2
$p(\mathbf{y} = C_2)$	probability of actual PV belonging to C_2

(where *PV* is short for “prominence values”, of course), we can calculate the product $p(\hat{\mathbf{y}} = C_1) \cdot p(\mathbf{y} = C_1)$ as the “hit rate” for the assumption that the PV belong to C_1, and we can do this similarly for C_2. This is equivalent to the calculation of the inner product when thinking of $p(\hat{\mathbf{y}} = C_1)$ and $p(\mathbf{y} = C_1)$ as vectors. A meaningful score for the accuracy of prominence class prediction can now be calculated by normalizing the sum of the inner products for both classes by the number of syllables in the utterance:

$$Score = \frac{\langle p(\hat{\mathbf{y}} = C_1), p(\mathbf{y} = C_1)\rangle + \langle p(\hat{\mathbf{y}} = C_2), p(\mathbf{y} = C_2)\rangle}{n} \cdot 100\% \,. \qquad (3.5.25)$$

Tab. 3.2 lists the achievable scores and the corresponding number of regressors used in the best-performing model for single speakers as well as for the “global set” of all 10 speakers from *Emo-DB*. Details on the selected regressors for each of these models are listed in appendix A.1.

Two aspects are striking about the results in Tab. 3.2:

1. There seem to be large differences in speaking style between different speakers. While only 2 featured predictors are sufficient to predict syllable accentuation with an accuracy of 80% for speaker 5, it takes 10 predictors to achieve the best result for speaker 8 which, in comparison, is even not just as good.

Speaker	**Absolute Values**		**Relative Values**	
	Score	No. Reg.	Score	No. Reg.
1	86.64%	7	85.98%	9
2	84.98%	8	85.88%	10
3	82.04%	6	85.02%	11
4	87.51%	9	91.18%	13
5	80.86%	2	81.08%	8
6	92.99%	11	100.00%	15
7	87.72%	9	84.17%	11
8	77.91%	10	75.48%	8
9	83.49%	5	82.45%	10
10	82.21%	10	76.55%	8
ALL	79.02%	5	75.55%	11

Table 3.2.: Best-performing sets of regressors for prominence prediction, both for individual speakers and globally.

2. For one half of the speakers, relative values word better than absolute values; for the other half, the opposite is true. Overall, absolute predictor values show a better performance than relative predictor values.

Before selecting the top-performing set of absolute predictors as the winner, let's have a closer look at the lower ranks to get an idea of how relevant the addition or the removal of single predictors might be. Tab. 3.3 lists the "top ten" scores out of all possible 32767 combinations and the corresponding sets of predictors. What we see is that some predictors appear very consistently throughout all sets, while others just pop up here and there. To arrive at a **robust set of regressors**, I have decided to make a majority decision amongst these ten sets, resulting in the following 5 regressors plus a linear offset:

Regressor	**Coefficient** $\hat{\beta}_p$	**Notation**
(offset)	5.789759	c_0
$P_{\Delta,l}$	0.024277	c_1
$L_{\Delta,r}$	0.162531	c_2
$L_{\div,l}$	0.619543	c_3
D	14.664055	c_4
$D_{\div,r}$	-0.739814	c_5

Syllable prominence will thus be calculated using this formula:

$$Pr = c_0 + c_1 \cdot P_{\Delta,l} + c_2 \cdot L_{\Delta,r} + c_3 \cdot L_{\div,l} + c_4 \cdot D + c_5 \cdot D_{\div,r}\ . \tag{3.5.26}$$

Score	**Regressors in 10 best-performing Global Models**														
	Pitch					Loudness					Duration				
	-	Δ_l	Δ_r	$\div_l$	$\div_r$	-	Δ_l	Δ_r	$\div_l$	$\div_r$	-	Δ_l	Δ_r	$\div_l$	$\div_r$
78.94%	0	**1**	0	0	0	0	0	**1**	**1**	0	**1**	0	0	1	**1**
78.92%	0	**1**	0	0	0	0	0	**1**	**1**	0	**1**	1	0	0	**1**
78.92%	0	**1**	0	0	0	0	0	**1**	**1**	0	**1**	0	1	1	**1**
78.90%	0	**1**	1	0	0	0	0	**1**	**1**	0	**1**	0	0	1	**1**
78.87%	0	**1**	0	0	0	0	0	**1**	**1**	0	**1**	0	1	0	**1**
78.84%	0	**1**	1	0	0	0	0	**1**	**1**	0	**1**	0	0	0	**1**
78.81%	0	**1**	0	0	1	0	0	**1**	**1**	0	**1**	0	0	0	**1**
78.80%	0	**1**	0	0	0	0	0	**1**	**1**	0	**1**	1	1	0	**1**
78.80%	0	**1**	0	0	0	0	0	**1**	**1**	0	**1**	0	1	0	0

Table 3.3.: Ten best-performing global sets of regressors for prominence prediction using absolute values.

Summarizing, we can state that absolute values of the auditive variables in general allow for better prediction of syllable prominence than relative values. A global regression model has been found which makes use of 5 auditive variables of which 4 are differences and quotients with respect to previous or following syllables, which confirms the assumption that prominence is synonymous with the *relative* salience of a syllable. The values of the regression coefficients $\hat{\beta}_p$ allow no conclusions on the importance of the single variables, since their ranges are completely different. The idea to perform a majority decision amongst the best 10 sets of predictors proved to be reasonable, since each of these ten sets contains at least one dispensable parameter which does not affect regression accuracy when being omitted.

Evaluation of Results

Yoon (2010) evaluated inter-speaker consistency on the presence or absence of pitch accent in a dataset of radio speakers reading the same texts. Analyzing *ToBI*-annotated labels[25], he found an average consistency of 79.8% for

[25] *ToBI*, short for *Tones and Break Indices*, is a standard for prosody labeling (Beckmann and Ayers-Elam, 1997).

“accented” vs. “non-accented” syllables over 3 female and 2 male speakers. This seems to be a good benchmark for the achievable accuracy of a “global”, i.e., non-speaker-specific model.

Al Moubayed et al. (2010) built prominence models for Swedish where the prominence levels of whole words rather than single syllables were to be estimated. Introducing their own interpretation of *fuzzyness* with a “maybe” class in-between “yes” and “no”, their best result for word prominence classification was about 69% correctly classified words for one single speaker. Interestingly, they report that the inter-annotator agreement between 4 persons was just about 69.2% as well.

The mentioned automatic prominence detector by Tamburini and Wagner (2007) uses the duration of syllable nuclei (in addition to intensity, pitch movements and spectral emphasis) as one of several features, which had been determined manually in advance. Though not being “fully” automatic for this reason, the approach is certainly comparable to what I have presented in the previous paragraphs; the authors report a prediction accuracy of $\rho = 0.71$ using Spearman’s rho as a measure[26].

In this context, the achieved speaker-independent regression accuracy of 79% can be viewed quite as a strong result.

[26] Spearman’s rho describes how well a relationship between two variables can be described with a(ny) monotonic function; this function does not have to be linear.

4. Deriving Prosodic and Paralinguistic Parameters

This chapter describes how descriptive scalar parameters are derived from the prosodic variables. The selection of parameters is linguistically or musically motivated. 27 melodic, rhythmic and paralinguistic parameters are described in detail. A complete list of prosodic and paralinguistic parameters can be found at the end of this chapter.

4.1. General Considerations

4.1.1. How "Free" is Free Speech?

Accentuation and Timing

Whether reading out a written sentence or speaking using our own words, the rhythmic structure is – up to a certain degree – determined by the used words and their corresponding pronunciation rules. It is thus important to understand that the speaker has limited degrees of freedom regarding both accentuation and timing of the syllables. As we want to find descriptive parameters of speech rhythm which eventually allow to discriminate not only between single speakers, but even between emotional states or stress levels, we should be aware of the remaining "degrees of freedom" a speaker has left when the words are given.

To get an idea of how similar (or not) syllable accentuation and timing can be when various speakers produce a given sequence of words, these two parameters are visualized in Fig. 4.1 for ten different speakers[1]. The positions (x-direction) of the black bars indicate the perceptual centers of the syllables,

[1] The sentence used reads *Ich will das eben wegbringen und dann mit Karl was trinken gehen* in "angry" emotion from the Emo-DB database.

while the bar heights represent the perceived prominences as labeled by a human annotator. In this figure, the time scale has been normalized for each speaker individually, such that the first and the last accentuated syllable occur at the same time instance. This was done to illustrate how the speakers distribute the syllables in time: when regarding the first prominent syllable as the rhythmic starting point (or the *metric onset*, cp. 4.3.1) and the last prominent syllable as the "rhythmic target", differences in syllable timing are most obvious with respect to these anchor points.

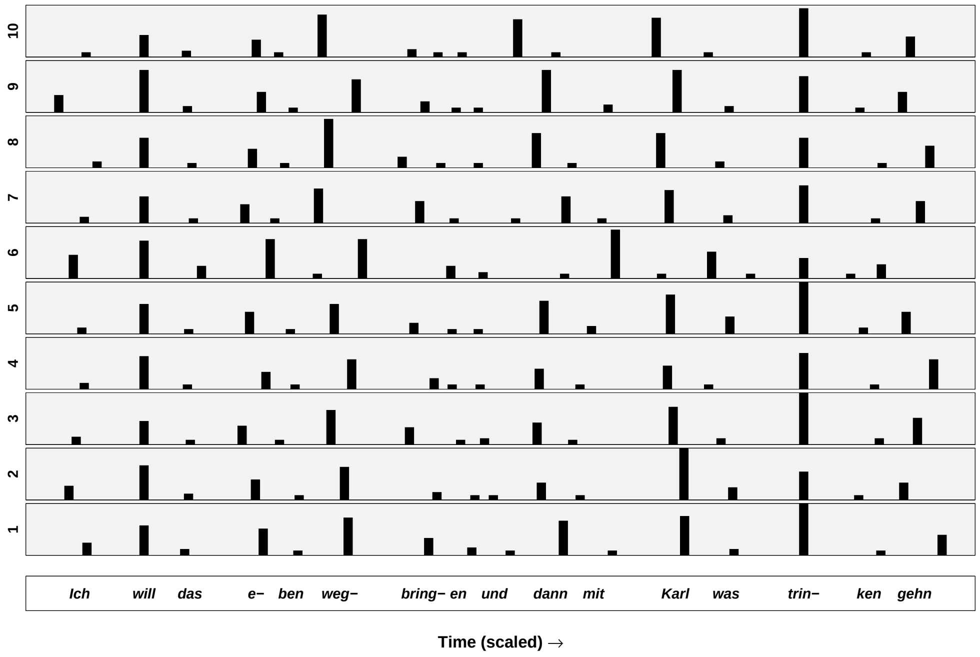


Figure 4.1.: Comparison of syllable accentuation and timing: perceptual centers (x-axis) and perceived prominences (y-axis) for the same sentence realized by 10 different speakers. Time axis normalized.

Looking at Fig. 4.1, we notice two things:

1. The perceived prominences of single syllables show only little variance across speakers. Although there are differences in absolute values, the proportions of consecutive syllable prominences are the same over all speakers. (Thus the third syllable, for example, is remarkably less prominent than its surrounding neighbors, syllables no. 2 and 4; which is true for speaker 6 as well as for all the others.) Indeed, it seems that the spoken content very much prescribes how much prominence a syllable is to be given by the speaker.

2. In contrast, syllable times vary considerably across speakers. This suggests a quite high degree of freedom in timing, which, in fact, accords

with our everyday speaking experience. Note that speaking tempo is ignored in Fig. 4.1 due to the normalized time axis.

Evidence for Isochrony?

The idea of *isochrony* on the syllable level is a basic principle in speech rhythm research and constitutes the division of languages into *syllable-timed*, *stress-timed* and *mora-timed* languages. In a strict sense, it is the postulate that all syllables of an utterance are distributed evenly in time. Despite the fact that, to date, no research has been published which shows any evidence for strict isochrony in whatever language, there are many signs that it exists on an impressionistic or perceptual level (Wagner, 2008).

This sentence above, freely translated, means: *I'll just bring that away and then go for a drink with Karl*, which are obviously two concatenated clauses. In both of these clauses, the prominent syllables appear quite regularly in groups of three; independently of the pause length before the word *and*. This might support the hypothesis that speakers tend to produce logical sub-units of a longer utterance in a rhythmically uniform way.

What is indeed observable in Fig. 4.1 is a phenomenon which linguists call *compensatory shortening*: the more non-prominent syllables are to be pronounced between two prominent syllables, the more the speaker tends to "compress" those non-prominent syllables in time in order to keep up the isochrony (Pike, 1945).

4.1.2. Temporal Aspects

The acoustic variables are calculated from the recorded speech files in temporal steps of 10 milliseconds, independent of the respective analysis window lengths (which are variable-specific for different reasons and may, e.g., depend on the minimum frequency for pitch tracking). We are thus provided with 100 values per second, which would mean an extreme overload for our sensory system if all this information was processed without perceptual prefiltering. Instead, the melodic and rhythmic impression of an utterance is created using the information from certain "anchor points" in the continuous speech signal flow.

Although a fundamental frequency is only apparent in vowels and voiced consonants, we perceive a continuous "melody" over the whole utterance, with the exception of intended speech pauses (→ 3.3.2). Depending on the nature of the melodic parameter to be extracted, the perceptually meaningful points are either local extrema of the curve (as used, e.g., for declination and peak shape), or one-per-syllable (in the case of utterance harmony).

For the rhythmic aspects of speech, these anchor points are the perceptual centers of the single syllables (→ 3.4). So, in the context of rhythm, the syllable is reduced to an *event* which is a point in time and has no length; its duration indeed contributes to its perceived prominence, but its temporal location is determined by its perceptual center.

4.2. Melodic Parameters

The melodic parameters presented in this section are scalar, linguistically and musically-motivated descriptors of speech characteristics which are calculated based on the continuous pitch contour (→ 3.3.2). Depending on the nature of the melodic parameter to be extracted, the perceptually meaningful points are either local extrema of this contour, or one-per-syllable.

4.2.1. Declination

In natural speech, most people tend to start speaking with a moderate pitch, which is then gradually lowered during the sentence. This phenomenon was first reported by Pike (1945). Although not indisputable amongst intonation researchers, the *declination* of an utterance can be measured as a prosodic property. Declination may also occur counter-intuitively, meaning that the fundamental frequency shows a positive trend[2]. To avoid ambiguities, in this thesis, positive declination values indicate a gradually rising pitch, while negative declination values represent a falling trend.

In the literature, the amount of declination is commonly estimated by fitting a straight line either through the local maxima (*topline*) or the local minima (*baseline*) of the pitch contour and calculating the angle of this linear regression line. As shown in the upper part of Fig. 4.2, these lines are not necessarily parallel, which suggests the calculation of an average declination value as a descriptive parameter:

$$declination = \frac{\measuredangle topline + \measuredangle baseline}{2} \,. \tag{4.2.1}$$

Based on the calculated declination value, the pitch contour is further *de-trended* by subtracting a (rising or falling) ramp with corresponding angle from the original pitch contour and subsequently centered around zero by subtracting the mean, resulting in some kind of “normalized” pitch contour. As visible from the lower part of Fig. 4.2, the first-order fits through both local maxima and minima are perfectly parallel to each other. It is intuitive, in my opinion, to consider this de-trended pitch contour as a suitable basis for other *level* and *range*-based parameters (see also the work of Patterson and Ladd (1999)).

[2] This was, amongst others, shown by Paeschke and Sendlmeier (2000) using the *Emo-DB* corpus.

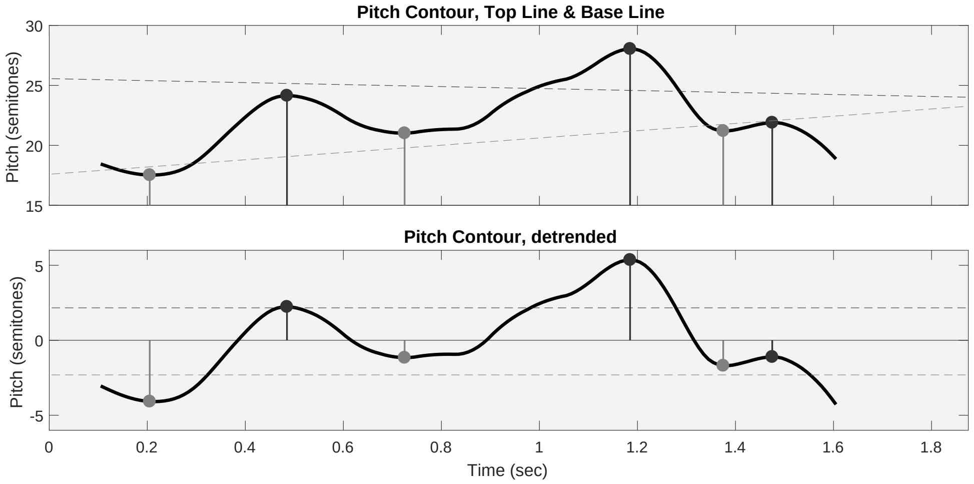


Figure 4.2.: Calculation of declination. Above: continuous pitch contour, local maxima (dark grey) and minima (light grey) with corresponding least-squares fits. Below: de-trended pitch contour centered around zero, local maxima (dark grey) and minima (light grey) with corresponding least-squares fits.

4.2.2. Pitch Onset

Pitch onset refers to the first "tone" of an utterance and is considered a general characteristic of a speaker, which means that it is believed to be virtually constant for individual speakers (Kehrein, 2002). Some studies even calculate declination as the difference of onset and *final low* (which is the last "tone" of that utterance)[3].

In any case, pitch onset is a characteristic prosodic property which I will calculate in three different ways:

1. **Absolute Pitch Onset**, which is the value of the first voiced frame in the continuous pitch contour, to test the hypothesis that pitch onset is constant, i.e., independent of the emotional state or stress level.
2. **Pitch onset with respect to Level**, which is a measure of how much the onset pitch value exceeds or falls below the average pitch value; given as the first voiced frame in the de-trended and normalized pitch contour.
3. **Pitch onset with respect to Final Low**, because both onset and final low are considered the "melodic anchor points" in a coherent linguistic expression, independent of declination as a general melodic trend.

[3] However, the example shown in Fig. 4.2 already suggests that this might be not an appropriate method; onset and final low are rather at the same pitch, although the pitch contour shows a clear positive slope.

4.2.3. Pitch Span

Rather than just calculating the standard deviation of all pitch values in an utterance, Patterson and Ladd (1999) proposed to calculate pitch range based on peaks and valleys in the pitch contour (→ 1.4.3, Fig. 4.4). Following their approach, the *pitch span* of an utterance is calculated as the difference of the average peak height and the average valley depth in the de-trended and normalized pitch contour (indicated by the dashed lines in Fig. 4.3). Doing so, I assume that a human listener implicitly considers the declination phenomenon when assessing the variation in pitch over an utterance.

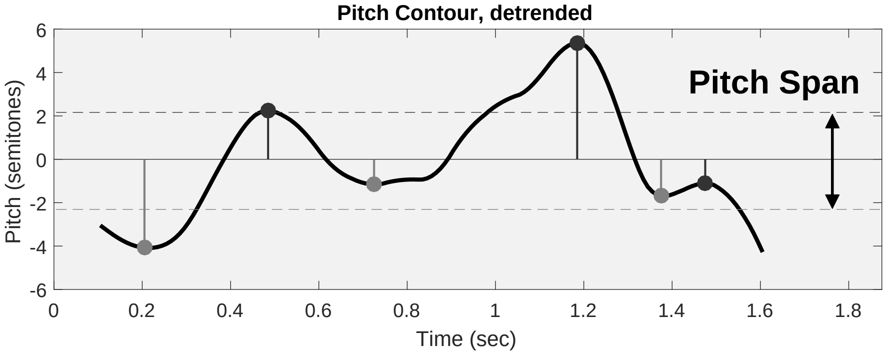


Figure 4.3.: Calculation of pitch span from the de-trended pitch contour. Shown are the local maxima (dark grey) and minima (light grey) as well as their average values (dashed lines).

I decided to call this parameter “span” rather than “range” to indicate that its calculation is somewhat different from the common linguistic pitch range measures. By using the average over all peaks and valleys, respectively, I do not favor one peak over another or claim linguistic boundary conditions to be fulfilled, as, e.g., a falling pitch trend over the utterance.

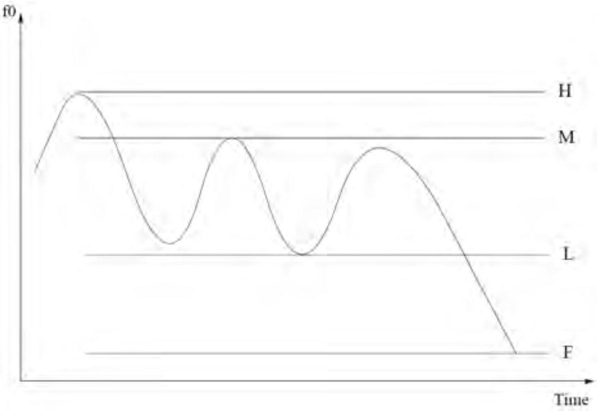


Figure 4.4.: Calculation of pitch range after Patterson and Ladd (1999).

4.2.4. Normalized Pitch Peak Extent

High dynamics in pitch are associated with expressiveness and — thought in an emotional context — also with arousal. While pitch span is a measure for the average range of pitch values during the utterance, the *normalized pitch peak extent* describes how much the highest pitch accent stands out of its environment. If you have a background in electric engineering, you might draw a parallel between this measure an the crest factor of a signal, which is the ratio of its peak value to its RMS value.

For the sake of simplicity, the normalized pitch peak extent is calculated from the de-trended and normalized pitch contour by determining the value of its highest peak, as shown in Fig. 4.5.

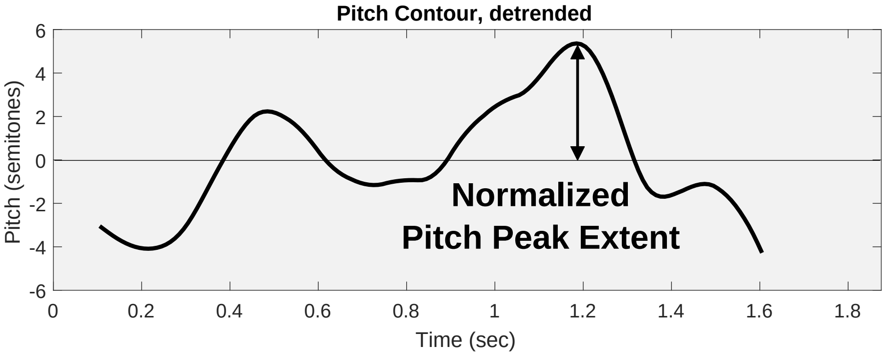


Figure 4.5.: Calculation of normalized pitch peak extent from the de-trended pitch contour.

4.2.5. Local Peak Dynamics

Another indicator of "dynamics" in the pitch contour are the absolute gradients between successive peaks and valleys, calculated as

$$peak\ dyn. = \frac{1}{N-1}\sum_{i=2}^{N}\frac{|\, f_{mel}[PV_i] - f_{mel}[PV_{i-1}]\,|}{PV_i - PV_{i-1}} \tag{4.2.2}$$

where PV_i is the time index of the i-th peak or valley. While pitch span and normalized pitch peak extent both characterize the range of values, we now also get an idea of the temporal aspects of the pitch movements using this measure.

4.2.6. Peak Timing and Peak Shape Ratios

Niebuhr (2007) studied the effects of different peak contour shapes on the meaning of an utterance by comparing peak rise and peak fall times in a categorical way (*slow* vs. *fast*). Local maxima in the pitch contour were compared with the vowel onset of an accentuated syllable and categorized as being either *early*, *medial*, or *late*. To do so, pitch and energy contours of single vowels were extracted manually and were compared to each other in the context of that very syllable. Examples for early, medial and late f_0 peaks are shown in Fig. 4.6: the solid vertical lines mark the boundaries of the syllable /maː/, while the dotted line in-between marks the increase in intensity at the beginning of the syllable nucleus. Note that the peak in fundamental frequency occurs mainly before, during, and after the accentuated syllable, respectively, in these three examples.

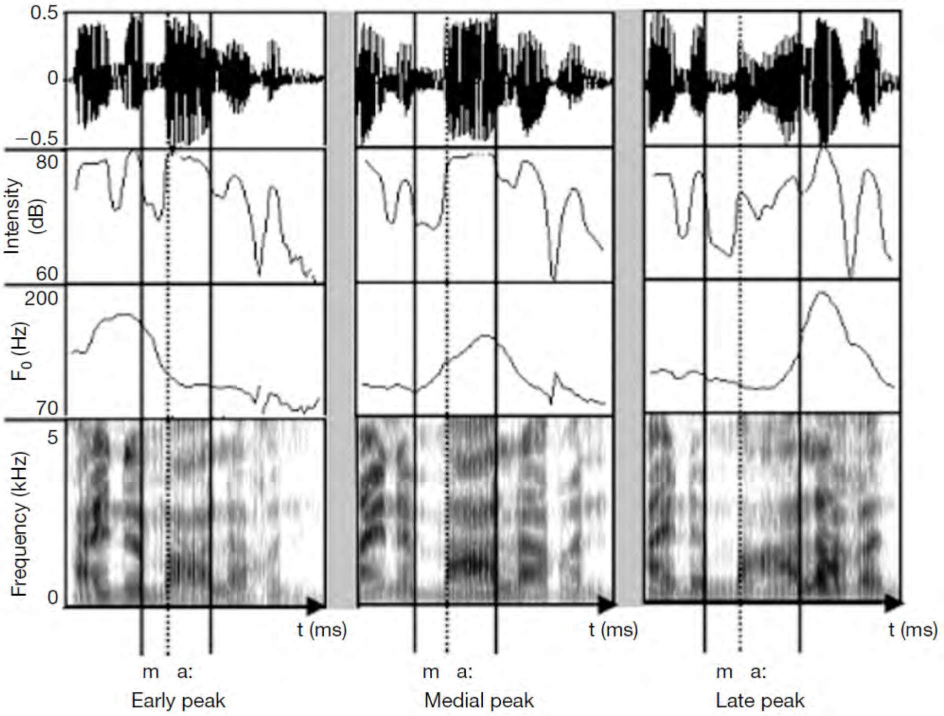


Figure 4.6.: Peak timing study by Niebuhr (2007): PCM audio, intensity and f_0 contours, and spectrogram (top to bottom) for three realizations of the utterance *Eine Malerin* ("a painter").

The hypothesis that pitch peak timing is a way of expressing the speaker's opinion on the spoken content is supported by other German linguists (Kohler, 2005). They claim that early peaks coincide with *known* or *accepted* information, while late peaks signal that the speaker disagrees with the subject of debate.

Following Niebuhr's approach, we can parametrize pitch and energy timing by comparing pitch peak times to those of the "closest highest" peaks in the loudness contour. The latter is a smoothed form of the measured short-term loudness track, which has been pre-processed in a similar way as the pitch track (→ 3.3.2). These **peak timing differences** are captured in terms of *average* and *dynamics* which are each calculated as follows:

$$PTD_{avg} = \frac{1}{N}\sum_{i=1}^{N} t_{P\_peak} - t_{L\_peak} \tag{4.2.3}$$

$$PTD_{dyn} = \frac{PTD_{std}}{PTD_{avg}} \tag{4.2.4}$$

where

$$PTD_{std} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N-1} \left(PTD_i - PTD_{avg} \right)^2} \quad .$$

What is crucial for the correct calculation of these parameters is the selection of the "appropriate" energy peak which belongs to the syllable nucleus. Due to the smooth course of the loudness contour, this works in a satisfactory manner[4], as also visible from Fig. 4.7.

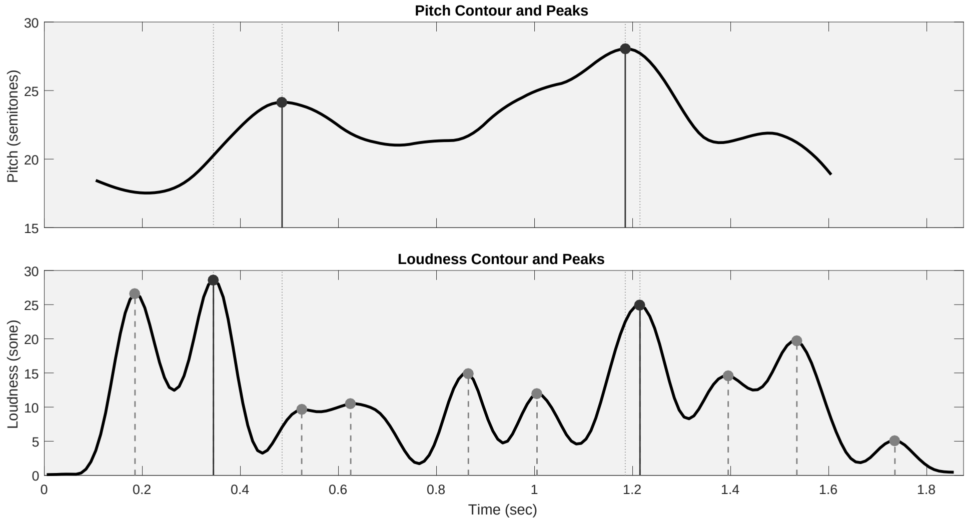


Figure 4.7.: Calculation of peak timing differences: pitch (above) and loudness (below) contours (above) and with their respective local maxima. Peaks taken for time difference calculations have dark gray markers.

With regard to the question *how "free" is free speech?* (→ 4.1.1), I think we can state that peak timing is definitely one of the freedoms a speaker has still available, regardless of pronunciation and accentuation rules.

[4] The energy peak selection performance has been checked in random samples; I have found no questionable allocation of energy peaks to pitch peaks.

Niebuhr also discovered that an early peak corresponds to a fast rise and a slow fall of the pitch contour around that peak, while a medial peak shows approximately equal rise and fall times, and a late peak is characterized by a slow rise and a fast fall. To account for this relationship, another pair of parameters is calculated in terms of *average* and *dynamics* of the **peak shape ratios** in the pitch contour which are given by the ratio of the respective rise and fall times:

$$PSR = \frac{t_{peak} - t_{valley,before}}{t_{valley,after} - t_{peak}} \quad . \qquad (4.2.5)$$

This approach is visualized in Fig. 4.8.

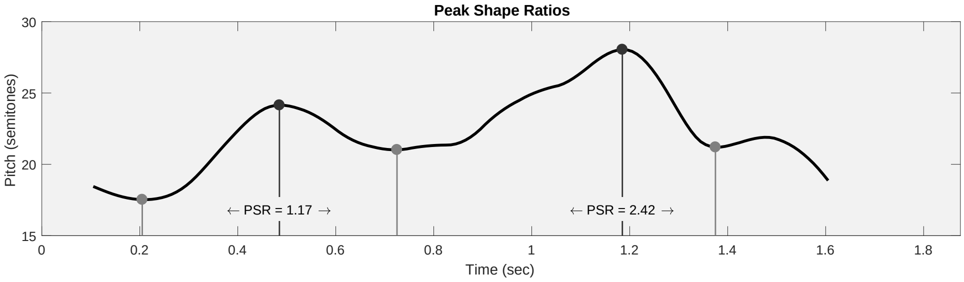


Figure 4.8.: Calculation of peak shape ratios: continuous pitch contour, local maxima (dark gray) and minima (light gray), corresponding peak shape ratios.

By analogy with peak timing differences, *average* and *dynamics* are calculated as scalar parameters

$$PSR_{avg} = \frac{1}{N}\sum_{i=1}^{N} PSR_i \qquad \text{and} \qquad PSR_{dyn} = \frac{PSR_{std}}{PSR_{avg}} \; . \qquad (4.2.6)$$

4.2.7. Utterance Harmony

At this point, we leave the background of classic phonology and enter the musical world, as I will describe my approach to capture the harmonic impression of an utterance. As already mentioned, the human voice is indeed a monophonic instrument and is usually not able to produce several tones at once; but nevertheless, one can still get a harmonic impression from successive tones produced within short periods of time. During my work on this thesis, I had to notice that, in the meantime, a similar idea had already been published by Yang and Lugger (2010). However, although the basic approach to arrive at a musical description of the relationship of tones in an utterance is similar[5], there are a few remarkable differences in the selection of the sounds which contribute to a harmonic impression.

[5] Mapping a frequency spectrum to a pitch class histogram is anyway a common technique in the field of music information retrieval and not a novel contribution by Yang and Lugger.

In musical terms, *harmony* is the interplay of musical tones, which are — in the Western world — restricted to 12 different tone qualities (C, C♯, D, D♯, E, F, F♯, G, G♯ A, A♯, B)[6]. So, the first task is to find out which *tones* are available in the sound we are about to analyze. This is commonly done by creating a so-called *pitch class histogram* which maps different frequencies belonging to the same pitch class (which is just a technical term for "musical tone") into one category; for example, an A for pitch values around 55Hz, 110Hz, 220Hz, 440Hz, et cetera. Yang and Lugger (2010) create such a histogram over the whole utterance, which obviously introduces a lot of noise due to the continuous nature of *speech melody*. As opposed to their approach, I use only the previously determined syllable pitch values (→ 3.5.2)[7] for harmony calculation, which are furthermore weighted with their corresponding prominence values before constructing the histogram.

Given an f_0 value, the corresponding pitch class can be determined in a straightforward way by calculating its corresponding MIDI pitch index,

$$p_{MIDI} = \text{round}\left(12 \cdot \log_2 \frac{f_0}{440}\right) + 69\,, \tag{4.2.7}$$

which is defined in such a way that all p_{MIDI} belonging to pitch class C are integer multiples of 12. Starting with C, every pitch class is thus assigned an index using the modulo function:

$$PC = (p_{MIDI} \bmod 12) + 1\,. \tag{4.2.8}$$

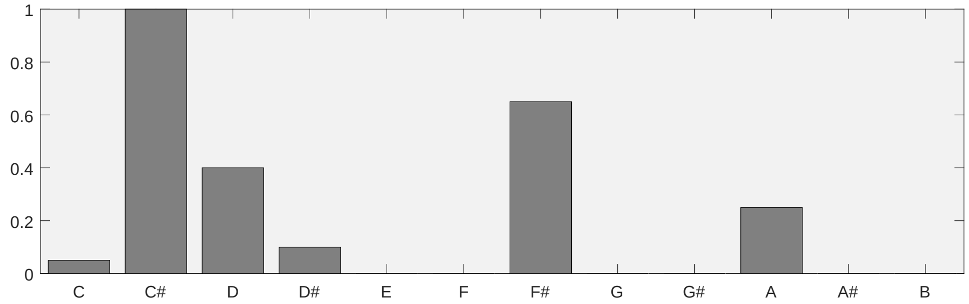


Figure 4.9.: Calculation of utterance harmony: weighted pitch class histogram for a sentence from the Emo-DB database.

The histogram is then created by adding the prominence value of each syllable to the corresponding dimension of a 12-dimensional vector which is subsequently normalized such that its maximum value equals one, as shown in Fig. 4.9.

[6] Without the context of a home key, we can assume enharmonic equivalence (E♭ = D♯).

[7] Note that, in this case, non-modified f_0 values are taken which are then translated into musical pitch; however, the determination of the perceived fundamental frequency of each syllable is following what has been described in section 3.5.2.

In order to determine the most likely chord which is composed of these tones, the pitch class histogram is correlated with a set of binary chord templates. As shown in Fig. 4.10, this set comprises 12 major and 12 minor triads as well as 3 diminished tetrachords and 4 augmented triads.

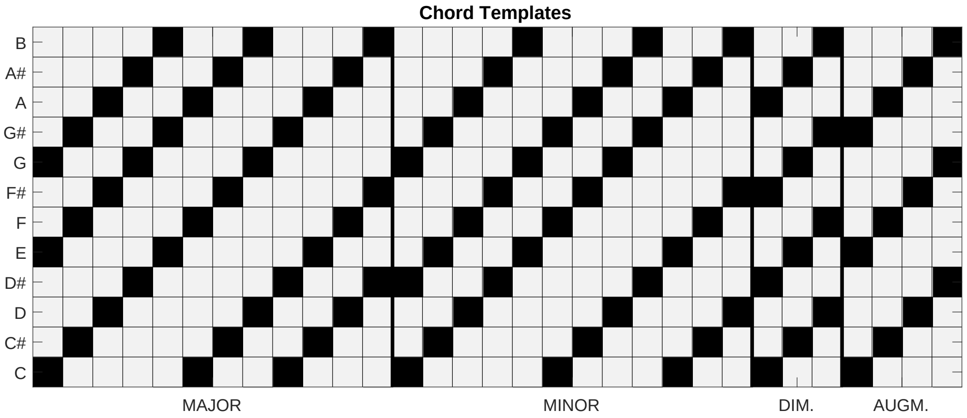


Figure 4.10.: Chord templates for harmony estimation: 12 major chords, 12 minor chords, 3 diminished chords, 4 augmented chords (from left to right).

Mathematically, the degree of correlation can be easily determined using a matrix multiplication,

$$corr = \mathbf{C}^T \mathbf{h}\,, \tag{4.2.9}$$

where **h** is a column vector containing the normalized pitch class histogram, and $\mathbf{C}^T$ is the transposed matrix visualized in Fig. 4.10 which is 1 where a black square is shown, and 0 otherwise. The result, depicted in Fig. 4.11, shows the highest correlation for `F♯ minor`, which perfectly fits to the three most prominent pitch classes in the histogram (`C♯` - `F♯` - `A`).

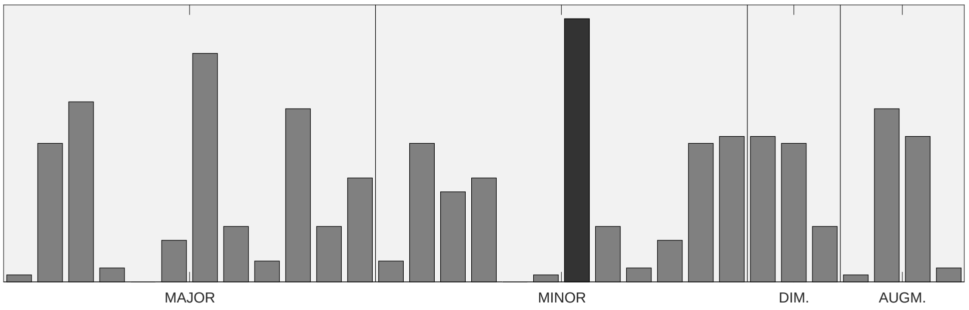


Figure 4.11.: Correlation with chord templates: 12 major chords, 12 minor chords, 3 diminished chords, 4 augmented chords (from left to right).

Two parameters are derived from the chord with the highest correlation; namely **chord identity** and **mode**, where the latter can be *major, minor, diminished* or *augmented.*

4.3. Rhythmic Parameters

The comprehension and measurement of speech rhythm is a central concern among linguists. However, it turned out that popular measures based on the percentage and the standard deviation of "vocalic" and "consonantic" time intervals (Ramus et al. (1999): $\%V$, ΔV, and ΔC; Grabe and Low (2002): *PVI*) primarily allow to distinguish between inherent rhythmic characteristics of different languages. As illustrated in Fig. 4.12, the stress-timed languages English and Dutch form a cluster in the upper left which is well separated from the syllable-timed languages cluster including Spanish, Italian, and French[8], while Japanese, as deputy for the mora-timed languages, is located far away from both clusters in the lower right of the $\%V/\Delta C$ plane.

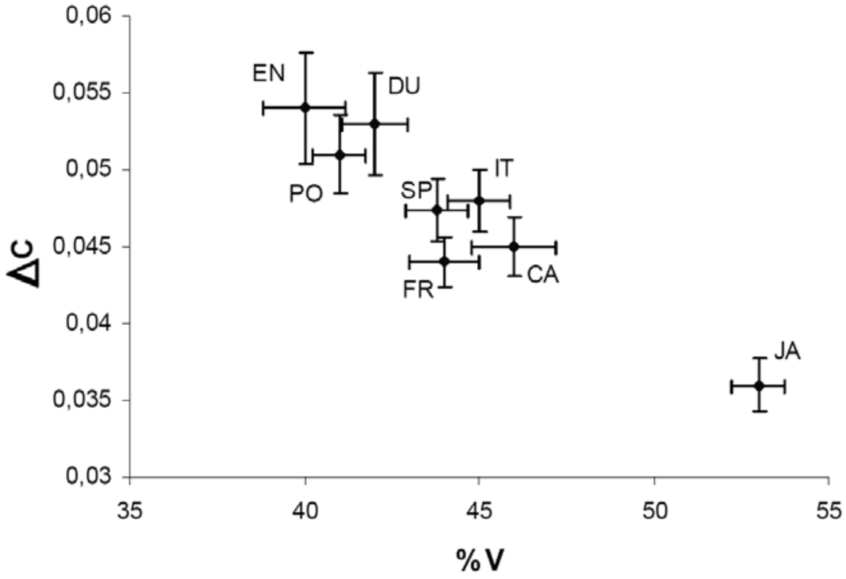


Figure 4.12.: Languages in the $\%V/\Delta C$ plane (from Ramus et al. (1999)).

Despite several interesting approaches to describe rhythm in speech — many of them of theoretical nature —, there is to date no common understanding in the scientific community on which characteristics of speech constitute its *rhythm* as the interplay of strong and weak beats grouped in a particular way. In my opinion, Cummins (2002) gets to the heart of it:

> *This variability raises the question of whether the kind of index proposed by Ramus, Grabe and others can meaningfully be said to capture anything about rhythm in speech. (...) More succinctly, where is the bomdi-bom-bom in %V?*

[8] Although clearly assignable to one of these two clusters, Polish and Catalan are considered *intermediate languages* between syllable-timed and stress-timed languages.

The rhythmic parameters presented in the following sections are based on a musical understanding of rhythm. This implies the assumption of one central property of musical rhythm which is not necessarily met by natural speech: the existence of a regular meter.

I will use the term "**accentuated syllable**" for those syllables which show a certain degree of relative prominence compared with other syllables in that utterance, to point out the difference between globally dividing prominence values into categories, independent of their occurrence in an utterance (→ 3.5.3), and judging relative prominence (= "accentuation") within the scope of an utterance. As an example, in the topmost graphic in Fig. 4.13 on page 103, the syllable at approx. 0.6s has a prominence of 17.5 which is definitely "prominent" from a global perspective, but it is shaded by the even more prominent syllable at 0.5s. Thus, the syllable at 0.5s is considered accentuated, while the syllable at approx. 0.6s is not.

4.3.1. Meter Fitting and Rhythmic Grid Creation

Any kind of rhythm requires the presence of a *meter*, meaning a continuous underlying pulse with periodical accents. Transferring this concept to speech, we implicitly assume that a speaker produces an utterance with an underlying metrical rhythm. Regarding the perceptual centers of the syllables as the beats of speech rhythm, this means that each syllable is either a multiple or a fraction of a fixed unit of time. While the assumption of a strictly continuous pulse on the syllable level might be true for poetry, it is restricted to accentuated syllables in both read prose and spontaneous speech (→ 4.1.1). The first step towards the creation of a *rhythmic grid* for an utterance is thus to estimate its meter; or rather: to fit the most likely regular pulse to the syllable beat.

Detecting "Accentuated" Syllables

The syllable beat is given by the p-centers of the syllables, of which the *accentuated* instances constitute the meter. The first challenge is now to find a threshold value for prominence level which separates "accentuated" from "non-accentuated" syllables. After some manual testing, it emerged that a constant threshold value is not applicable even for different spoken sentences of the same speaker. The solution is an adaptive threshold which is not only a function of a syllable's prominence value, but also of the distance to the surrounding syllables. It is calculated as follows:

1. Each syllable is represented by a single impulse of height 1 at its p-center which is multiplied by its prominence value.
2. This series of impulses is then convolved with a triangular window[9], resulting in several overlapping triangles of which the highest values in each time step form the threshold curve. One could interpret this as some abstract kind of "prominence masking".
3. Those scaled impulses which do not fall below the threshold curve (i.e., which are not "masked") are regarded as being accentuated. If more than 60% of the syllables are marked accentuated, the width of the triangles is increased by a factor of 1.2 and the procedure is repeated from step 2.

The topmost graphic in Fig. 4.13 displays the syllable p-centers and their prominences as vertical bars and the threshold curve as a dotted line. Those syllables which are considered accentuated are marked black. Thinking only of these accentuated syllables, one could get the impression that there is indeed some sort of immanent temporal regularity.

Finding the Most Likely Meter

The temporal differences Δt between the accentuated syllables will of course not be perfectly equal or integer multiples of each other. Our goal thus must be to find a common Δt_q (where the q stands for "quantized") as a compromise which requires the least quantization effort; that is, which introduces as little temporal shifts as possible. If we had a continuous, sinusoidal function, we could make use of its auto-correlation function to find the underlying fundamental period (→ 3.3.1). By playing a simple trick, we can create a pseudo-sinusoidal function from the pattern of prominence values over time: by convolving the single impulses with a sinusoidal kernel function of sufficient width, it should be possible to obtain usable results.

The middle graphics in Fig. 4.13 show such a pseudo-sinusoidal signal (left) and its ACF (right) using a kernel width corresponding to the minimum Δt between the accented syllables. As the convolution kernel, several popular windowing function in digital signal processing may be appropriate; I have chosen the *Hann window*

$$w_{Hann}[n] = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) = \sin^2\left(\frac{\pi n}{N-1}\right) \tag{4.3.10}$$

due to its rather basic cosine shape which produces no discontinuities at the edges.

[9] An initial width of 500ms has been determined as suitable in many cases.

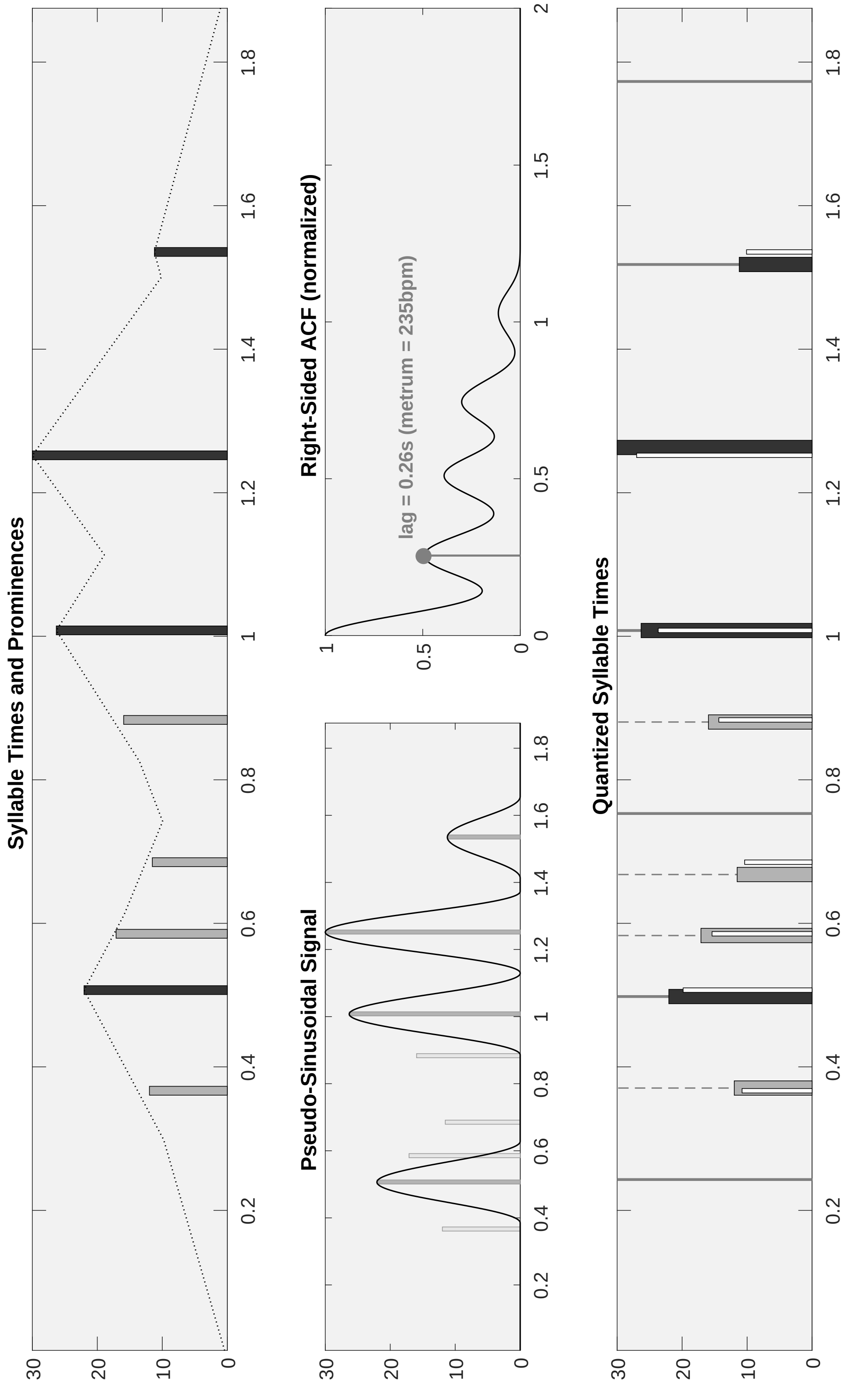


Figure 4.13.: The main steps of rhythmic grid creation. Top: syllable prominences and threshold function; “ .

By determining the lag of the second highest peak, we get the optimum $\Delta t_{q,opt}$ with respect to the lowest quantization cost.

Creating the Rhythmic Grid

In order to fit the meter beats to the accentuated syllables best possible, a regular grid is created, starting from $t_0 = 0$ in steps of $\Delta t_{q,opt}$ to a t_1 which is greater or equal to the p-center of the last syllable in the utterance. Subsequently, each of the syllables at $t = t_i$ is assigned to the closest position $t_{i,q}$ in the grid, and the overall quantization cost is calculated as

$$cost(\tau) = \sum_{i=1}^{N} \left| \, t_i - t_{i,q}(\tau) \, \right| \cdot P_i \tag{4.3.11}$$

where P_i is the prominence of the i-th of N syllables, such that shifting syllables of high prominence is penalized more than shifting less prominent syllables, and $\tau_0 = 0$ is a relative offset which is used in the remaining steps. An iterative procedure is now started by shifting the grid in steps of 1ms and calculating the quantization cost again in each step, up to $\tau_{max} = \Delta t_{q,opt}$. Doing so, we end up with a series of time shifts and the corresponding quantization costs. Now it is a trivial task to pick the τ_{opt} which leads to the lowest cost.

Next, the locations of the unaccented syllables in-between the quantized, accentuated syllables are analyzed and the best-fitting *micro grid* is determined in a similar manner as described above: if there are unaccented syllables, five different ways to subdivide the interval between the accentuated syllables (2, 3, 4, 6, and 8 divisions) are tested by creating the respective micro grid and calculating the quantization cost. In contrast to above, we are not aiming at finding an optimum time shift, but the optimum number of subdivisions. This procedure is performed for every interval between two accentuated syllables individually, accounting for the fact that the number of syllables as well as their approximate position are determined by the words to be said, rather than being chosen freely by the speaker.

The undermost graphic in Fig. 4.13 shows the original syllable positions as narrow, white bars, and the quantized positions as wide bars. The solid vertical lines mark the meter grid, and the dashed vertical lines indicate the micro grids.

Three different parameters are calculated based on this rhythmic grid, namely *meter regularity*, *tatum feel*, and *tempo*.

4.3.2. Meter Regularity

Now that we have forced the syllable p-centers (PC) into a regular temporal grid, we can judge the inherent regularity of the non-quantized rhythmic pattern by formulating a measure based on the quantization cost.

The meter regularity of an utterance is calculated as follows:

$$R = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{|\, t_i - t_{i,q} \,|}{\Delta t_q} \,. \tag{4.3.12}$$

Again, t_i and $t_{i,q}$ are the times of the i-th syllable p-center before and after quantization, respectively, and N is the total number of accentuated syllables. If the pattern had been perfectly regular before quantization, there was no quantization effort and R would equal 1. The maximum possible shift for a single syllable is $\frac{1}{2}\Delta t_q$, but this will never happen for all accentuated syllables, because in that case, another optimum time shift τ_{opt} would have been found. Values for R typically range between [0.9..1], the lowest value found in the *Emo-DB* database was 0.78.

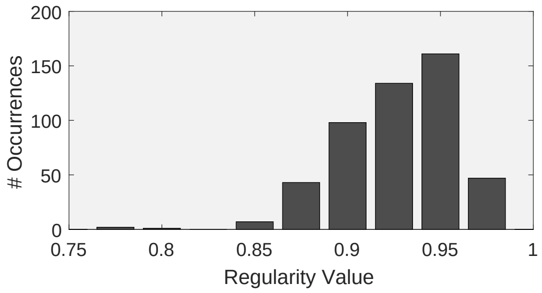


Figure 4.14.: Histogram of regularity values from all *Emo-DB* sentences.

4.3.3. Speech Tempo

In the literature, *speech rate* is often measured in syllables per second, which is comprehensible since the syllable is the backbone of speech rhythm in a linguistic sense. Moreover, since many subjects of analysis require single syllables to be identified either manually or automatically, such a measure can be calculated in a convenient way.

Although this kind of speech rate reflects a general temporal property and may indicate well if speaker A has been reading a sentence “faster” or “slower” compared to speaker B, it does not capture the speed of the pulse in a musical sense. Rhythm perception experiments are often designed as tapping tasks where test persons are asked to tap their fingers while listening to different rhythms. I am convinced that test persons in a hypothetical tempo estimation experiment would use the same technique to capture some kind of "pulse" in speech.

Having the rhythmic grid available, the calculation of speech tempo *in a musical sense* (thus a divergent name for this parameter) is as simple as

$$tempo = \frac{60}{\Delta t_q} \quad \text{... [bpm]} \tag{4.3.13}$$

where *bpm* refers to the musical unit of "beats per minute".

4.3.4. Tatum Feel

Now that we have captured the meter both in its regularity and tempo, we should have a look on the remaining parts which constitute the rhythm of speech: the unaccented syllables. By the way they subdivide the interval between two successive metric beats, they give character to a rhythm. A regular meter beat would never be called "swinging" or "syncopated" — it's all in the fine structure in-between the meter. In music research, the term *tatum* has taken root for the finest temporal unit which is perceived by a listener in the context of a rhythm; the term has been formed by Bilmes (1993) (who was searching for a synonym for this kind of *temporal atom*, by the way, and came up with "tatum" in honor of the legendary Jazz pianist Art Tatum).

Although pronunciation rules prescribe which syllables are to be given more prominence than others and thus the number of non-accentuated syllables between two metric pulses is pre-determined (see discussion in section 4.1.1), the speaker has some degree of freedom in the timing of these syllables. As described above, the micro-rhythmic grid is created by fitting an evenly distributed number of beats, dividing the temporal space between two metric pulses into 2, 3, 4, 6, or 8 pieces of same size.

Due to the fact that meter estimation is highly vulnerable to misdetections in terms of halved or doubled tempo, it seems reasonable to just make a basic distinction between **binary** and **ternary** tatum feel by checking if the majority of "best fitting" subdivisions is an integer multiple of 2 or of 3.

$$tatum = \begin{cases} 'binary' & \text{if} \quad mode(subdivisions) \in \{2, 4, 8\} \\ 'ternary' & \text{if} \quad mode(subdivisions) \in \{3, 6\} \end{cases} . \tag{4.3.14}$$

4.3.5. Prominence Dynamics

The parameters discussed up to now mainly consider the aspect of syllable timing; the *dynamics* of a rhythm pattern given through the interaction of more and less accentuated syllables have not yet been covered explicitly.

Intuitively, it seems to make sense to calculate the standard deviation of all prominence values in an utterance; just as with peak timing differences and peak shape ratios (→ 4.2.6), this standard deviation is normalized by the average prominence level:

$$PD = \frac{\sigma_{Prom}}{\mu_{Prom}} \,. \tag{4.3.15}$$

4.3.6. Legatoness

Remember the discussion on what constitutes the perceived length of a syllable in section 3.2.4? The perceived syllable length will in most cases be shorter than the duration between the syllable boundaries (which I will call the *full syllable* in this context), because these boundaries normally mark the onset of a syllable, but not explicitly the end of the previous one.

This has inspired me to calculate the *legatoness* of a syllable as the ratio of perceived to full syllable length. It is a measure of "how legato" a syllable has been uttered, and its value can never exceed 1 by definition. The scalar measure used for classification is the average legatoness over an utterance,

$$legatoness = \frac{1}{N}\sum_{i=1}^{N}\frac{l_{perceived}}{l_{full}} \,, \tag{4.3.16}$$

with l_x denoting syllable length. Typical values of legatoness range from [0.70..0.85], the lowest and highest values for sentences from *Emo-DB* are 0.60 and 0.95, respectively.

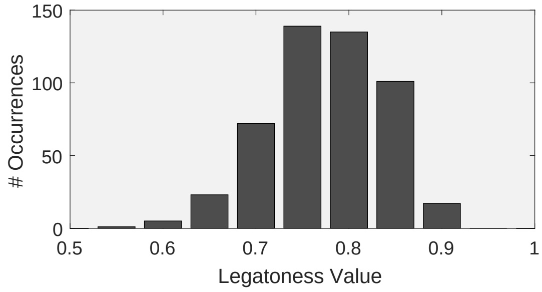


Figure 4.15.: Histogram of legatoness values from all *Emo-DB* sentences.

4.4. Paralinguistic Parameters

All the speech parameters we got to know so far are descriptors for dynamic phenomena in speech. Melody and rhythm are inherently *dynamic*, and scalar measures for these variables describe some sort of "global trend" over the utterance. Paralinguistic parameters, on the contrary, are rather static over the course of an utterance; they capture a general impression of the speakers' "voice sound".

We have learned that the characteristic "tone" of a sound is called *timbre* and that this term is hard to define exactly (→ 1.2.1). We have further learned that voicing characteristics which are associated with glottal oscillation patterns are summed up under the term *voice quality*. We will become familiar with some paralinguistic parameters in the following sections, and we will learn that they are not necessarily spectral characteristics of speech.

4.4.1. Roughness

Roughness is the impression we get when a musical sound is modulated in amplitude or frequency in the range of approximately 15Hz to 300Hz. As a psychoacoustical measure, it is commonly used to estimate the (non-)pleasantness of sounds or to evaluate the sound quality of "noisy" sounds as engine noise or the sound of an electric shaver. Roughness is measured in *asper*, which is the Latin word for "roughness". 1 asper is defined as the roughness produced by a 1kHz tone of 60dB SPL which is 100% amplitude-modulated with a modulation frequency of 70Hz (Zwicker and Fastl, 1999). Roughness is thus no spectral feature, although its effect contributes to the timbre of a sound. It depends on the center frequency of the sound as well as on the modulation frequency and the modulation index. Experiments have shown that the just noticeable difference for roughness is about 17% (Daniel and Weber, 1997).

Several models for roughness estimation have been formulated and further developed over the years. The most recent approach, to my knowledge, is that from Höldrich and Pflüger (1999) which is based on the calculation of effective modulation indexes in critical bands. An excellent description of their roughness calculation has been provided by Sontacchi (1998) which I will try to summarize in the following[10]:

[10] The model is formulated in a very general way with a lot of variable factors. The formulas in this description already include fixed values. Referring to Höldrich and Pflüger (1999), they have been like this: $t = 2$, $p = 1$, $q = 0$, and $s = 1$.

1. The PCM signal is cut into overlapping frames of 200ms which are each weighted with a Hanning window.

2. Two auditory filters which imitate the frequency response of the outer and middle ear are applied.

3. In accordance with earlier models (Aures, 1985; Daniel and Weber, 1993), the pre-filtered signal is fed into a critical-band filterbank with 47 overlapping channels ($z_i = (0.5 \cdot i)$Bark, $\Delta z = 1$Bark), whereby the contribution of every spectral component to each band is calculated individually and the threshold in quiet is considered.

4. To determine the effective modulation index of the single bands, their envelope spectra are weighted with modified Aures curves (Aures, 1985) as well as two weighting curves which are independent of the carrier frequency and which have been determined in listening tests, before being transformed back into the time domain. The effective modulation index m_i of the i-th critical band is then obtained by dividing the RMS value of the weighted envelope by its DC value.

5. The specific roughness in each band is calculated using another, band-specific weighting function $g(z_i)$ to take the dependency on the carrier frequency into account:

$$r_i = m_i^2 \cdot g(z_i)\,. \tag{4.4.17}$$

6. Finally, the overall roughness value for the current frame is calculated by considering the cross-correlation between neighboring critical bands to avoid overestimation:

$$R = c \cdot \sum_{i=1}^{47} \left(r_i \cdot k_{i,i-2}\right) \tag{4.4.18}$$

   where $k_{i,i-2}$ is the cross-correlation coefficient between the i-th and the $(i-2)$-th band, and c is a calibration factor to ensure that the 1kHz/60dB tone with f_{mod}=70Hz equals 1 asper.

As for peak shape ratios and peak timing differences (→ 4.2.6), *average roughness* and *roughness dynamics* are calculated as scalar parameters from this time series of roughness values; the latter being defined as the mean normalize by the standard deviation.

4.4.2. Sharpness

Broadly speaking, *sharpness* describes the proportion of high-frequency content in a sound mixture. Just as roughness, it is used to describe the sensory pleasantness of a sound, and just as loudness, it is a ratio quantity in that sense that one sound can be "twice as sharp" as another. The main influence on the sharpness of a sound is its overall spectral envelope, whereas the spectral fine structure play a minor role (Zwicker and Fastl, 1999).

The spectral envelope can be described by the specific loudness pattern over critical bands (→ 3.2.3). A popular model for sharpness has been formulated by Zwicker (1982) who uses the weighted first moment of the specific loudness divided by the total loudness:

$$S = 0.11 \cdot \frac{\int_{z=0}^{24Bark} N' g(z) z \, dz}{\int_{z=0}^{24Bark} N' \, dz} \tag{4.4.19}$$

where

$$g(z) = \begin{cases} 1 & \text{for } 0 < z \leq 16Bark \\ 0.066 \cdot e^{0.171z} & \text{for } 16 < z \leq 24Bark \end{cases}$$

and the factor of 0.11 normalizes the sharpness of a narrow-band noise (≤ 1Bark) with a center frequency of 1kHz at 60dB SPL to 1 *acum*, which is the Latin word for "sharpness". The shape of the weighting function $g(z)$ demonstrates the influence of high frequencies on the impression of sharpness.

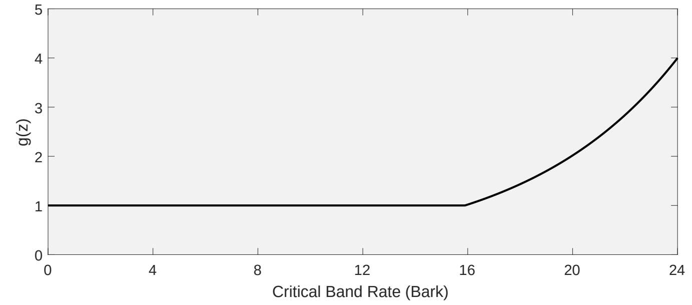


Figure 4.16.: Weighting curve $g(z)$ for sharpness calculation after Zwicker (1982).

4.4.3. Harmonics-to-Noise Ratio

In section 3.3.1, we have learned that the auto-correlation function of a signal can be used to find the most prominent frequency in the complex signal mixture by determining the time lag of the second largest peak in the ACF which is the reciprocal of the fundamental frequency.

The auto-correlation at $\tau = 0$ is the integral over the signal multiplied with itself which is equal to the signal's power. In the same way, one can argue that the normalized autocorrelation at the second largest peak indicates the relative power of the harmonic signal component (Boersma, 1993). This is plausible when looking at Fig. 4.17 which shows three different signals and their corresponding right-sided autocorrelation functions (ACFs). The ACFs have been normalized such that the maximum value at $\tau = 0$ equals 1. The first signal is an extract from an *Emo-DB* sentence which comprises the syllable /la/ spoken by a male speaker (upper left). Its corresponding ACF clearly shows a harmonic structure, and the second largest peak has a value of approx. 0.32 (upper right). Next, we have a pure sinusoid of equal length which shows perfect periodicity in the ACF, and its second largest peak (as all the others) equals 1. Finally, white noise[11] has an ACF which shows one single peak at zero lag and virtually no periodicity at all, so the second largest peak is very small.

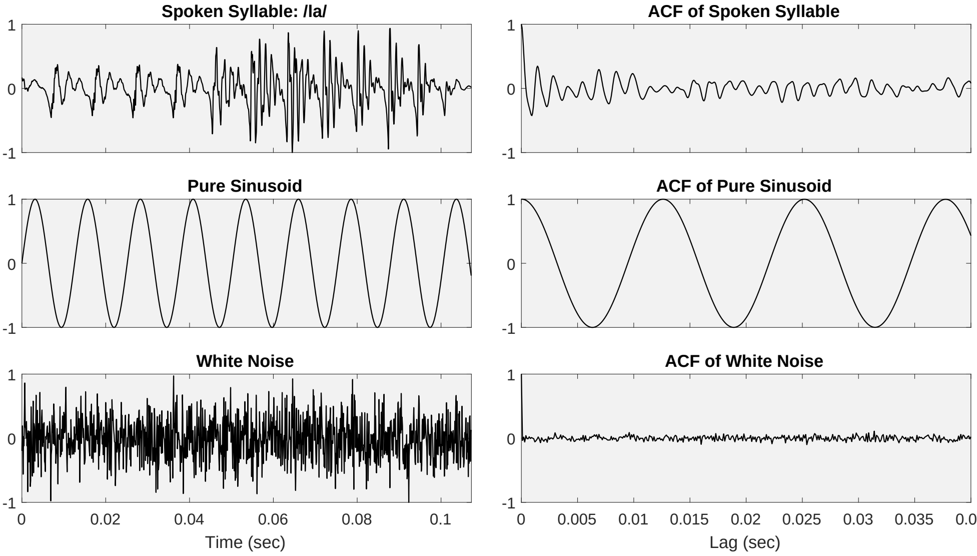


Figure 4.17.: Different signals (left) and their normalized right-sided auto-correlation functions (right). Top: speech signal excerpt, voiced syllable /la/. Middle: pure sinusoid with f = 500Hz. Bottom: white noise.

If we denote the normalized auto-correlation function of the signal $x(t)$ with r'_x and the lag τ of the largest peak except for the zero lag with τ_{max}, we can — following Boersma (1993) — write the logarithmic *harmonics-to-noise ratio* in dB as

$$HNR = 10 \cdot \log_{10} \frac{r'_x(\tau_{max})}{1 - r'_x(\tau_{max})} \,. \tag{4.4.20}$$

[11] White noise is a random signal which is absolutely uncorrelated.

4.4.4. Voice Quality Measures

The term *voice quality* is often associated with disordered voice, but is not necessarily restricted to this pathologic sense. Voice quality generally refers to voicing characteristics, which are associated with different vibratory patterns of the glottis. Depending on the aperture of the arytenoid cartilages[12], different phonation types are realized (Ladefoged, 1971), as sketched in Fig. 4.18. The most common phonation type is *modal* which yields maximum vibration of the vocal cords through an optimal combination of airflow and glottal tension.

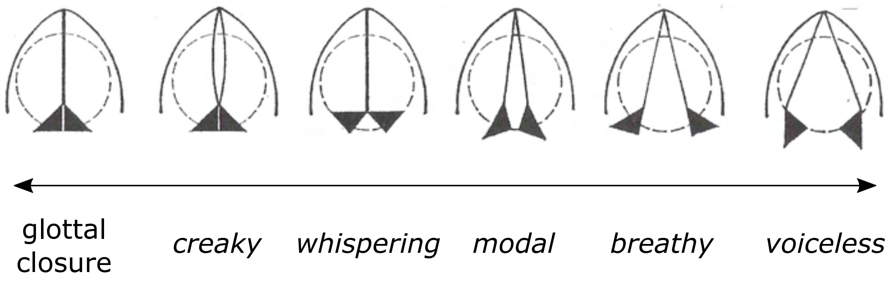


Figure 4.18.: Continuum of phonation types from closed to open glottis (after Ladefoged (1971), pictures taken from (WikimediaCommons, 2005)). The triangles represent the arytenoid cartilages, the lines connected to these triangles are the vocal cords.

Voice quality is usually judged by voice therapists using established scales, e.g., the German *RBH Index* for the assessment of roughness, breathiness and hoarseness (Nawka, 1987). Though being rated subjectively, there are correlations with the psychoacoustic parameters *roughness* and *harmonics-to-noise ratio*.

Some voice quality attributes are commonly associated with emotional terms, such as a breathy voice with intimacy, a whispery voice with confidentiality, or a creaky voice with boredom (Laver, 1980). However, different experimental studies came up with different mappings between voice quality and emotional categories (Gobl and Chasaide, 2003).

Describing the Glottal Excitation Pattern

To better understand the nature of the parameters which will be presented in the following paragraphs, we should take a quick look on the phonation process itself. Fig. 4.19 sketches a glottal cycle during modal phonation. As visible in the upper plot, the modulation of the glottal area (approximately)

[12] The arytenoid cartilages are a part of the larynx and control the movement of the vocal folds.

follows a triangular shape. Due to the inertia of the air mass inside the glottis, the resulting airflow caused by the subglottal pressure is skewed to the right. The lower plot shows the resulting output sound. Interestingly, the intensity of the output sound is mainly driven by the abrupt closing of the glottis (Honda, 2008). Incomplete glottal closure results in a "softer" voice due to less intense excitation of the vocal tract.

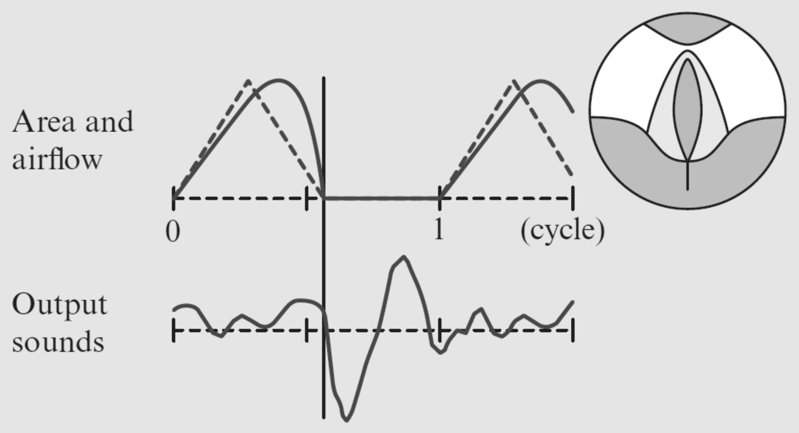


Figure 4.19.: Sketch of a glottal cycle during modal phonation (from Honda (2008)). Above: glottal area (dashed) and airflow through the glottis (solid). Below: output sound. The vertical line marks the glottal closing.

The glottal excitation pattern can be assessed with an *electroglottograph* (EGG) which indirectly measures the motion of the vocal cords during speaking through changes in electrical conductivity between two electrodes placed directly on the throat (Lecluse et al., 1975). Stevens and Hanson (1995) showed that it is also possible to derive voice quality measures directly from the acoustic signal. Their measures are based on spectral properties of the **excitation spectrum**: if we are able to decompose the speech signal into a source part and a filter part (see 3.1), it must be possible to subtract the contribution of the filter part from the recorded speech signal to end up with the glottal excitation signal. The influence of the vocal tract is modeled by the contribution of the first four formants to the Fourier spectrum. Following Fant (1970), the effect of the i-th formant F_i with bandwidth B_i at frequency f can be expressed as

$$V(f, F_i, B_i) = 20 \log \frac{F_i^2 + (\frac{B_i}{2})^2}{\sqrt{\left((f - F_i)^2 + (\frac{B_i}{2})^2\right)\left((f + F_i)^2 + (\frac{B_i}{2})^2\right)}} \,. \tag{4.4.21}$$

Using this formula, the vocal-tract compensated amplitude spectrum $|\tilde{X}(f)|$ can be calculated from the classic Fourier amplitude spectrum by subtracting the contributions by the first four formants:

$$\left|\tilde{X}(f)\right| = |X(f)| - \sum_{i=1}^{4} V(f, F_i, B_i) . \tag{4.4.22}$$

The acoustic parameters Stevens and Hanson introduced are basically quotients of harmonic peak amplitudes in the vocal-tract compensated amplitude spectrum. Since amplitude is commonly drawn on a logarithmic scale (in decibels), these quotients turn into differences. In accordance with the relevant literature on voice quality, I will use these symbols in the following:

H_1, H_2	Amplitudes at f_0 and $2 \cdot f_0$
$F_{1,p}$, $F_{2,p}$, $F_{3,p}$	Frequencies of spectral peaks close to F_1, F_2, F_3
$A_{1,p}$, $A_{2,p}$, $A_{3,p}$	Amplitudes of spectral peaks close to F_1, F_2, F_3

These are the measures of voice quality and their acoustic correlates:

- The **Open Quotient (OQ)** is that fraction of the glottal period during which the glottis is open. It corresponds to the difference in amplitude between the first two harmonic peaks[13] in the excitation spectrum, $\tilde{H}_1 - \tilde{H}_2$.
- The **Glottal Opening (GO)** describes how wide the glottis opens over the cycle. Its acoustical correlate is the amplitude of the first formant in relation to the amplitude of the first harmonic, $\tilde{H}_1 - \tilde{A}_{1,p}$.
- **Skewness (SK)** is the degree of asymmetry of the glottal flow curve. The more skew, the more abrupt the glottis closes. This is reflected in the amplitude of the second formant, again related to the first harmonic ($\tilde{H}_1 - \tilde{A}_{2,p}$).
- The **Rate of Closure (RC)** corresponds to the velocity of the glottal closure and can be approximated by the difference in amplitude between the third formant and the first harmonic, $\tilde{H}_1 - \tilde{A}_{3,p}$.
- If the glottis is not completely closed during phonation, this **Incompleteness of Closure (IC)** leads to a loss of energy in the F_1 range which is reflected in the bandwidth of the first formant (B_1). Since F_1 and thus B_1 depend highly on the vowel produced, a normalized measure is calculated as $\frac{B_1}{F_1}$.

[13] The fundamental frequency is regarded as the first harmonic, while $2 \cdot f_0$ (the first *overtone*) is the second harmonic.

Figure 4.20 shows an original short-time amplitude spectrum with marked values at f_0, $2f_0$, F_1, F_2, and F_3 (in gray) as well as the adjusted values at f_0, $2f_0$, $F_{1,p}$, $F_{2,p}$, and $F_{3,p}$ (in black). It is clearly visible that the vocal-tract compensated values are not only lower than the original values, but sometimes also significantly adjusted in frequency, especially regarding the difference between F_1 and $F_{1,p}$.

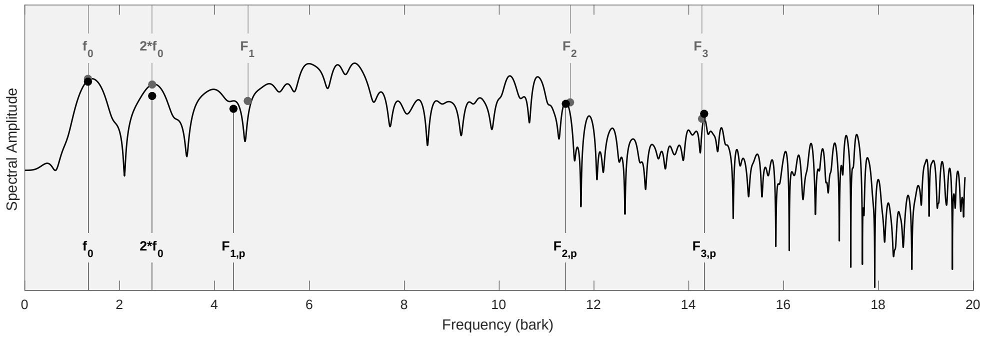


Figure 4.20.: Spectral properties for voice quality measure calculation: vocal tract-compensated amplitudes at f_0, $2 \cdot f_0$, F_1, F_2 and F_3.

Calculation of Voice Quality Measures

Lugger et al. (2006) implemented their own versions of Stevens' and Hanson's voice quality parameters and successfully used them for emotion recognition under several background noise conditions. They argue that, just as for *incompleteness of closure*, also the measures based on spectral amplitude differences should be normalized by the corresponding frequency differences to remove vowel dependency. I agree on their line of argumentation that spectral gradients will characterize the shape of the glottal spectrum better than amplitude ratios alone; the five voice quality measures are thus calculated in the following way:

$$OQG = \frac{\tilde{H}_1 - \tilde{H}_2}{f_0} \tag{4.4.23}$$

$$GOG = \frac{\tilde{H}_1 - \tilde{A}_{1,p}}{F_{1,p} - f_0} \tag{4.4.24}$$

$$SKG = \frac{\tilde{H}_1 - \tilde{A}_{2,p}}{F_{2,p} - f_0} \tag{4.4.25}$$

$$RCG = \frac{\tilde{H}_1 - \tilde{A}_{3,p}}{F_{3,p} - f_0} \tag{4.4.26}$$

$$IC = \frac{B_1}{F_1} \, . \tag{4.4.27}$$

The four corresponding gradients of the example from Fig. 4.20 are visualized in Fig. 4.21. Note that the y-axis has been scaled for better readability, thus the gradients seem to be steeper than in Fig. 4.20.

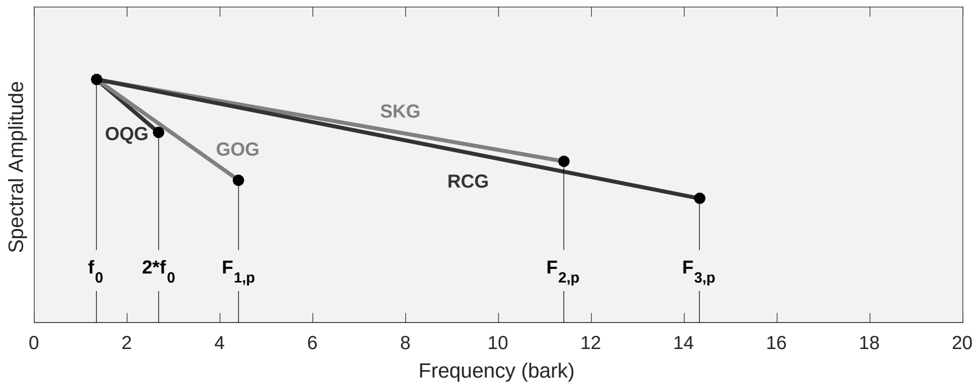


Figure 4.21.: Gradients calculated from spectral amplitude differences, after Lugger et al. (2006).

4.4.5. What about ... ?

If you have a background in digital signal processing, you might miss one or another parameter which is frequently featured in other publications. Let me explain why I have not considered the following potential speech parameters in this work:

- **Mel-Frequency Cepstral Coefficients (MFCCs)** are very popular in speech recognition, because they represent many spectral characteristics of a sound in a compact way. However, MFCCs are *segmental* features, since they change significantly with every single phone (Nwe et al., 2003), and thus depend on the spoken content.

- **Fluctuation strength** is one of the four "elementary auditory sensations" in psychoacoustics, next to loudness, roughness, and sharpness. It is similar roughness, but describes the perception of slower amplitude-modulated sounds (up to f_{mod} = 20Hz). Since speech sounds fluctuate very rapidly, this modulation frequency range is not of interest for speech analysis.

- **Jitter** and **shimmer** describe cycle-to-cycle variations in the oscillation of the vocal cords. Jitter characterizes irregularities in frequency, while shimmer is a measure for deviations in amplitude; both are thus measures of frequency and amplitude modulation phenomena, which is well captured by the sophisticated calculation of roughness.

Index	Category	Parameter Name
1	melodic	Pitch Onset (absolute)
2	melodic	Pitch Onset (w.r.t. Level)
3	melodic	Pitch Onset (w.r.t. Final Low)
4	melodic	Normalized Pitch Peak Extent
5	melodic	Pitch Span
6	melodic	Declination
7	melodic	Peak Shape Ratios: Average
8	melodic	Peak Shape Ratios: Dynamics
9	melodic	Local Peak Dynamics
10	melodic	Peak Timing Differences: Average
11	melodic	Peak Timing Differences: Dynamics
12	melodic	Utterance Harmony: Chord
13	melodic	Utterance Harmony: Mode
14	rhythmic	Meter Regularity
15	rhythmic	Tatum Feel
16	rhythmic	Prominence Dynamics
17	rhythmic	Tempo
18	rhythmic	Legatoness
19	paralingustic	VQM – Open Quotient Gradient
20	paralingustic	VQM – Glottal Opening Gradient
21	paralingustic	VQM – Skewness Gradient
22	paralingustic	VQM – Rate of Closure Gradient
23	paralingustic	VQM – Incompleteness of Closure
24	paralingustic	Roughness: Average
25	paralingustic	Roughness: Dynamics
26	paralingustic	Sharpness
27	paralingustic	Harmonics-to-Noise Ratio

Table 4.1.: Complete list of prosodic and paralinguistic parameters. *VQM* stands for *Voice Quality Measures*, "averages" are arithmetic means and "dynamics" are standard deviations normalized by the means.

5. Results

Classification is the process which assigns a sample to a specific category based on a variety of the sample's features. These features must be chosen carefully to obtain good classification results. In this chapter, the best-performing set of prosodic and paralinguistic parameters for classification of both emotional speech and speech under stress are presented.

5.1. Classification and Parameter Selection

Now that we have found 27 different parameters which describe *the way we say it* up to a certain degree, we are interested in finding out how well these parameters reflect the emotional state or the stress level of the person speaking. Because we have learned from the literature that this is a non-trivial task at all, we try to keep things rather simple by sticking to emotional categories on the one hand and to a before/after comparison of stressful events.

5.1.1. Classification

Classification is the process which assigns a given sample to a specific category based on a variety of features. In the context of this work, a sample is an sentence from the *Emo-DB* or the *IEM-PSD* databases; and the categories are the seven basic emotions or the three regions before/during/after a potential stressful event, respectively. The features are given by the 27 prosodic and paralinguistic parameters which have been calculated for each utterance. In mathematical terms, we project each utterance onto one point in a 27-dimensional parameter space where *similar* objects form clusters of data points, while *dissimilar* objects are farther away from each other. A simple example is given in Fig. 5.1, where apples and bananas are represented by two features including color and shape (expressed by the length-to-width ratio). Obviously, color is not a very strong feature for classification purposes, since their value ranges overlap. On the contrary, the "lengthiness" of a fruit allows a clear distinction between apples and bananas. One could draw a straight line between the two clusters of triangles and circles which perfectly separates them in space.

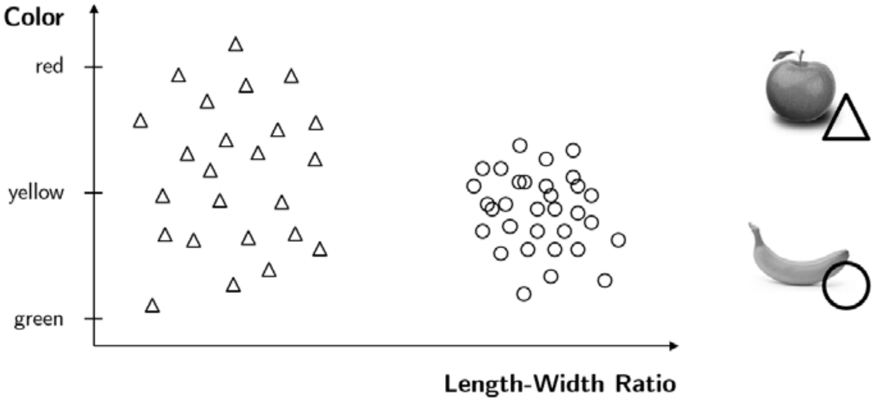


Figure 5.1.: Example for classification: apples (triangles) and bananas (circles) are represented by their colors and shapes as data points in a two-dimensional parameter space.

There are two ways to characterize the distribution of data points in the parameter space. **Parametric classification methods** describe the data distribution through *probability density functions* (pdf), which minimizes the probability of a wrong decision and — theoretically — always results in an optimum solution. However, it must be ensured that the data actually follows the assumed distribution, which is not always the case. Usually, the pdf is assumed to be multivariate Gaussian, such that a cluster of data points is described by means and covariances for each dimension of the parameter space. **Nonparametric classification methods** use the distances between single data points as a criterion. If an unknown fruit was to be classified into the "apple" or the "banana" class in the example mentioned above, it would be represented by an additional data point in Fig. 5.1, and one would either calculate the *likelihood* of belonging to one or the other class, or simply measure the distance to the closest data points in order to assign a class label to that fruit.

The ultimate goal in classification is in either case to subdivide the parameter space into separate regions which each belong to one class and to formulate mathematical models which describe the *decision boundaries* between the classes. Classification is a fundamental task in many areas of computer science, and thus the number of methods is manifold. The complexity of this field is, in my opinion, best summarized by Batliner et al. (2011a):

> *Finding, fine-tuning, and evaluating classifiers is a broad topic in its own; although there might be preferences to use one or the other*

> *approach in specific fields — such as emotion recognition, it generally suffers from too many degrees of freedom: a strict comparison across studies is practically never possible. Statements such as "it has been proved that classifier X is superior to classifier Y", should never be generalized. Often it only means that there has been more fine-tuning for X than for Y. In the long run, it might turn out that specific models and classifiers based on them are — on the average — better suited for emotion recognition. However, searching for an optimal classifier alone will not be a panacea; it will not improve unsatisfying recognition rates to such an extent that the intended application will be successful. Anyway, it should be mandatory to document the steps explicitly, e.g., whether a cross-validation has been done speaker-independently or in a speaker-dependent way. This statement holds similarly for comparison across whole studies: what never should be done is simply to compare recognition rates between two studies. Such performance depends crucially on too many factors which have not been standardized yet.*

A comprehensive review on classifiers used for emotional speech recognition has been provided by Anagnostopoulos et al. (2015), including Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Decision Trees, and k-Nearest Neighbor distance classifiers (kNN). I will not compare different classifiers against each other, as this is not the focus of my work; I have chosen Support Vector Machines as a classifier due to their ability to minimize the empirical classification error while maximizing the *geometric margin* between the classes at the same time (→ 5.1.5).

5.1.2. Parameter Selection

What is crucial in any case is the selection of relevant parameters for classification. The set of prosodic and paralinguistic parameters has been compiled in all conscience, following the relevant literature in the fields of linguistics and psychoacoustics as well as some "musical" properties of intonation and speech rhythm. There is, however, no evidence at this point that these parameters contain any descriptive power with respect to the emotional state of the speaker; in other words: we may *assume* that, e.g., speech rhythm regularity is affected by emotions, but we do not *know* if this is the case.

If we include weak or meaningless parameters in our analysis, we add a lot of noise to the results — think of the fruit color which does not allow to separate apples from bananas if it's not "red" by chance. The difference in distances or probabilities will be more concise if only calculated along the x-axis rather than in two-dimensional space. In addition, we have to be aware of the *curse of dimensionality* (Duda et al., 2012): when the dimensionality of the abstract parameter space increases, the volume of that space increases exponentially, and the available data become sparse in comparison. Moreover, the classifier will tend to over-fit to the data. As a consequence, the classification results are specific to the training data used and thus not of general use. The goal is therefore to find the smallest possible set of parameters which each show the highest possible difference in value for one or the other emotion. This is also a common task in computer science, often referred to as *feature selection*[1].

In mathematical terms, we have a parameter set

$$X = \{x_i \mid i = 1..N\} \ ,$$

and we are looking for a sub-set

$$Y_M = \{x_{i1}, x_{i2}, \ldots, x_{iM}\} \, (M < N)$$

such that

$$\{x_{i1}, x_{i2}, \ldots, x_{iM}\} = argmax \, [J \, \{x_i \mid i = 1..N\}] \ ,$$

where $J \, \{x_i \mid i = 1..N\}$ represents a target function which could be both a statistical measure or the accuracy of a classification result obtained using the sub-set Y_M[2].

On the one hand, it is by far not sufficient to perform classification with single parameters one-by-one, because this would completely ignore the covariances between the parameters. On the other hand, a *full-factorial* search through the parameter space where all $2^N - 1$ possible combinations of parameters are evaluated would result in 134,217,727 calculation cycles, which is obviously infeasible to do. The compromise is called **sequential parameter selection**: starting from an empty set of parameters,

$$Y_0 = \emptyset \, ,$$

we sequentially add parameters which — in combination with those already selected — maximize the value of the target function:

$$x^+ = \underset{x \in (X \setminus Y_k)}{argmax} \, [J(Y_k \cup x)]$$

$$Y_{k+1} = Y_k \cup x^+ \ .$$

[1] Its counterpart, *feature extraction*, transforms the high-dimensional parameter space into a lower-dimensional space using techniques like, e.g., Principal Component Analysis (PCA). Since we want to preserve the meaning of the parameters in order to answer the question which parameters are suitable for emotion and stress recognition, feature selection is the method of choice.

[2] Depending on the formulation of J, the condition for the ideal subset could also be written using the $argmin()$ operator

This iterative method is aborted once the target function saturates. While conceptually simple and straightforward, this concepts suffers from *nesting*: once a feature is part of the set, it can not be removed again. An alternative approach is to start with the full parameter set and to sequentially remove single parameters; in this context, we talk about sequential *forward* and *backward* selection methods. There also exists a variety of extensions to these methods to overcome the nesting problem.

5.1.3. Training and Testing

A *classifier* is an algorithm which classifies unknown objects into a number of predefined classes based on certain classification rules. These rules are implemented into the classifier by *training* it using data with known class affiliation. Depending on the distribution of data points in the training set, the classifier might be able to separate the training data perfectly or not; we can calculate its **in-sample classification performance** either as a measure of suitability of the chosen algorithm for that classification problem, or as a measure of data separability.

This, however, doesn't tell us anything about its **out-of-sample classification performance**, which is the precision in classification of unknown data. The classifier might have adapted perfectly to some characteristics of the training data which are no generalities, and thus might show poor performance when faced with independent data. To avoid the risk of *overfitting*, a common trick is to partition the training data set into complementary subsets, training the classifier on one subset and validating it on the other. In *k-fold* ***cross-validation***, multiple cycles of training and testing are performed using different partitions, averaging the validation results over the cycles.

5.1.4. Parameter Preprocessing

Among the 27 prosodic and paralinguistic parameters listed in Tab. 4.1 on page 117, there are two *categorical* parameters which are, in this way, not suitable for numerical classification:

- **Tatum Feel** is either *binary* or *ternary*, which can be translated into the numbers 2 and 3, respectively. By doing this, we indeed invalidate the rhythmic meaning of these two terms, but we still might find differences in speaking style between emotions or stress levels if there are significant deviations between single classes. For example, an average value of 2.17 could be interpreted as “rather binary”, while 2.89 would be “rather ternary”.

- **Utterance Mode** can be *major*, *minor*, *diminished*, or *augmented*. When assigning numeric values to each of these categories, we should keep in mind that the difference between two values is equivalent to some kind of similarity between the two corresponding classes. A reasonable assignment might thus be the following:

augmented	1
major	2
minor	3
diminished	4

  Thinking of triads, all neighboring chords differ in one tone only, and their harmonic characteristics are “as similar as possible”. I am sure we will all agree that augmented and diminished harmonies are the “most dissimilar” items from this list and thus belong to both ends.

In addition to this, unfortunately, **Utterance Chord** turned out to be inappropriate at all. The relationship between major and minor chords in diatonic function is of *relative* kind, meaning that any chord is “musically close” to some other chord by, e.g., forming its parallel chord (as $C - Am$), regardless of its absolute identity, say *C* or *F#*. Some kind of musical distance could be coded using chord distances on the circle of fifths, but its circular shape cannot be mapped onto absolute positions in the abstract parameter space. As a consequence, chord quality has to be neglected, while its mode will be used as describe above.

5.1.5. Support Vector Machines

Support Vector Machines (SVM) are referred to as a *large margin classifier*, aiming at the creation of decision boundaries with the largest distance to the nearest training data point of any class. The idea behind maximizing this functional margin is to lower the generalization error of the classifier, that is, making it as robust as possible.

This is sketched in Fig. 5.2 which shows the optimum decision boundary versus two sub-optimal decision boundaries (small plots) in terms of a maximum margin for an exemplary distribution of data points. The boundary is surrounded by two parallel, equidistant margin lines (dashed) which touch the closest data points from both classes. These data points are the only relevant points for the positioning of the boundary, and they are called the *support vectors* (thus the name).

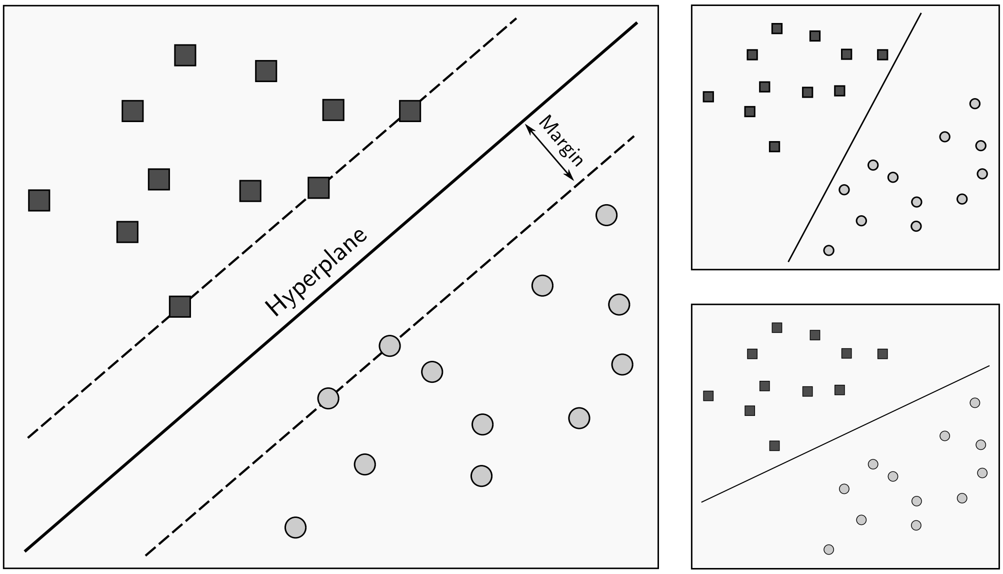


Figure 5.2.: Support Vector Machine classification: distribution of data points belonging to two different classes and different ways to draw a decision boundary: optimal (left) and suboptimal cases (right).

In this simple two-dimensional example, the data points are separated by a straight line. In the general p-dimensional case, each data point is characterized by a p-dimensional vector, and the decision boundary is a $(p-1)$-dimensional hyperplane. An SVM is thus a binary, linear classifier which can be adapted to the multi-class, nonlinear case in the following way:

- We either train M binary classifiers which separate one class from all the others (*one-versus-all*), or we train one classifier for each pair of classes, ending up with $\binom{M}{2}$ classifiers (*one-versus-one*).
- Misclassifications are allowed, but penalized with an adjustable weight.
- The original feature space is transformed into a higher-dimensional feature space where the data are linearly separable; this is known as the "kernel trick" and has been introduced by Boser et al. (1992).

Further details on Support Vector Machines can be found, for example, in the SVM tutorial by Burges (1998), and on `www.kernel-machines.org`.

5.2. The Prosody of Emotional Speech

5.2.1. Investigated Speech Data

The *Berlin Database of Emotional Speech*, or simply *Emo-DB*, was created by Burkhardt et al. (2005). Ten different German sentences (5 long, 5 short) were produced by 10 actors (5 male, 5 female) in 6 different emotions (*angry*, *anxious*, *bored*, *disgusted*, *happy*, *sad*) as well as in a *neutral* version. The sentences, listed in Tab. 5.1, are emotionally neutral and consist of everyday vocabulary, such that they make sense in every realized emotion. For some sentences, several versions in the same emotion have been recorded, resulting in a total of approximately 800 sentences.

1	Der Lappen liegt auf dem Eisschrank. *The cloth is lying on the frigde.*
2	Das will sie am Mittwoch abgeben. *She wants to hand it in on Wednesday.*
3	Heute abend könnte ich es ihm sagen. *Tonight, I could tell him.*
4	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück. *The black piece of paper is up there besides the piece of wood.*
5	In sieben Stunden wird es soweit sein. *In seven hours, it will be ready.*
6	Was sind denn das für Tüten, die da unter dem Tisch stehen? *What about these bags under the table?*
7	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter. *They just carried it up, and now they are going down again.*
8	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht. *The last weekends, I always went home to see Agnes.*
9	Ich will das eben wegbringen und dann mit Karl was trinken gehen. *I just want to take this away, and then go for a drink with Karl.*
10	Die wird auf dem Platz sein, wo wir sie immer hinlegen. *It will be at the place where we always put it.*

Table 5.1.: Sentences from Emo-DB with English translation.

These recorded sentences were subsequently presented to 20 test subjects who had to recognize the emotion and were asked to rate the authenticity of the acted emotions. Only those sentences were selected which showed a correct recognition rate of at least 80% and which were considered "authentic" by at least 60% of the listeners. The database finally comprises 535 sentences in total and comes with phonetic transcriptions both on the phoneme and on the syllable level. It is publicly available via the internet: `http://www.expressive-speech.net/emodb/`.

While checking the integrity of the data, I found out that 42 of the 535 sentences have missing or corrupt syllable transcriptions, so they have been neglected for everything reported so far in this thesis. Altogether, 493 sentences are left for analysis. Fig. 5.3 gives an overview on which sentence is available from which speaker in which emotion. As obvious from these charts, the emotions *disgusted* and *sad* are very scarce, and also *anxious* and *happy* just exceed 50% of the theoretically available number of sentences (which is the product of the number of sentence types and the number of speakers).

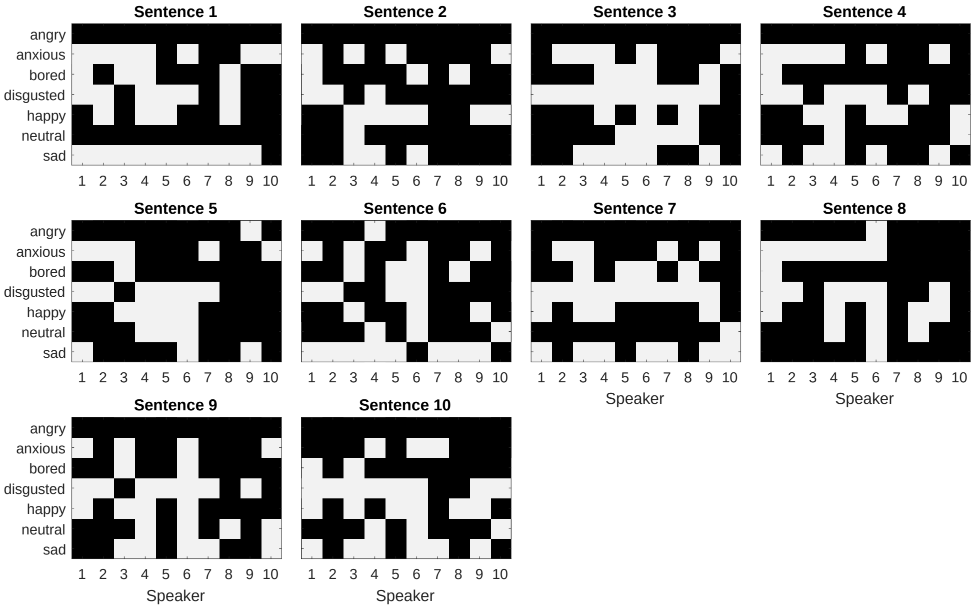
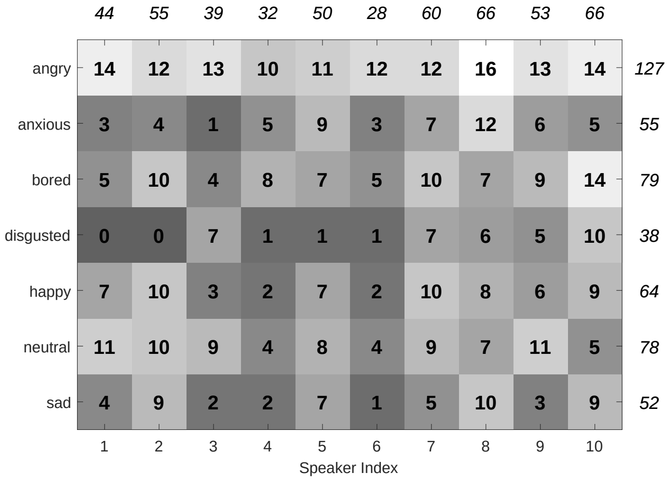


Figure 5.3.: Distribution of sentences over speakers and emotions. A black square indicates that this sentence is available by this speaker (x dimension) and in this emotion (y dimension).

The relevant statistics are shown in Fig. 5.4, which is the distribution of emotions over speakers. For two speakers, not one single *disgusted* sentence is available, such that specific classifiers for these speakers could only distinguish between 6 instead of 7 classes. In any case, it will not be possible to apply cross-validation in speaker-specific classification due to the small number of samples in single classes.



	1	2	3	4	5	6	7	8	9	10	
	44	*55*	*39*	*32*	*50*	*28*	*60*	*66*	*53*	*66*	
angry	14	12	13	10	11	12	12	16	13	14	*127*
anxious	3	4	1	5	9	3	7	12	6	5	*55*
bored	5	10	4	8	7	5	10	7	9	14	*79*
disgusted	0	0	7	1	1	1	7	6	5	10	*38*
happy	7	10	3	2	7	2	10	8	6	9	*64*
neutral	11	10	9	4	8	4	9	7	11	5	*78*
sad	4	9	2	2	7	1	5	10	3	9	*52*

Figure 5.4.: Distribution of emotions over speakers. The numbers indicate the overall amount of sentences a speaker has produced in that emotion, the colors correspond to the numbers (dark = low, bright = high) and help to find structure within the data. Italic numbers outside the box show the sums over all speakers and emotions, respectively.

5.2.2. Relevant Parameters

Parameter Scaling

Since one of the parameters turned our to be inappropriate for classification (section 5.1.4), 26 out of 27 prosodic and paralinguistic parameters are collected for 493 sentences in total. A check for not-a-number values (NaN) causes 9 sentences to be removed, as well as 2 sentences due to infinitely high values for at least one parameter. Finally, 482 sentences from *Emo-DB* are available for further analysis.

In addition to the absolute parameter values, relative parameters are calculated with respect to two different references:

a) A reference value for each parameter is calculated by taking the **average over all sentences** from that speaker, aiming at estimating his or her “normal” speaking style. These values will be referred to as “rel. avg.” in the result tables.
b) A reference value for each parameter is calculated by taking the **average over all *neutral* sentences** from that speaker, making use of the fact that *Emo-DB* provides a designated reference speaking style anyway. These values will be referred to as “rel. neutral” in the result tables.

In both cases, the absolute parameters are one-by-one divided by their respective reference values and subsequently checked for outliers which might occur due to, e.g., division by very small numbers. If outliers are detected, the corresponding sentences are not considered in the classification, such that the total number of sentences available is reduced. As the parameter *peak timing differences: dynamics* turns out to produce a vast amount of outliers through scaling, this parameter is rejected to prevent the deletion of the majority of sentences, such that only 25 parameters are left when performing classification with relative values.

- When taking **averaged** parameter values as the reference, 81 sentences have to be removed due to outliers, such that 401 sentences in 7 classes are available for classification.
- When taking ***neutral*** parameter values as the reference, the 77 *neutral* sentences[3] have to be ignored for obvious reasons. Only 1 single outlier is detected, and 404 sentences are available for classification.

Speaker-Specific Classification Results

To investigate how suitable our prosodic and paralinguistic parameters are for emotion recognition, SVM classifiers are trained based on the speech data of each speaker. Due to the limited number of samples in single classes, no cross-validation can be performed, such that the result will only represent the *in-sample classification accuracy* which does not allow to draw general conclusions from it. The classification accuracy is calculated by the *Correctly Classified Ratio* (CCR) which simply relates the number of correctly assigned samples to the total number of samples,

$$CCR = \frac{\text{\# correct assignments}}{\text{\# samples}} \,. \tag{5.2.1}$$

Both sequential forward and backward parameter selection are performed using the SVM classification error $(1 - CCR)$ of the current parameter set as the optimization criterion. If more than one set of parameters leads to the same minimum classification error, the smallest of these sets is chosen to be the best-performing parameter set.

The results for each of three parameter scaling methods, given in Tab. 5.2, show that perfect classification of emotions is possible for each speaker, using just a fraction of the full parameter set. Please note that the classification problem is reduced to 6 classes for speakers 1 and 2 due to missing *disgusted* sentences, and that the neglected *neutral* emotion for the "averaged w.r.t. *neutral*" parameters also reduces the number of classes by one, so that we end up with just 5 classes for speakers 1 and 2, and 6 for the others.

[3] One of the 78 *neutral* sentences from Fig. 5.4 has previously been removed due to NaN or Inf values.

Speaker	*absolute*		*rel. avg.*		*rel. neutral*	
	CCR	**nP**	**CCR**	**nP**	**CCR**	**nP**
1 (m)	100%	8	100%	6	100%	6
2 (f)	100%	8	100%	5	100%	4
3 (f)	100%	9	100%	7	100%	6
4 (m)	100%	4	100%	5	100%	3
5 (m)	100%	9	100%	7	100%	9
6 (m)	100%	5	100%	4	100%	5
7 (f)	100%	11	100%	9	100%	9
8 (f)	100%	11	100%	8	100%	11
9 (m)	100%	8	100%	7	100%	6
10 (f)	100%	8	100%	7	100%	7

Table 5.2.: Correctly classified ratios and corresponding numbers of parameters used for every single speaker and all three scaling types. *CCR* = correctly classified ratio, *nP* = number of parameters in best-performing set, *m* and *f* indicate the gender of the speakers.

Since all classifiers achieve the best performance value possible, a direct comparison seems to be meaningless; but what we can do is to compare the number of parameters contained in the best-performing set. Apparently, less parameters are necessary to achieve the same classification result when using relative values; at least for those calculated with respect to average values[4]. I would see this as a sign of quality, since a smaller parameter set suggests a higher potential for generalization.

To see which of our 26 parameters are contained in the best-performing sets, let's have a look at Fig. 5.5 which displays the CCRs for each speaker in the lower plot and the corresponding set of parameters above the bars, both for absolute parameter values. What attracts attention is the fact that there is little consistence between the parameter sets. One might detect similarities between speakers 2 and 5 as well as between 1 and 10, but in general, the composition of the parameter sets is very diverse. This observation is in accordance with the commonly stated fact that between-speaker differences in absolute speech parameter values often exceed between-emotion differences, which clearly limits the performance of speaker-independent classifiers using absolute values.

[4] The feature sets built from relative values w.r.t. the *neutral* emotion are also smaller (or equal) in size than their corresponding absolute-value parameter sets, but this might be caused by the reduced number of classes; I thus avoid a clear statement here.

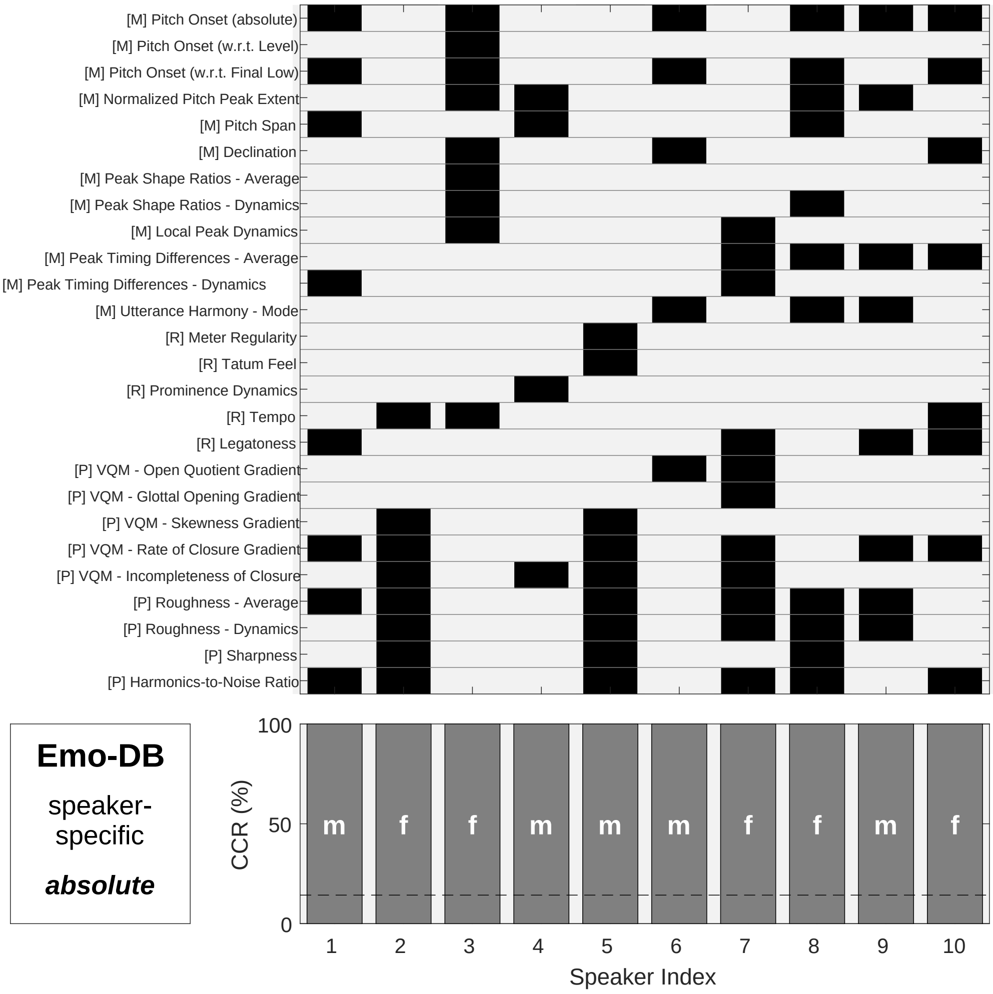


Figure 5.5.: Parameter selection for speaker-specific SVM classifiers using **absolute** parameters. Upper plot: black markers indicate that a parameter is part of the best-performing set for that speaker. Lower plot: all classifiers achieve a CCR of 100%; *m* and *f* denote the gender of the speaker.

Although there is no parameter which would have never been used in any of the best-performing parameter sets, the rhythmic parameters seem to be significantly less important for emotion recognition than the others — at least when using absolute parameter values. The emotions of some speakers can primarily be distinguished by voice quality parameters (e.g., speakers 2, 5,and 7), others mainly differ in melodic parameters (e.g., speaker 3), and still others make use of parameters from all. The most often selected parameters are *absolute pitch onset, rate-of-closure gradient, average roughness,* and *harmonics-to-noise ratio.* Interestingly, *pitch onset w.r.t. final low* is often selected together with the absolute pitch onset value, which suggests that these values are not redundant, but add some new information instead.

However, things are a little different when looking at the parameter sets for classification with relative parameter values (with respect to average values). As visible in Fig. 5.6, *average roughness* is still the most often selected parameter and also the pitch onset-related parameters seem play a major role again; but now, also rhythmic parameters including *tempo* and *legatoness* gain in importance.

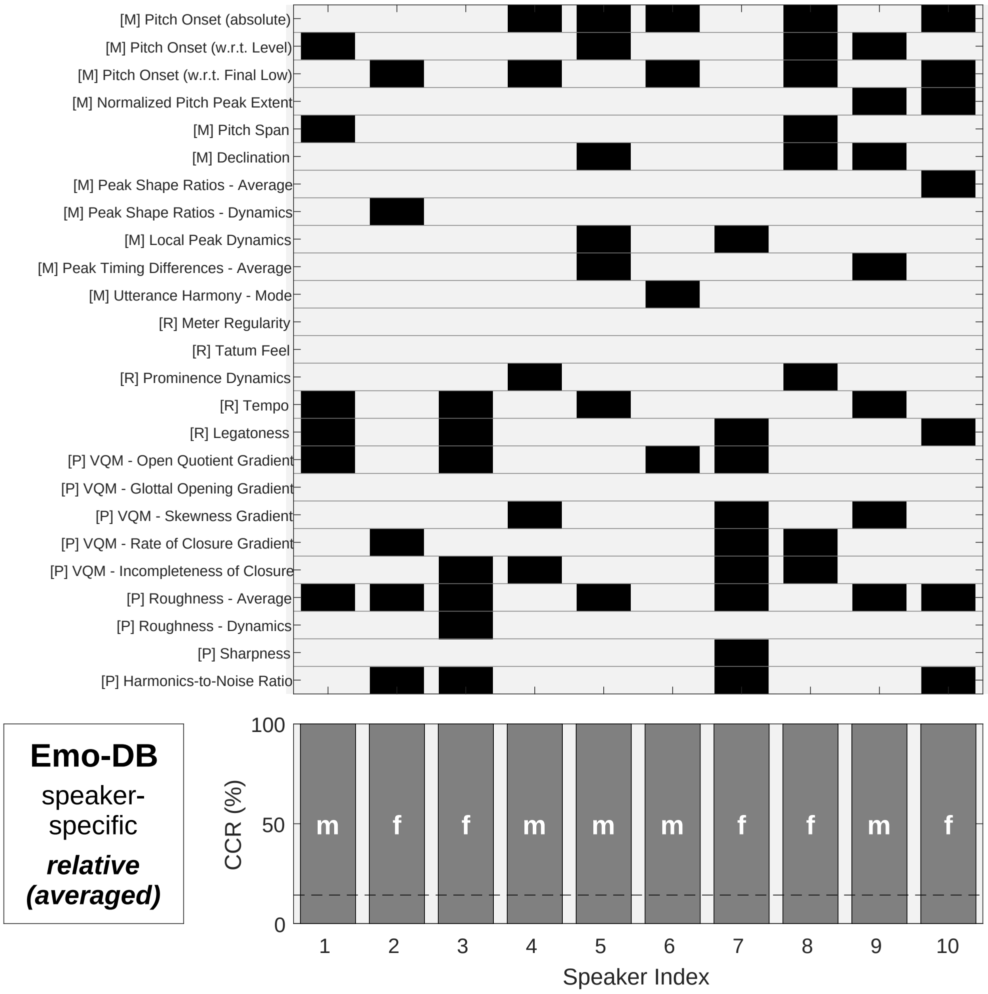


Figure 5.6.: Parameter selection for speaker-specific SVM classifiers using **relative** parameters (w.r.t. **average** values). Description: see above.

Again, a different picture emerges for the relative parameters with respect to the *neutral* emotion (Fig. 5.7): *pitch onset*-related parameters do not seem to be of great importance and *legatoness* remains the only relevant rhythmic parameter, whereas paralinguistic parameters including the *open-quotient gradient*, *sharpness*, and *harmonics-to-noise* ratio dominate the parameter sets.

It is questionable if these results give valid evidence on which parameters best

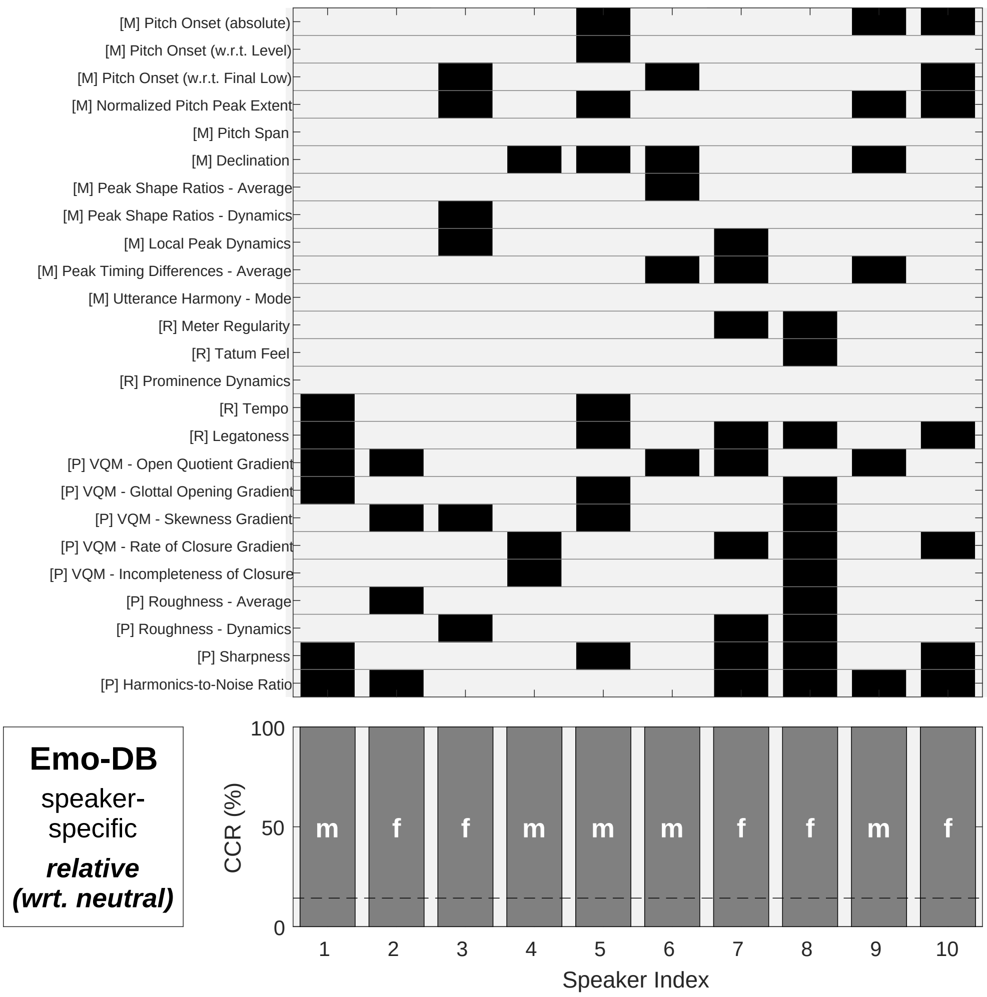


Figure 5.7.: Parameter selection for speaker-specific SVM classifiers using **relative** parameters (w.r.t. ***neutral*** **emotion**). Description: see above.

reflect cognitive stress, as there are significant differences between the two types of relative parameters which have been calculated using two different kinds of reference data, but with the same goal, namely to represent deviations from the "normal" speaking style. We might, however, name several parameters which are rarely present in all three collections of feature sets:

- Peak Shape Ratios
- Meter Regularity
- Tatum Feel
- Prominence Dynamics

The fact that three of these parameters are rhythmic descriptors might suggest that speech rhythm is not as much affected by emotions as speech melody or timbre, but one could as well question if the descriptors of speech rhythm proposed in this thesis capture the *rhythm* phenomenon in an appropriate way.

There is no evidence that gender plays a role in the parameter selection process at all.

Global Classification Results

To investigate if there is a global set of parameters which allows emotion classification with satisfying precision, the same approach as described above is followed again, this time considering all available sentences for one single SVM classifier. To prevent the parameter selection algorithms from getting caught in local minima[5], they are forced to go through the complete parameter set, even if the criterion gets worse again. Now that we have a sufficient number of samples in each class, we can perform cross-validation. The parameter k is set to 10, which is not only a commonly chosen value, but definitely makes sense regarding the sample sizes.

The results for the three parameter scaling types are given in Tab. 5.3. Shown are CCRs of the best-performing parameter sets for both parameter selection methods as well as the number of parameters in the corresponding sets (*nP*).

Scaling	nC	Dir.	CCR	nP	nP 95%
absolute	7	fw	64.11%	24	10
		bw	67.43%	22	13
rel. avg.	7	fw	70.32%	16	12
		bw	69.58%	15	8
rel. neutral	6	fw	69.31%	19	9
		bw	68.56%	12	7

Table 5.3.: Correctly classified ratios and corresponding number of parameters using a global classifier for all three scaling types. *nC* = number of classes, *fw* = forward selection, *bw* = backward selection, *nP* = number of parameters in best-performing set, *nP 95%* = number of parameters in reduced set which still yields 95% of the best set's CCR.

[5] Parameter selection algorithms commonly stop adding or removing parameters if the criterion does not further improve — they have found a minimum in the objective function. It might, however, be the case that there is a global, "better" minimum a few steps ahead which is never found if the algorithms stop at the current point.

To evaluate how meaningful the last added parameters really are, the selection history is traced back to the point where the CCR just exceeds 95% of its highest value. As an example, the best set for absolute parameters found with forward selection includes 24 parameters and yields a CCR of 64.11%. With just 10 parameters, we still get 61.20% CCR, which is 95.47% of the best value. In other words, one can estimate how many parameters contribute to the first 95% of the information (*nP 95%*), and how many to the last 5%.

When looking at the percentages of correctly classified samples, we should keep in mind that chance level is only 14.3% for 7 classes and 16.7% for 6 classes. Apparently, the best result has been found using forward selection of relative parameters w.r.t. average values, which is nearly 5 times better than chance. Fig. 5.8 illustrates the sequential parameter selection process for this best result.

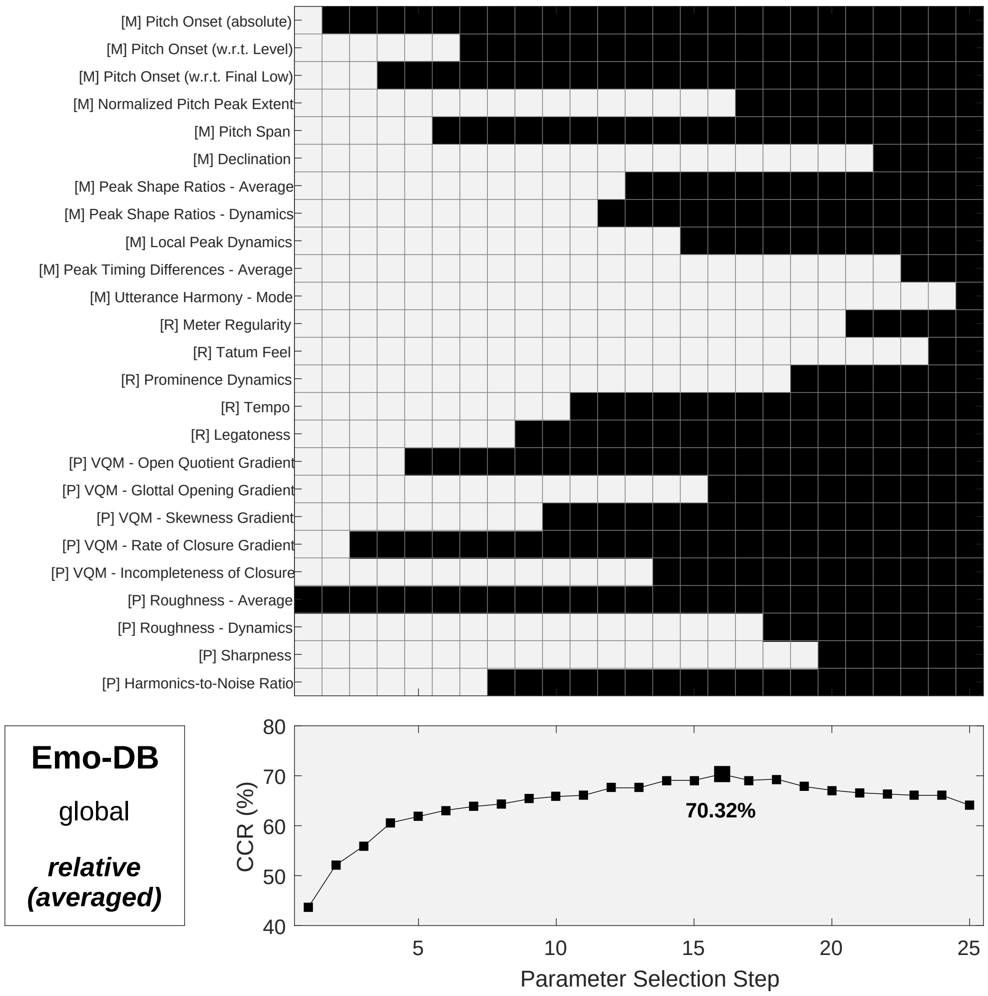


Figure 5.8.: Sequential forward parameter selection for the global SVM classifier using **relative** parameters (w.r.t. **average** values). Description: see above.

The lower plot in Fig. 5.8 shows the CCR for each parameter selection step, while the evolution of the parameter set is illustrated in the upper plot, where black squares mark those parameters which are in the set at the current step.

When comparing this result to those for the other scaling methods and parameter selection directions (which have been moved into appendix A.2 for better clarity[6]), it is apparent that these different approaches deliver divergent results. To provide an overview, I have analyzed which parameters are selected how often, and how often they are part of a best-performing parameter set. In Fig. 5.9, we see the *overall selection frequency* on the left side, that is, how many times a parameter has been part of the set over all selection steps, with the theoretical maximum being $(6 \cdot 25 =)$ 150 times. On the right side the *occurrence in best sets* is displayed, where the maximum is obviously 6.

The following parameters are always part of the **best-performing parameter set**, regardless of scaling type or selection direction:

- **Pitch Onset** (both absolute and with respect to final low)
- **Local Peak Dynamics**
- **Speech Tempo**
- **4** out of 5 **Voice Quality Measures** (except for GOG)
- **Average Roughness**
- **Harmonics-to-Noise Ratio**

The overall impression from the parameter selection results for a global *Emo-DB* classifier is that *Pitch Span, Utterance Harmony, Meter Regularity, Tatum Feel,* and *Sharpness* have only little descriptive power for the emotional state of the speakers. The fact that some of these parameters are part of the best-performing set in 50% of the cases is not of great relevance, since these sets are relatively large (see Tab. 5.3), and the increments in CCR introduced by adding these parameters are negligible.

[6] The figures start from page 160.

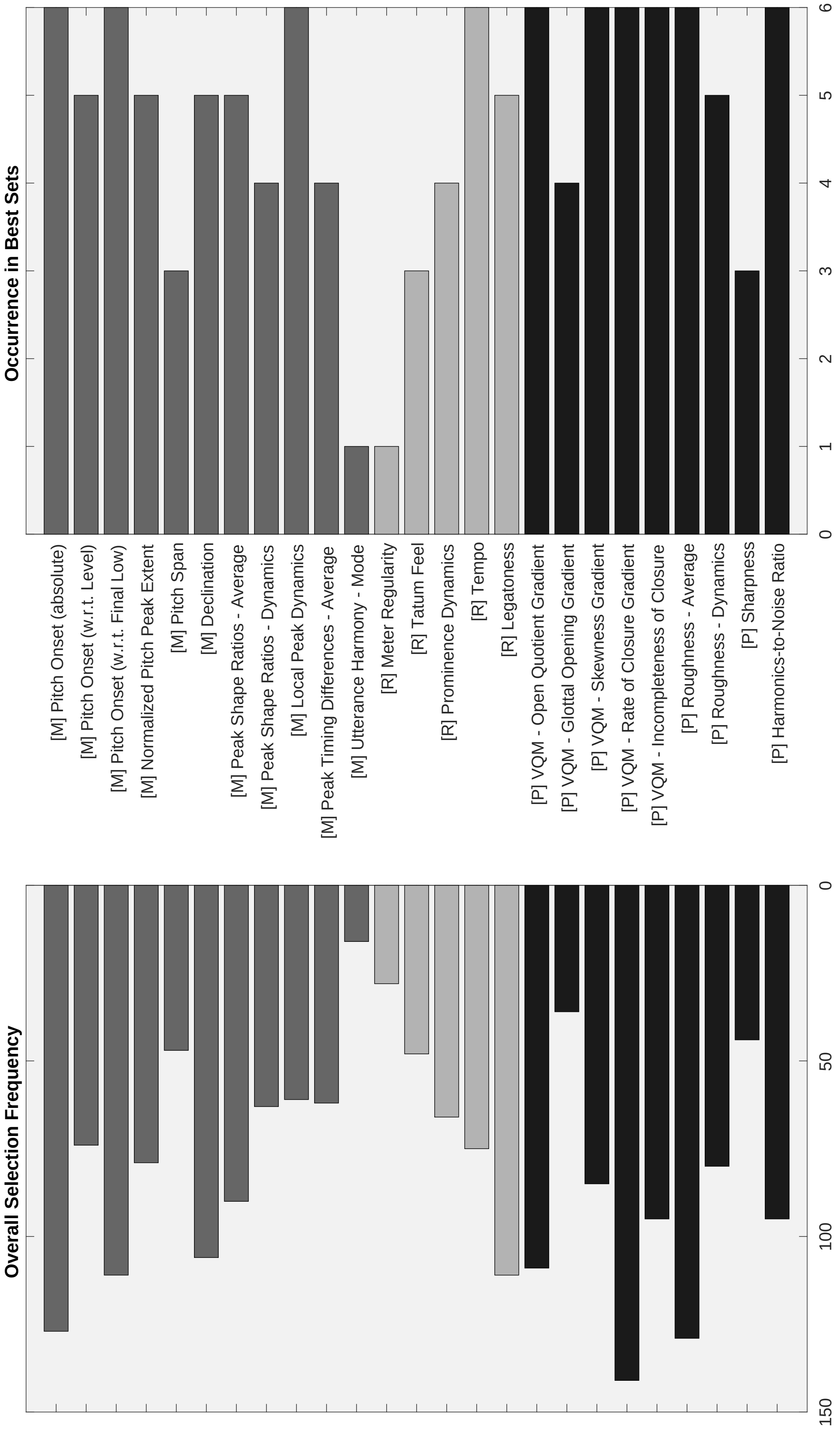


Figure 5.9.: Overview on parameter selection for the global SVM classifier.

5.3. The Prosody of Speech Under Stress

5.3.1. Investigated Speech Data

The *IEM Pilot Speech Database* (*IEM-PSD*) has extensively been introduced in chapter 2. According to the flight program, there are 18 potential stress-invoking events in total. As a “ground truth” for the actual stress the participants were exposed to, their heart rates have been measured over the whole experiment, such that we are able to calculate heart rate variability (HRV) parameters for any time period we like.

As a reminder, Tab. 5.4 lists all events from the flight plan which shall be analyzed in this section

Event ID	**Description**
0a	Reference flight: Takeoff
0b	Reference flight: Landing
1a	Flight 1: APU fails
1b	Flight 1: Runway change during taxi
1c	Flight 1: Engine #2 hot start trial
1d	Flight 1: Takeoff
1e	Flight 1: TCAS alert
1f	Flight 1: Engine #1 flame out
1g	Flight 1: Landing / Engine #2 fire
2a	Flight 2: Takeoff
2b	Flight 2: Engine #1 seizure
2c	Flight 2: Generator #2 failure
2d	Flight 2: Landing / Gear collapse
3a	Flight 3: Engine #1 flame out / Takeoff abortion
3b	Flight 3: Engine #1 fire
4a	Flight 4: Takeoff
4b	Flight 4: Engine flame out
4c	Flight 4: Landing

Table 5.4.: List of events from *IEM-PSD*.

HRV Parameters as a Reference for Stress Level

Unfortunately, biological signals have a substantially lower temporal resolution in comparison with acoustical signals. The pulse rate in rest is typically [50..100] beats per minute for adults, so one can imagine that it takes some time before measures of frequency or variability can be calculated. Instantaneous heart rate as a descriptor for stress level would be just as valuable as a few individual pitch values would be (namely, not at all). The common approach for calculating HRV parameters in cardiology is to use overlapping analysis windows of 5 minutes, resulting in a time series of parameter values for observation periods of several hours (Malik et al., 1996).

This has two consequences:

- We will not be able to assign a “reference stress level” to every utterance; conversely, we will have to assign several utterances to a stress level.
- The immediate reaction to a stressful event — which is indeed visible in the instantaneous heart rate — is typically decayed within 30 seconds. If we captured this reaction using a 5-minute window, we would not be able to distinguish between the three states *before*, *during*, and *after* the event due to temporal smearing of the parameter values. We thus have to perform **event-based analysis** with analysis windows as short as possible.

From a theoretical point of view, the minimum window length is determined by the lowest frequency band considered in the analysis. The LF band goes down to 0.04Hz (→ 2.3.2), so the minimum window length is 25 seconds; to be on the safe side, I have chosen 30 seconds. This enables us to capture the immediate stress reaction provoked by the event separate from the physiological steady-state phases before and after the event, as sketched in Fig. 5.10.

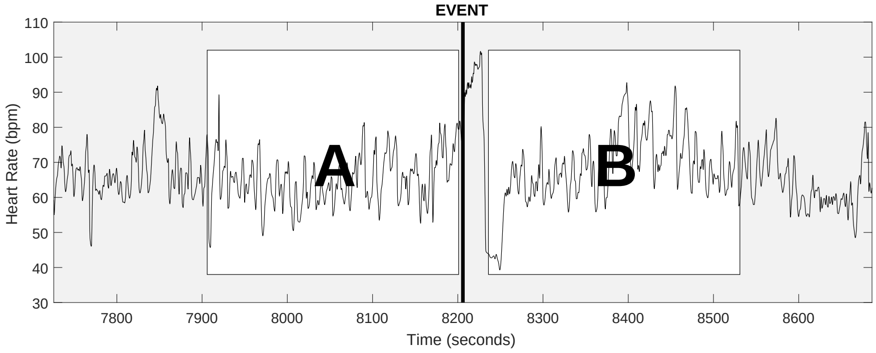


Figure 5.10.: Event-based analysis: HRV parameters are calculated in windows of 30 seconds within 5 minutes before the event (A) and 5 minute after the event (B), leaving a 30-second “cool-down phase” immediately after the event.

For both of these regions before and after an event, a set of HRV parameters is calculated (see section 2.3.2). To investigate if the individual events have introduced stress on the pilots, the *in-sample classification accuracy*[7] of these two HRV parameter sets is determined for each speaker and each event individually. If classification accuracy is low, it can be concluded that the related event has had no significant impact on that pilot. Fig. 5.11 displays HRV classification accuracies for each speaker and each event as vertical bars, and the average accuracies over all speakers are given numerical across each group of bars[8].

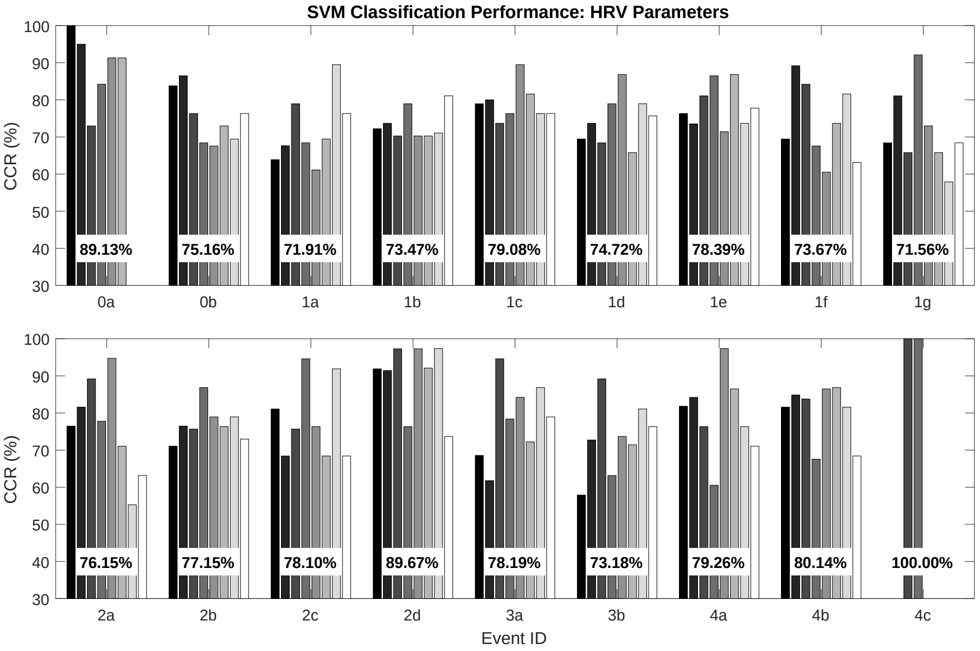


Figure 5.11.: *IEM-PSD* classification performance of "before" vs. "after" using HRV parameters for all subjects and events. Different shades of gray indicate different speakers; the numbers are the average classification performance over all subjects.

Selected Events from *IEM-PSD*

Since both quality and extent of a stress reaction are highly individual, the stress reaction of each speaker must be evaluated individually. Classification algorithms, however, need a sufficient number of data values in order to gain some explanatory power. As a consequence, we have to specify a threshold

[7] This means that the training data and the test data are identical, and we want to find out how well the data can be separated in the first place.

[8] Please note that the HRV data "before event 0a" in session 4 and "after event 4c" in sessions 1/3/4 turned out to be invalid, so they had to be neglected.

value for the required number of utterances before as well as after the event. The higher we set this value, the stronger the classification result, but also so much lower the number of events which fulfill this condition.

The number of available events as a function of the minimum number of utterances is depicted in Fig. 5.12. It seems reasonable to set this threshold value to 10, ending up with 28 events in total. These 28 events in total mean 9 unique events, unfortunately only from the first two flights and the reference flight (Events 0a, 1a, 1b, 1c, 1d, 1e, 2a, 2b, 2c; see Fig. 2.2 on page 35.)

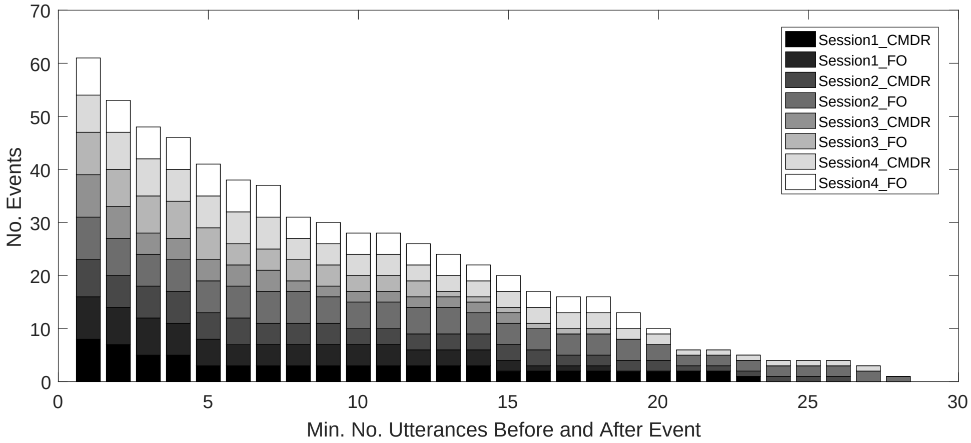


Figure 5.12.: Number of available events, broken down by different speakers, as a function of the minimum number of utterances both before and after the event.

5.3.2. Relevant Parameters

For the classification of speech under stress, I will follow a similar approach as for emotional speech (→ 5.2.2). In this case, there is no effective “neutral” state available, so there will be only one type of relative values. At the same time, classifiers can be created not only globally or speaker-specifically, but also *event-specifically*.

Parameter Scaling

Just as for *Emo-DB*, 26 prosodic and paralinguistic parameters are investigated as potential descriptors of stress level. Altogether, 1063 utterances are available for further analysis, after having removed 34 utterances due to erroneous parameter values.

In addition to the absolute parameter values, relative parameters are calculated with respect to the average over all sentences from that speaker. This is done by dividing the absolute parameters one-by-one divided by their respective reference values and subsequently checking the resulting values for outliers which might have occurred due to, e.g., division by very small numbers. If outliers are detected, the corresponding sentences are not considered in the classification, such that the total number of sentences available is reduced. At the end, 173 utterances had to be removed due to outliers, such that 890 of them are still available for classification.

Speaker-Specific Classification Results

Using the same methodology as for *Emo-DB*, SVM classifiers are trained based on the speech data of each speaker. Due to the limited number of samples in single classes, no cross-validation can be performed, such that the result will only represent the in-sample classification accuracy. Both sequential forward and backward parameter selection are performed using the SVM classification error $(1 - CCR)$ of the current parameter set as the optimization criterion. If more than one set of parameters leads to the same minimum classification error, the smallest of these sets is chosen to be the best-performing parameter set.

The results are given in Tab. 5.5. We see that, for 3 out of 8 speakers, there is no single event which fulfills both quality criteria (required stress impact and sufficient number of speech files) at the same time. In contrast to the *Emo-DB* results, the SVM is not able to classify the speech samples perfectly in most of the cases. However, the percentage of correctly classified samples is not only significantly about chance level, but generally on par with the reference CCRs from the heart rate variability parameters. Relative parameters are, on the whole, superior to absolute parameters, due to the fact that they achieve better classification rates with fewer parameters (with one exception for each of these two statements). This trend coincides with what we observed for the speaker-specific classifiers built on *Emo-DB* speech data[9].

Please note that these speaker-specific classifiers implicitly are also event-specific classifiers, since there is not more than 1 event per speaker at which the HRV parameters were able to classify “before” vs. “after” better than 80% CCR, while at the same time at least 10 utterances were available from both classes.

Let’s also have a look on the parameters which form the best-performing sets, displayed in Fig. 5.13. We can again observe considerable individual differences between speakers which make use of a wide variety of parameters, re-

[9] The classifiers achieved 100% CCR in all cases, but the number of parameters needed to achieve this result was clearly lower for relative parameters (→ 5.2.2).

Speaker	Events	*absolute*		*relative*		CCR HRV
		CCR	nP	CCR	nP	
S1: CMDR	2c	84.21%	17	91.67%	14	81.08%
S1: F/O	—	—	—	—	—	—
S2: CMDR	2a	83.78%	12	83.87%	15	89.19%
S2: F/O	2c	86.36%	20	84.09%	13	94.59%
S3: CMDR	—	—	—	—	—	—
S3: F/O	—	—	—	—	—	—
S4: CMDR	1a	100.00%	19	100.00%	15	89.47%
S4: F/O	1b	91.67%	12	92.59%	9	81.08%

Table 5.5.: Correctly classified ratios and corresponding numbers of parameters used for every single speaker and both scaling types. *CCR* = correctly classified ratio, *nP* = number of parameters in best-performing set.

spectively. However, some commonalities are also apparent: melodic parameters are not strongly represented, with the exception of *average peak timing* (in the absolute parameters) and the *dynamics of peak shape ratios*. Apart from the *glottal opening gradient*, voice quality measures do not dominate the parameter sets as in the *Emo-DB* results, whereas some rhythmic parameters perform surprisingly well, including *prominence dynamics*, *tempo*, and *legatoness*.

It is notable that the numbers of parameters needed to achieve the best-possible result is significantly higher for speech data from *IEM-PSD* than for *Emo-DB* data. On average, absolute parameter sets consist of 16 parameters (*Emo-DB*: 8), and relative parameter sets include about 13 parameters (*Emo-DB*: 6-7). One might question if this is due to the different nature of speech under stress compared to emotional speech, or if the reason for this difference might rather be found in "acted vs. natural speech".

Event-Specific Classification Results

Since the different events in the flight plan are assumed to put different kinds of demand on the participants, their impact on the pilots' speech is expected to vary as well. To investigate this, *event-specific global classifiers* are created for each of the selected events. The classification results for each of these classifiers are given in Tab. 5.6; the number of speakers contained in the training

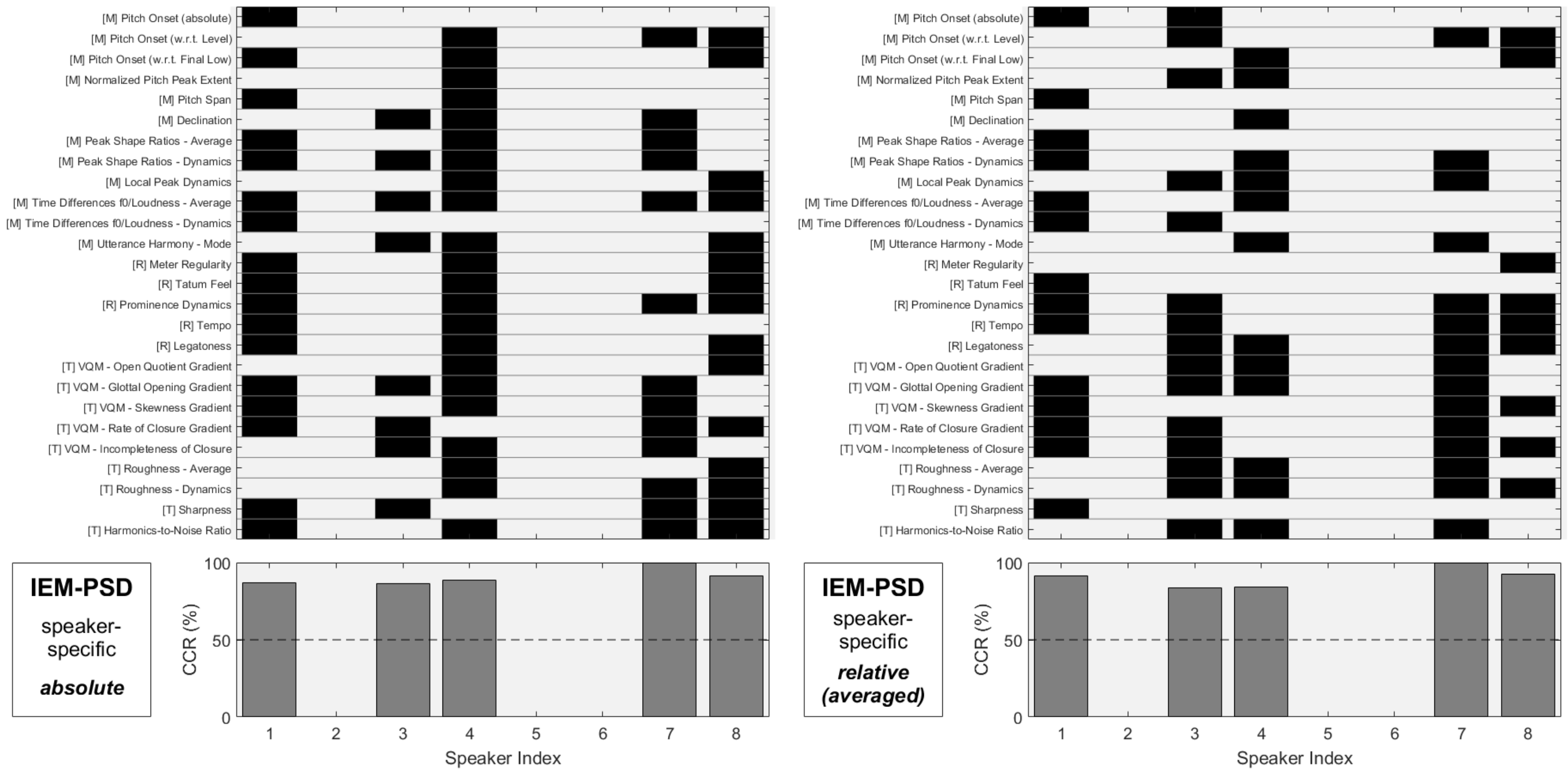


Figure 5.13.: Parameter selection for **speaker-specific** SVM classifiers using both **absolute** (left) and **relative** (right) parameters. Upper plots: black markers indicate that a parameter is part of the best-performing set for that speaker. Lower plots: correctly classified ratios (CCR) achieved with the bets-performing set. The dashed line represents chance level.

data for the classifier depends on the amount of utterances produced before and after the event[10].

Event	Speakers	*absolute* CCR	*absolute* nP	*relative* CCR	*relative* nP	∅ CCR HRV
0a	3	66.67%	18	74.42%	17	72.97%
1a	1, 2, 4, 7, 8	66.33%	16	72.08%	6	73.15%
1b	5, 6, 7, 8	66.89%	13	69.23%	11	73.17%
1c	7	74.42%	7	72.50%	9	76.32%
1d	1, 2, 6, 8	62.04%	14	80.23%	8	71.15%
1e	7, 8	81.82%	22	78.26%	16	75.73%
2a	1, 3	79.75%	16	76.71%	16	82.83%
2b	1, 2	74.29%	16	79.37%	13	73.76%
2c	1, 4, 8	75.00%	12	76.85%	13	81.37%

Table 5.6.: Correctly classified ratios and corresponding numbers of parameters used for every single event and both scaling types. *CCR* = correctly classified ratio, *nP* = number of parameters in best-performing set.

For these results, a direct comparison of CCRs for speech parameters and HRV parameters is not reasonable, because the average classification accuracy of the HRV parameters has been determined for each participant individually and the ∅ *CCR HRV* value given in Tab. 5.6 is just the average accuracy for all subjects involved. The correctly classified ratios for the speech parameters, in contrast, are *not* speaker-specific, but result from global classifiers trained with data from that specific event. We can again state that the number of necessary parameters for best-possible classification is generally lower when using relative parameters. On average, the classification accuracy is also better (CCR = 75.52%, opposed to 71.91% for absolute parameters). As obvious from Fig. 5.14, there is no obvious tendency towards one or the other category — melodic, rhythmic, paralinguistic — of parameters, due to the relatively big size of the best-performing parameter sets. Single parameters, however, including the *dynamics of peak timing* and *meter regularity*, are rather seldom to appear in one of the "best sets". Three events attract attention because of their relatively small sets of best-performing parameters with regard to relative parameters, (that is, events *1a*, *1c*, and *1d*), but even in these cases, no clear trend towards one or the other category is visible.

[10] The second condition from the speaker-specific classifiers, that the HRV parameters must at least show a classification accuracy of 80%, has been dropped due to the fact that Tab. 5.6 would otherwise be identical with Tab. 5.5, since there has been only one event per

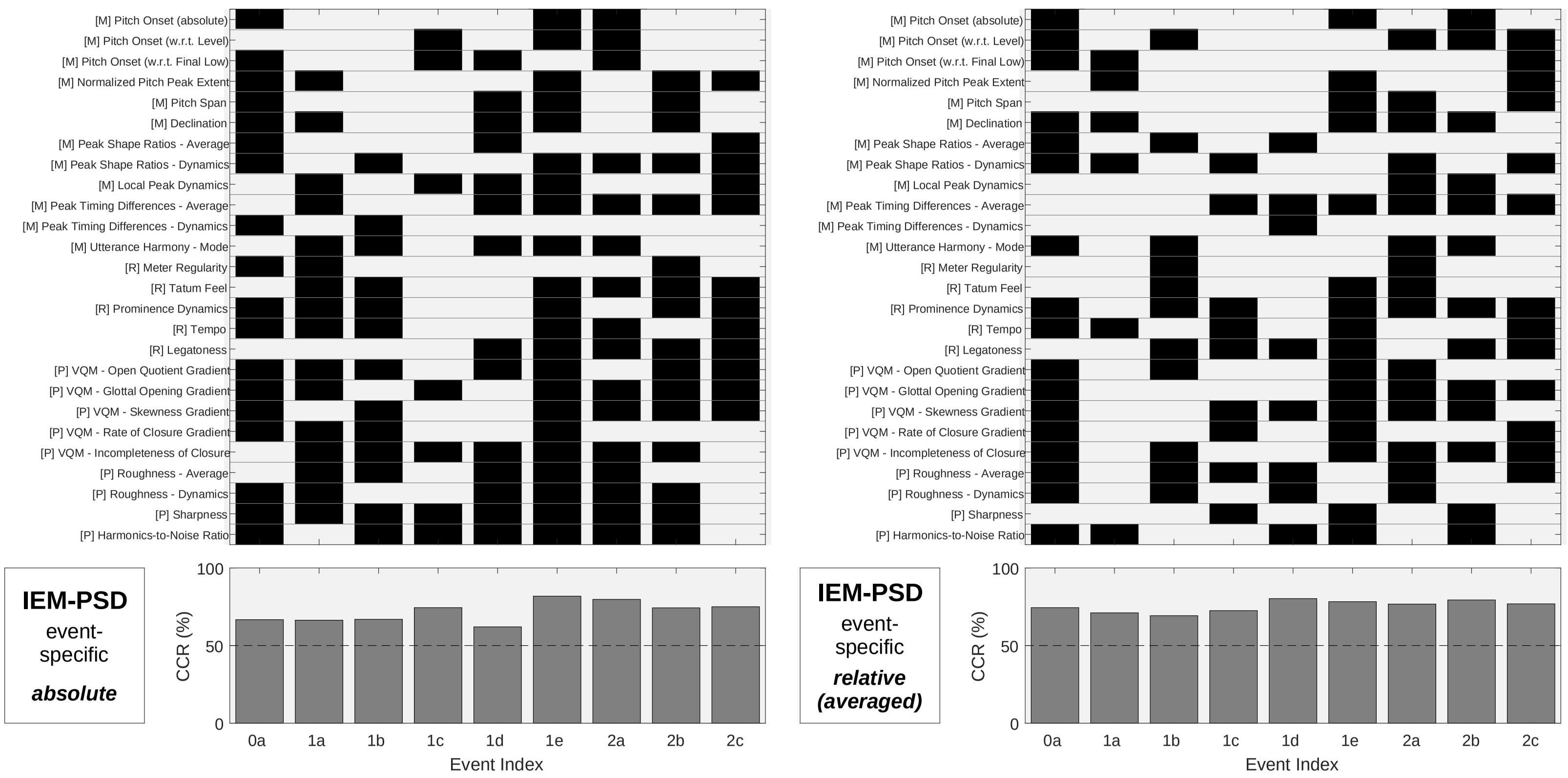


Figure 5.14.: Parameter selection for **event-specific** SVM classifiers using both **absolute** (left) and **relative** (right) parameters. Upper plots: black markers indicate that a parameter is part of the best-performing set for that event Lower plots: correctly classified ratios (CCR) achieved with the bets-performing set. The dashed line represents chance level.

Global Classification Results

Finally, let's see how well a global classifier would perform on the pilot speech database. We should not raise great expectations on its classification accuracy due to the frequently mentioned inter-personal differences in stress responses as well as general speaking style, but it is nevertheless interesting how far we can get with this.

For the sake of completeness, I have also trained SVM classifiers on all 9 selected events, independent of the ground truth given by the HRV parameters. For some of these events, we have only little evidence for actual stress impact. This means that we are looking for changes in speech parameters which may have not changed significantly, because there was no driving force which coud have changed the cognitive state of the speaker; in other words, we introduce a lot of noise. As visible from Tab. 5.7 ("selected" category), the classification accuracy just exceeds chance level.

Events	Scaling	Dir.	CCR	nP	nP 95%
selected	absolute	fw	53.12%	15	1
		bw	53.93%	6	1
	relative	fw	57.01%	15	1
		bw	56.20%	4	2
qualified	absolute	fw	60.73%	16	12
		bw	65.97%	10	7
	relative	fw	64.91%	16	5
		bw	70.76%	10	4

Table 5.7.: Correctly classified ratios and corresponding number of parameters.

But things change when we only consider "qualified" events which fulfill the condition of showing a classification accuracy of at least 80% in the HRV parameters. The best result is achieved with a set of 10 parameters found using sequential backward selection of relative parameters. When tracing the history of this selection procedure (as shown in Fig. 5.15) and read it from right to left, we will discover that three paralinguistic parameters (**harmonics-to-noise ratio** and **2 voice quality measures**) form the basis of the parameter set, supplemented by one rhythmic (**legatoness**) and two melodic (**local peak dynamics** and **dynamics in peak shape ratios**) parameters. This set of 6 relative parameters already facilitates classification of qualified IEM-PSD events with an accuracy of 69.01%.

speaker which fulfilled both conditions.

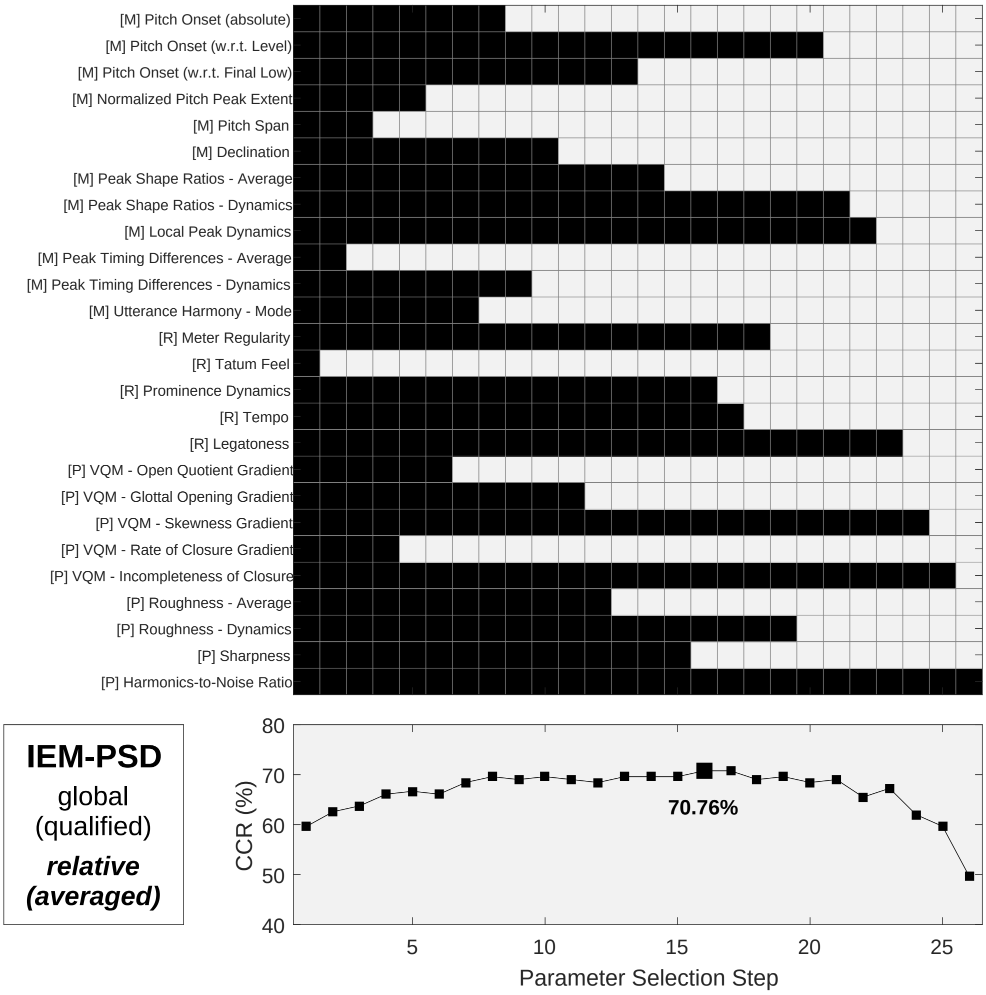


Figure 5.15.: Sequential **backward** parameter selection for all **qualified** events from *IEM-PSD* using **relative** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

5.4. Discussion

When comparing the classification results for *Emo-DB* and *IEM-PSD*, we should keep in mind that we do not just have “emotions” on the one side and “stress” on the other, but also (or should I say *rather*?) prompted sentences in acted emotional states here and free speech in situations of actual stress there.

The **classification results** presented in this chapter should be seen as indications of descriptive power of single speech parameters, but not as reliable statements regarding the absolute numbers. I have put little efforts in tuning the SVM parameters, and I have not considered alternative classification methods, since the scope of my research is something different. I think that Batliner’s statement on the sense in comparing classification rates between different studies (see the quote on page 120) is true. Nevertheless, it is worth having a look at the classification accuracies in terms of CCRs for both databases.

Speaker-specific classifiers for the *Emo-DB* data manage to assign the correct emotion label for 100% of the sentences with 6.5 relative parameters on average (see Tab. 5.2 on page 130). Keeping in ind that chance level is just about 14% for 7 categories, this is an impressive statement and confirms that the prosodic and paralinguistic parameters presented in this thesis indeed describe *the way we speak*, even if the acted emotions might have been performed in a very “intense” way — what fosters their differentiation — and considering that in-sample classification generally leads to highly specific results.

In contrast, the speaker-specific classifiers (which were implicitly also event-specific) for the *IEM-PSD* data only show an average classification accuracy of 90.44% and need twice as much speech parameters to do so (see Tab. 5.5 on page 143) — and in this case, chance level is at 50% for a binary decision. This is not just as impressive in absolute numbers, but one can imagine that this classification task is by far more difficult. To my knowledge, there are no other studies on speech under stress which actually apply a classification algorithm to differentiate between “stress” and “non-stress” conditions; statistical measures as p-values of ANOVA analyses are, in my opinion, difficult to grasp. In any case, the investigated prosodic and paralinguistic parameters have certainly proven their general applicability for stress recognition, since they do not only reach, but even outperform the established HRV parameters in classification accuracy (90.44% vs. 87.08% CCR for speaker-specific classifiers).

Interestingly, the best *global* classifier for speech under stress (70.76% using 10 parameters on qualified events, see Tab. 5.7 on page 147) performs on par with the best global classifier fo emotional speech (70.32% using 16 parameters, see Tab. 5.3 on page 134).

When we compare the **best-performing parameter sets** for both *Emo-DB* and *IEM-PSD* data, different pictures emerge for speaker-specific and for global classifiers. While the "best sets" for speaker-specific classification in the emotional speech domain are just half as big as their speech-under-stress counterparts, the opposite is true for the parameter sets for global classification. A possible conclusion would be that emotional expression is, in general, more speaker-specific than stress responses are — at least regarding the verbal reaction.

While the speaker-specific parameter sets for *Emo-DB* data show clear focuses towards one or the other category of speech features (melodic, rhythmic, or paralinguistic), the speaker-specific parameter sets for *IEM-PSD* are very diverse, probably due to the big size of most of the "best sets". However, the global sets for both databases have a union set of 5 parameters, which is, after all, half of the best-performing set for the speech-under-stress data.

One unambiguous result of these investigations is that a smaller set of relative parameters yields higher classification accuracies than a larger set of absolute parameters. This was expected in advance, because it is an open secret in language research that individual differences in speaking style between speakers usually exceed global differences in speaking style under different conditions.

6. Summary

In this thesis, I have presented a concept for the calculation of 27 prosodic and paralinguistic parameters to describe *the way we speak* in an objective way. A consistent methodology has been implemented which builds on a theoretical model of "prosody" and extends it by the use of auditive variables which take human perception of pitch, loudness, and timing into account. A central aim of my work was to bring together diverse approaches and insights from different fields of research, including linguistics, speech-language pathology, digital signal processing, music information retrieval, and machine learning; I hope to have met my own claims by considering all kinds of different approaches and by providing clear descriptions and explanations throughout this thesis.

On my journey from the raw signal as captured by a microphone to the several high-level parameters which describe certain qualities of melody, rhythm, or timbre, some innovations had to be made in order to bridge one or the other divide. A self-tuning algorithm for the blind detection of syllable boundaries has been developed which automatically detects the (most probable) number of syllables in an utterance and which furthermore considers the perceived syllable length as a relevant parameter for duration perception. Another novel contribution is made by the presented technique to derive a continuous pitch contour from a series of fundamental frequency values and by a rule-based method to determine the pitch of a syllable. Finally, the musical interpretation of *speech rhythm* prompted me to force the perceptual centers of the syllables into a regular rhythmic grid, which is also an original contribution. Some of the presented parameters are already known, others are commonly evaluated manually, but not automatically, and again others are my own invention.

These prosodic and paralinguistic parameters have further been investigated regarding their ability to differentiate between different emotional states or different levels of cognitive stress. The latter has been made possible by the creation of a speech database with speech under stress in the cockpit, which features professional airline pilots in a class D full flight simulator. This database is publicly available free of charge for academic research purposes and will hopefully add value to the scientific community.

The classification results show that speech parameters are indeed suitable indicators of both emotional state and stress level. We experience immanent limitations due to individual differences in speaking style as well as strong parameters which appear in all best-performing parameter sets. As presumed, relative changes in speech parameters with respect to a speaker's "normal" speaking style are more meaningful than absolute values. A remarkable fact is that, for the classification of speech under actual stress, speaker-specific classifiers outperform the well-established parameters of heart rate variability.

> *"Es war sehr schön, es hat mich sehr gefreut."*
>
> — Franz Joseph I of Austria, 1830-1916

A. Supplemental Figures and Tables

A.1. Prominence Regression

Distribution of Labelled Prominence Values in *Emo-DB*

The following figures show the distribution of labelled prominence values per syllable for all 10 different sentences from the *Emo-DB* database, over all speakers and emotions, as Box-Whisker plots.

The bold line shows the median (which is the 2nd quartile or the 50% percentile), the box represents the range between the 1st (25%) and the 3rd (75%) quartile, and the whiskers indicate the complete range of the data without outliers.

The ordinate shows values in *Emo-DB* notation from 0 to 3, which correspond to the prominence levels [0..30] in the linguistic standard notation. The dotted line at 12.5 marks the soft threshold for the fuzzy regression analysis.

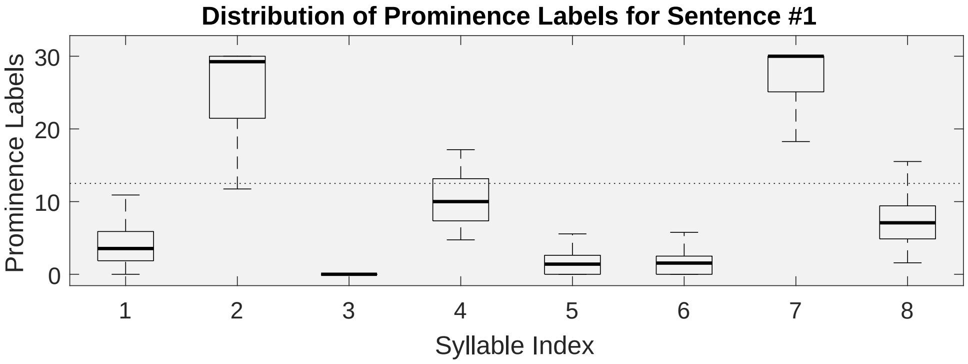


Figure A.1.: Labelled prominence values from *Emo-DB* for sentence 1.'

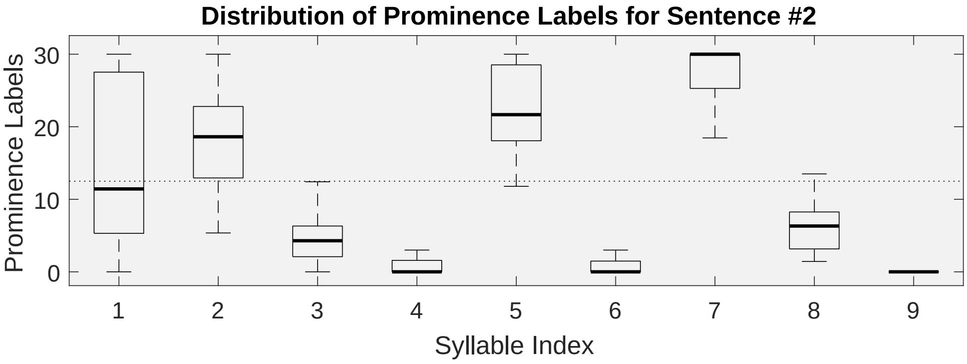


Figure A.2.: Labelled prominence values from *Emo-DB* for sentence 2.

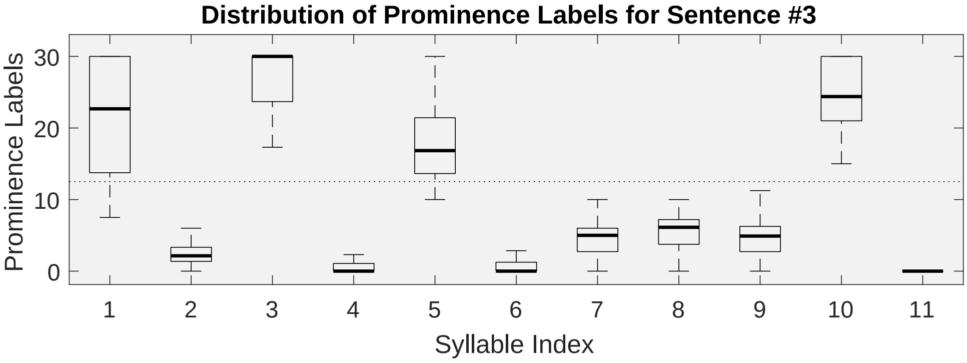


Figure A.3.: Labelled prominence values from *Emo-DB* for sentence 3.

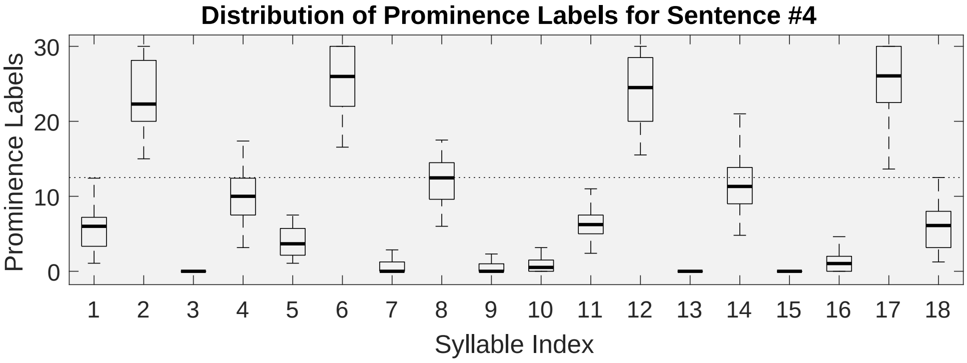


Figure A.4.: Labelled prominence values from *Emo-DB* for sentence 4.

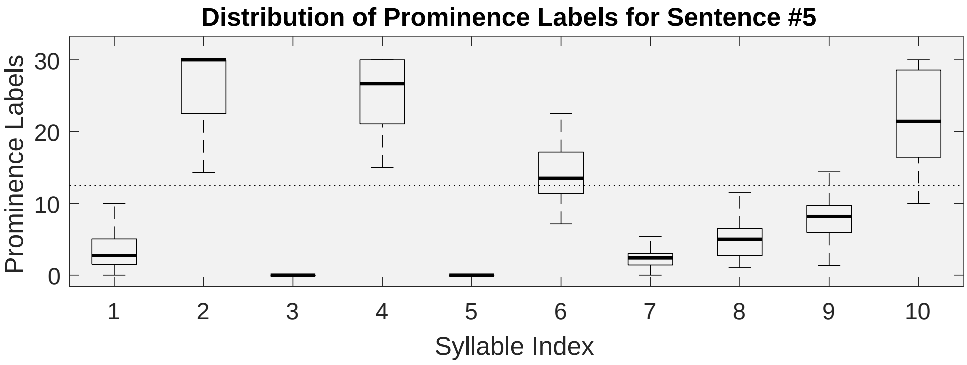


Figure A.5.: Labelled prominence values from *Emo-DB* for sentence 5.

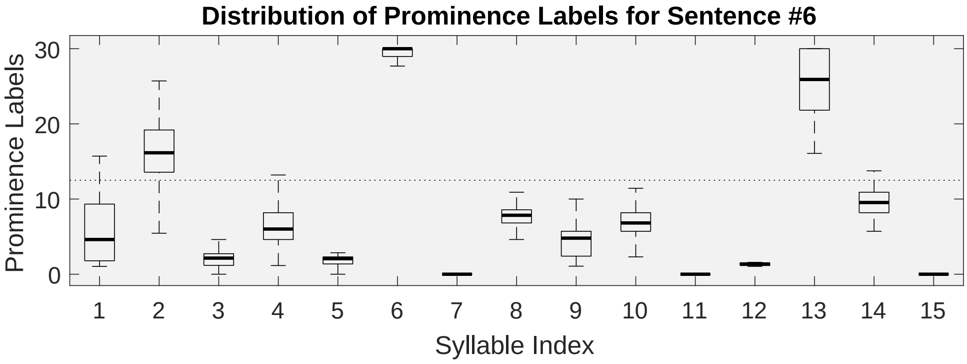


Figure A.6.: Labelled prominence values from *Emo-DB* for sentence 6.

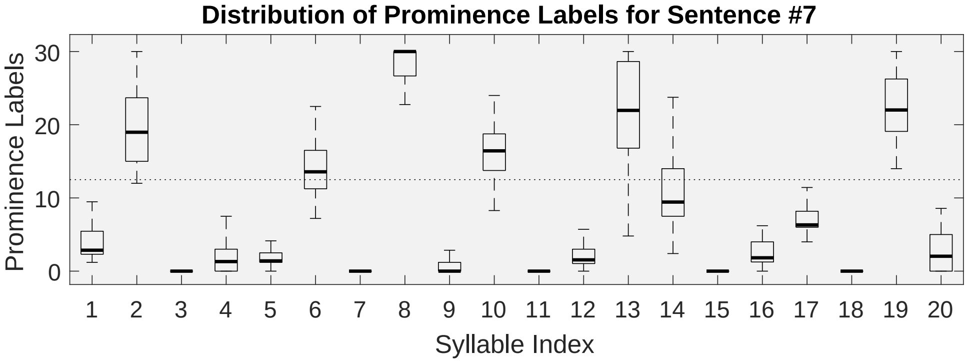


Figure A.7.: Labelled prominence values from *Emo-DB* for sentence 7.

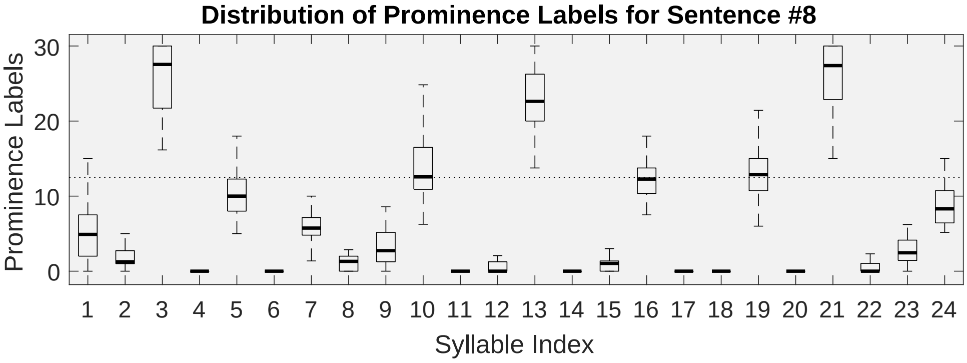


Figure A.8.: Labelled prominence values from *Emo-DB* for sentence 8.

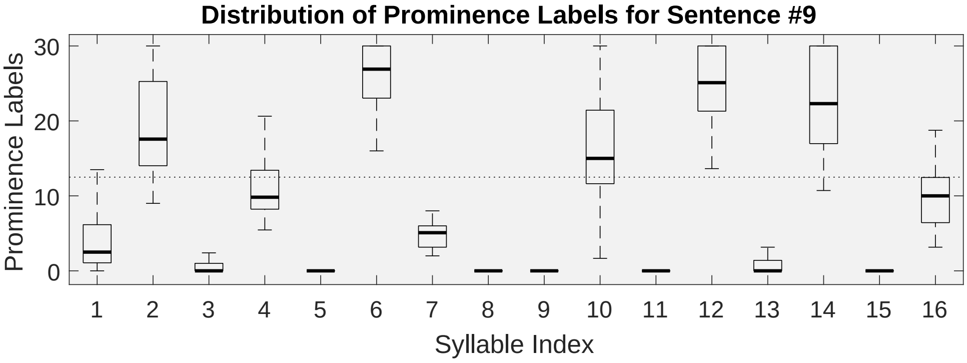


Figure A.9.: Labelled prominence values from *Emo-DB* for sentence 9.

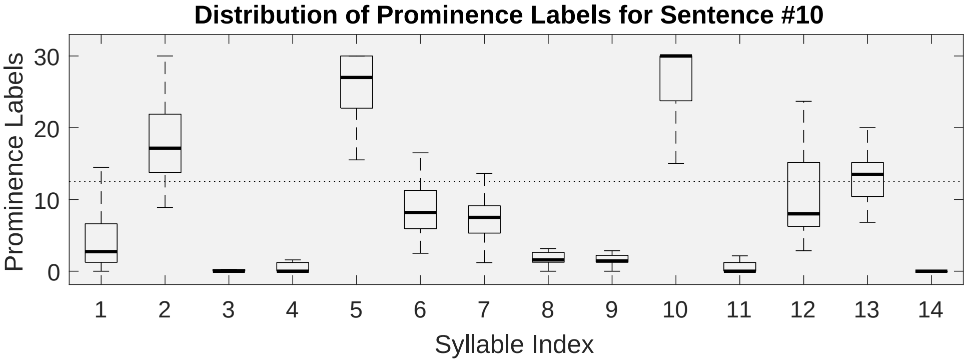


Figure A.10.: Labelled prominence values from *Emo-DB* for sentence 10.

Regression Accuracies for Speaker-Specific and Global Models

Speaker	Score	Regressors in best-performing Model														
		P	$P_{\Delta,l}$	$P_{\Delta,r}$	$P_{\div,l}$	$P_{\div,r}$	L	$L_{\Delta,l}$	$L_{\Delta,r}$	$L_{\div,l}$	$L_{\div,r}$	D	$D_{\Delta,l}$	$D_{\Delta,r}$	$D_{\div,l}$	$D_{\div,r}$
1	86.64%	0	1	1	1	0	0	0	0	1	1	0	1	1	0	0
2	84.98%	1	0	0	0	1	1	1	0	0	1	1	1	1	0	0
3	82.04%	0	0	1	0	1	1	0	1	0	1	1	0	0	0	0
4	87.51%	1	1	0	0	1	0	1	0	1	0	1	1	0	1	1
5	80.86%	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
6	92.99%	0	1	0	1	0	0	1	1	1	1	1	1	1	1	1
7	87.72%	0	1	1	1	0	1	1	0	0	1	1	1	0	1	0
8	77.91%	1	0	0	1	0	1	0	0	1	1	1	1	1	1	1
9	83.49%	0	1	0	0	1	0	0	0	1	1	0	0	0	1	0
10	82.21%	1	1	1	0	0	1	1	1	0	1	0	1	1	1	0
ALL	79.02%	0	1	0	0	0	0	0	1	1	0	1	0	0	1	0

Table A.1.: Best-performing sets of regressors for prominence prediction, both for individual speakers and globally, using **absolute values**.

Speaker	Score	Regressors in best-performing Model														
		P	$P_{\Delta,l}$	$P_{\Delta,r}$	$P_{\div,l}$	$P_{\div,r}$	L	$L_{\Delta,l}$	$L_{\Delta,r}$	$L_{\div,l}$	$L_{\div,r}$	D	$D_{\Delta,l}$	$D_{\Delta,r}$	$D_{\div,l}$	$D_{\div,r}$
1	85.98%	0	1	1	1	0	0	0	1	0	1	1	0	1	1	1
2	85.88%	0	1	1	1	1	1	1	1	1	1	0	0	0	1	0
3	85.02%	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0
4	91.18%	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1
5	81.08%	1	1	0	1	0	0	0	1	0	1	1	1	0	0	1
6	100.00%	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	84.17%	0	1	1	1	0	0	1	1	1	1	1	1	0	1	1
8	75.48%	1	0	1	1	0	0	0	1	1	1	1	0	1	0	0
9	82.45%	1	1	0	1	1	0	1	0	1	0	0	1	1	1	1
10	76.55%	0	0	1	1	1	1	0	0	0	1	1	1	0	1	0
ALL	75.55%	1	1	1	0	0	1	1	0	1	1	1	1	1	1	0

Table A.2.: Best-performing sets of regressors for prominence prediction, both for individual speakers and globally, using **relative values**.

A.2. *Emo-DB* Parameter Selection

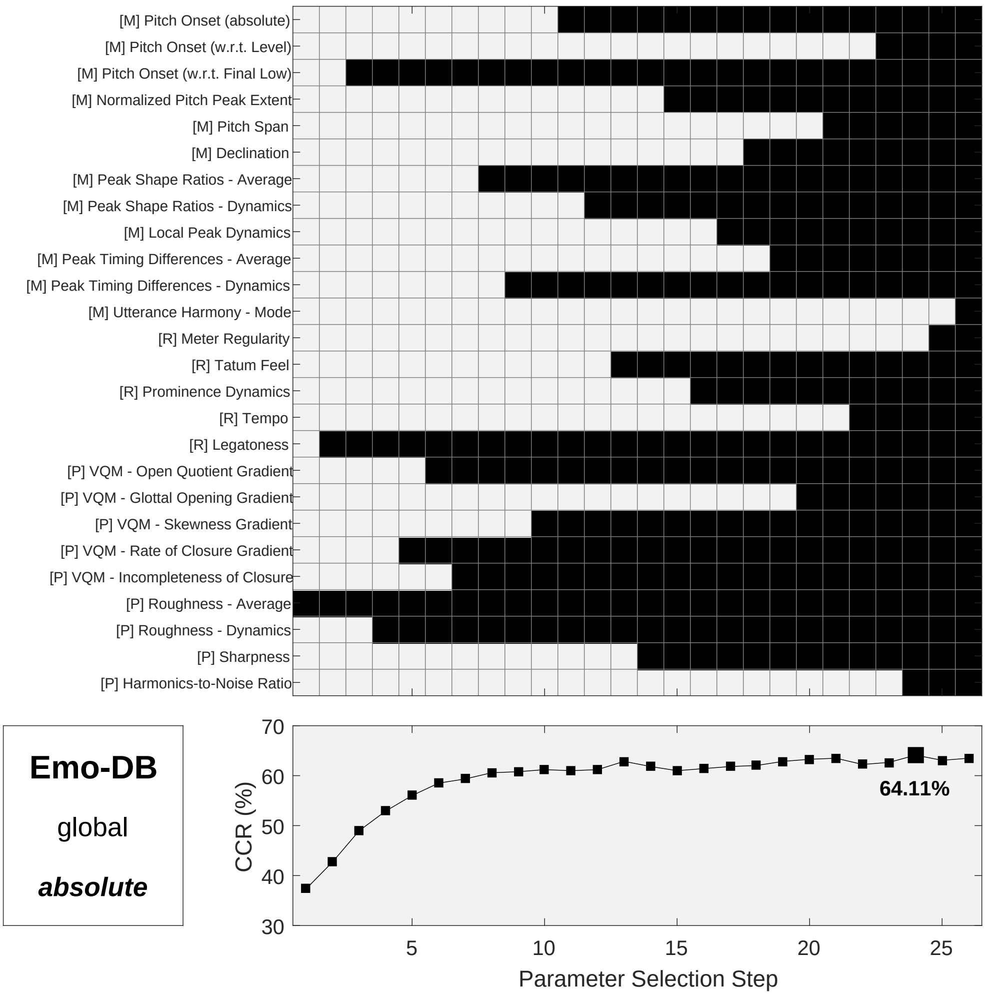


Figure A.11.: Sequential **forward** parameter selection for *Emo-DB* using **absolute** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

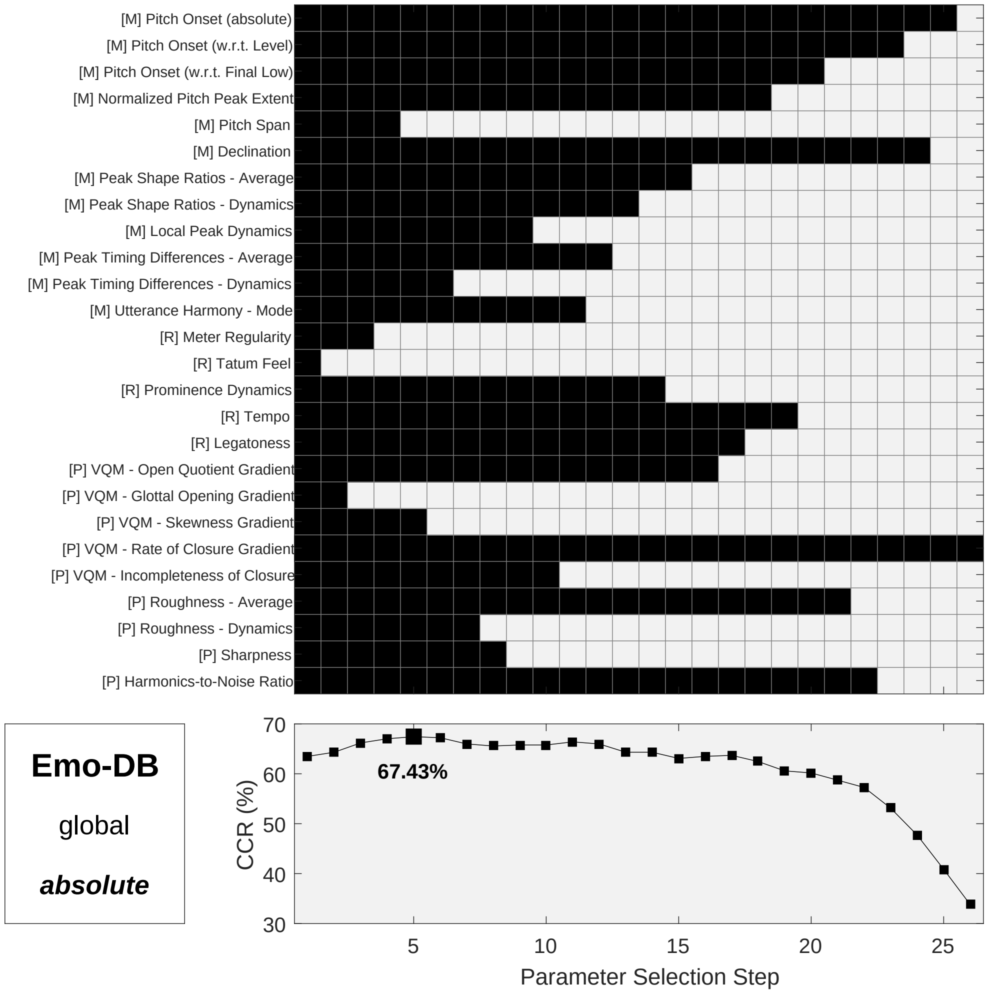


Figure A.12.: Sequential **backward** parameter selection for *Emo-DB* using **absolute** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

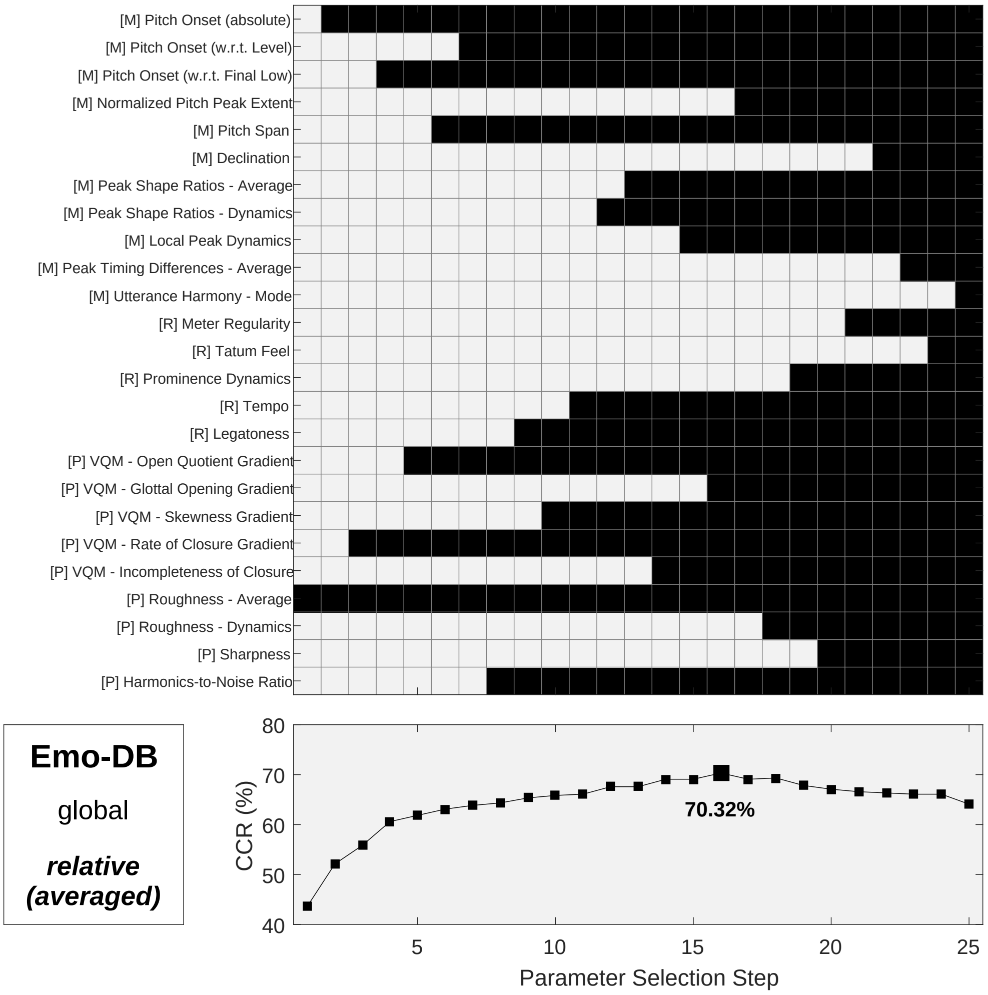


Figure A.13.: Sequential **forward** parameter selection for *Emo-DB* using **relative** parameter values w.r.t. **average** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

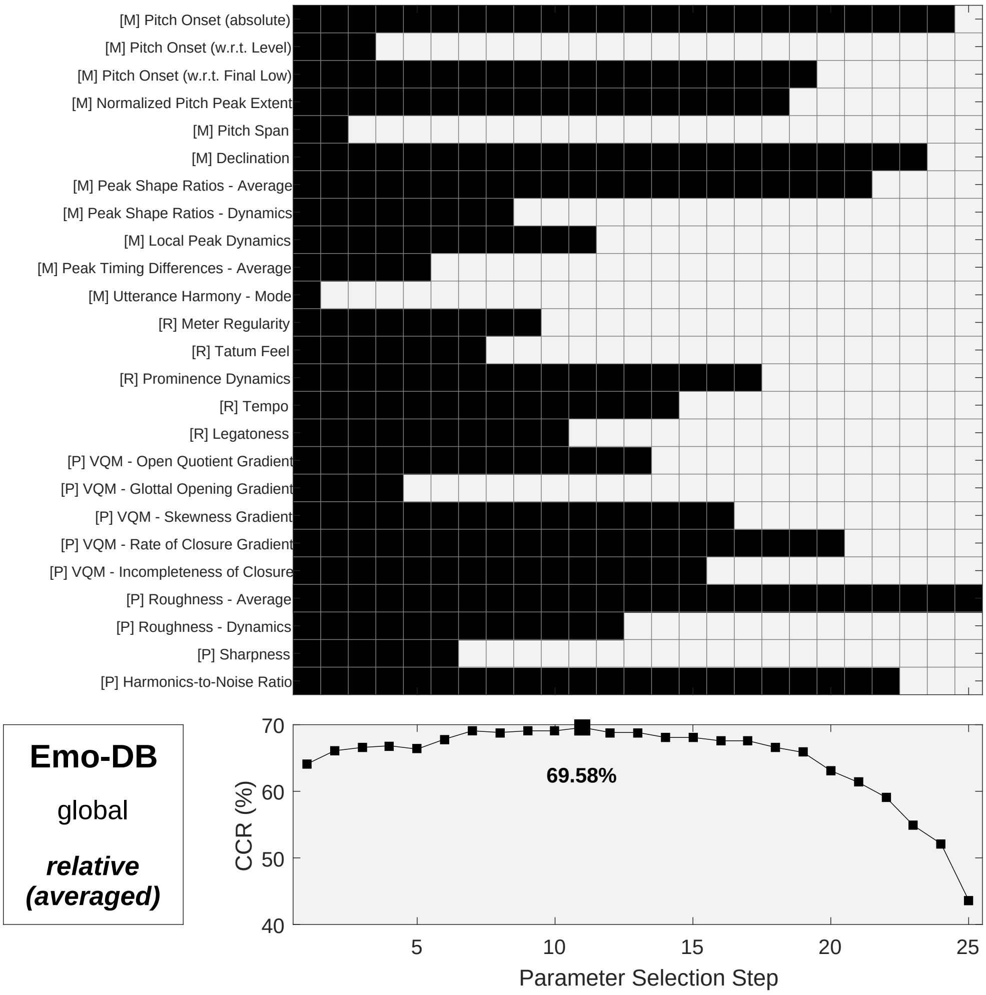


Figure A.14.: Sequential **backward** parameter selection for *Emo-DB* using **relative** parameter values w.r.t. **average** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

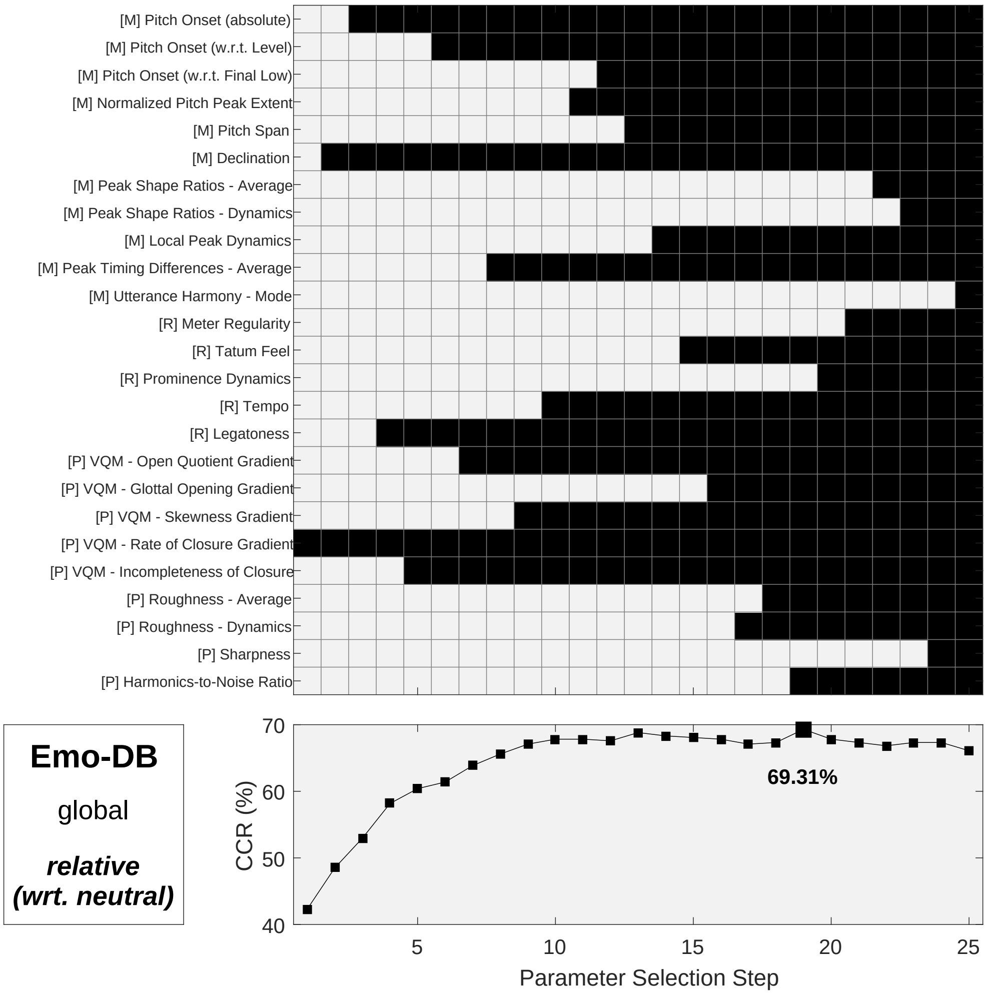


Figure A.15.: Sequential **forward** parameter selection for *Emo-DB* using **relative** parameter values w.r.t. the **"neutral"** emotion. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

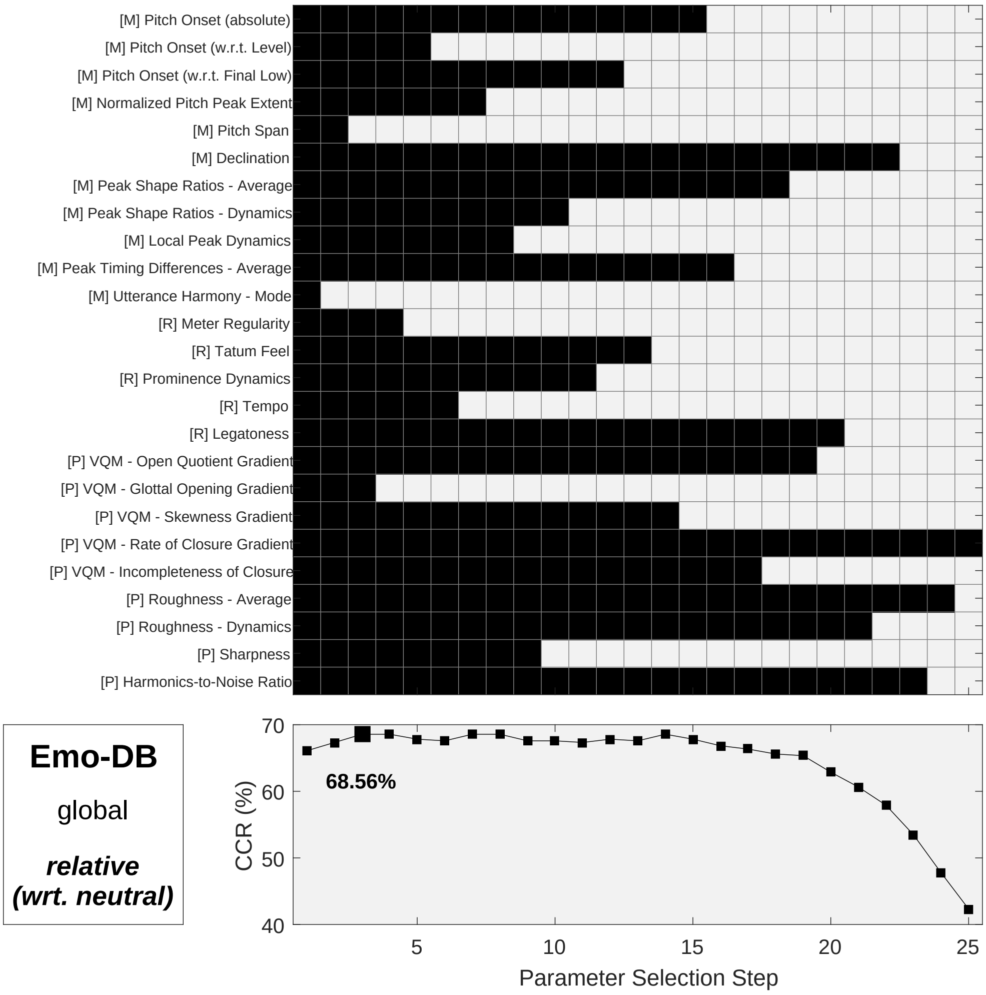


Figure A.16.: Sequential **backward** parameter selection for *Emo-DB* using **relative** parameter values w.r.t. the **"neutral"** emotion. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

A.3. *IEM-PSD* Parameter Selection

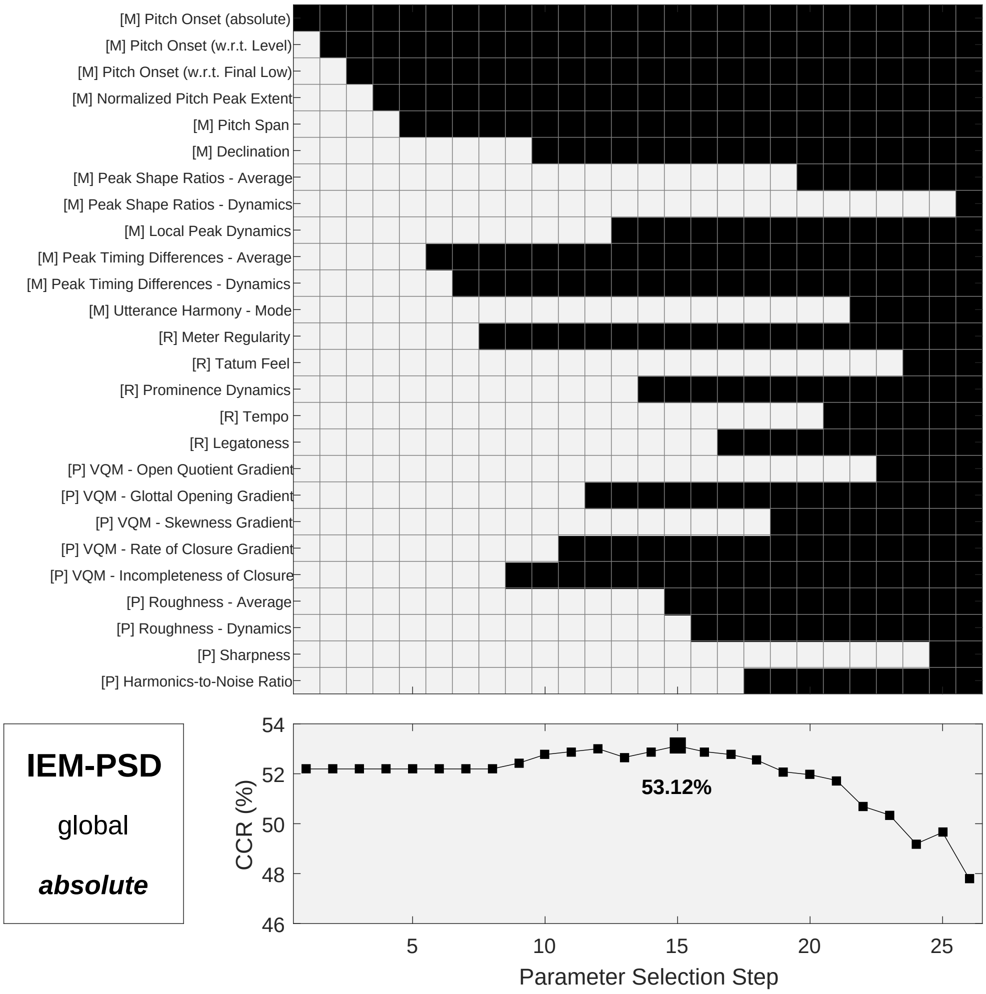


Figure A.17.: Sequential **forward** parameter selection for all **selected** events from *IEM-PSD* using **absolute** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

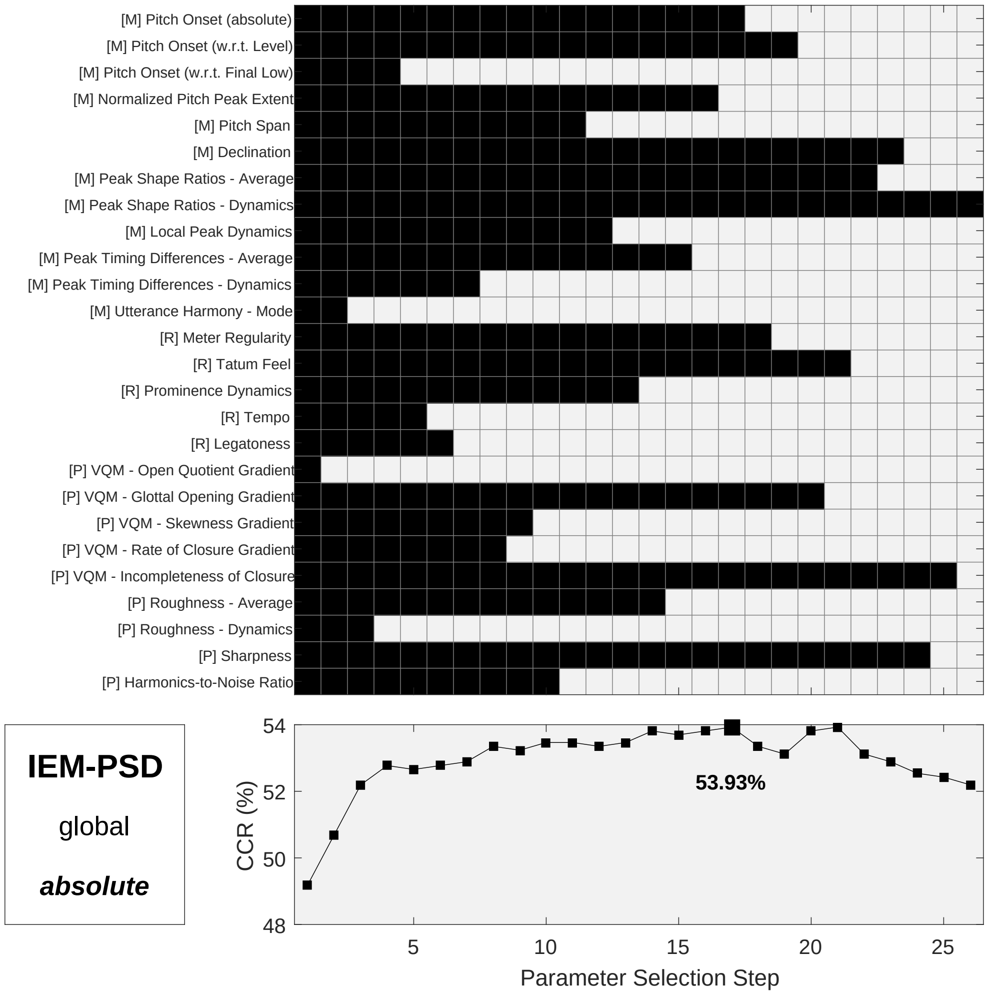


Figure A.18.: Sequential **backward** parameter selection for all **selected** events from *IEM-PSD* using **absolute** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

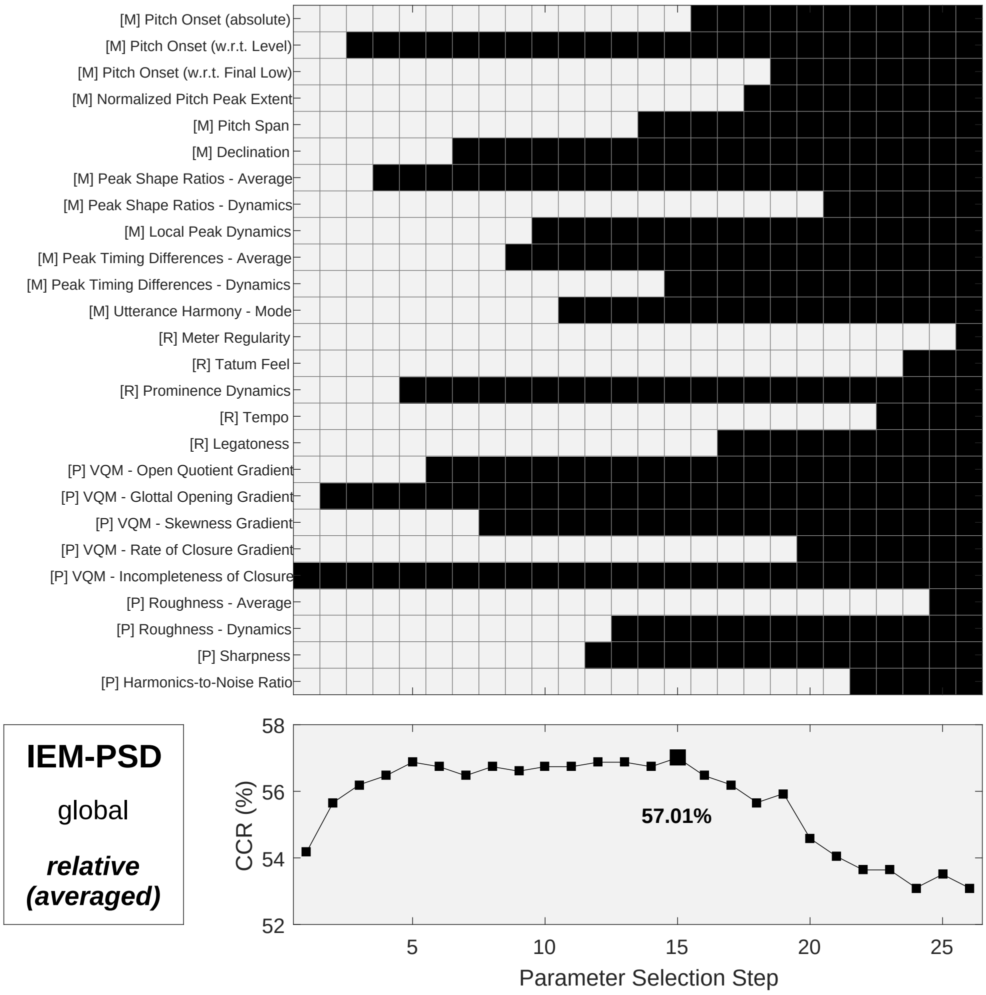


Figure A.19.: Sequential **forward** parameter selection for all **selected** events from *IEM-PSD* using **relative** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

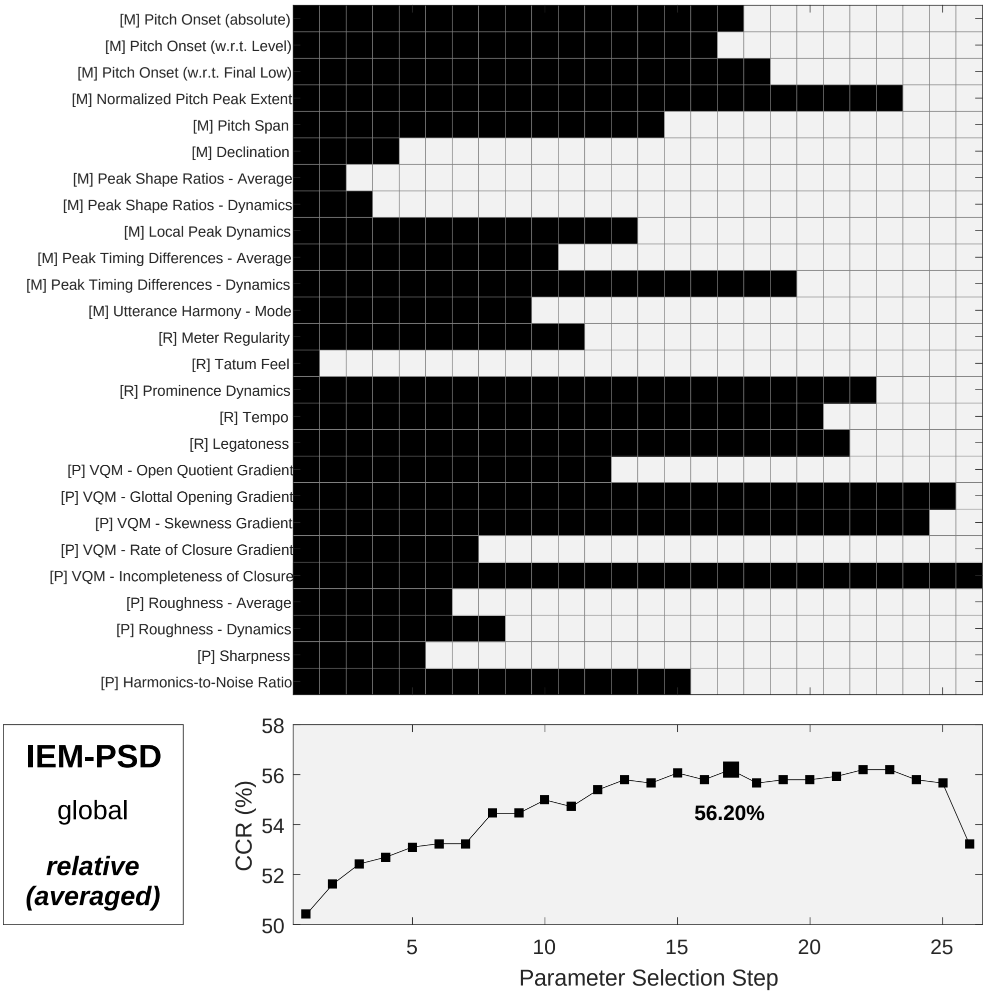


Figure A.20.: Sequential **backward** parameter selection for all **selected** events from *IEM-PSD* using **relative** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

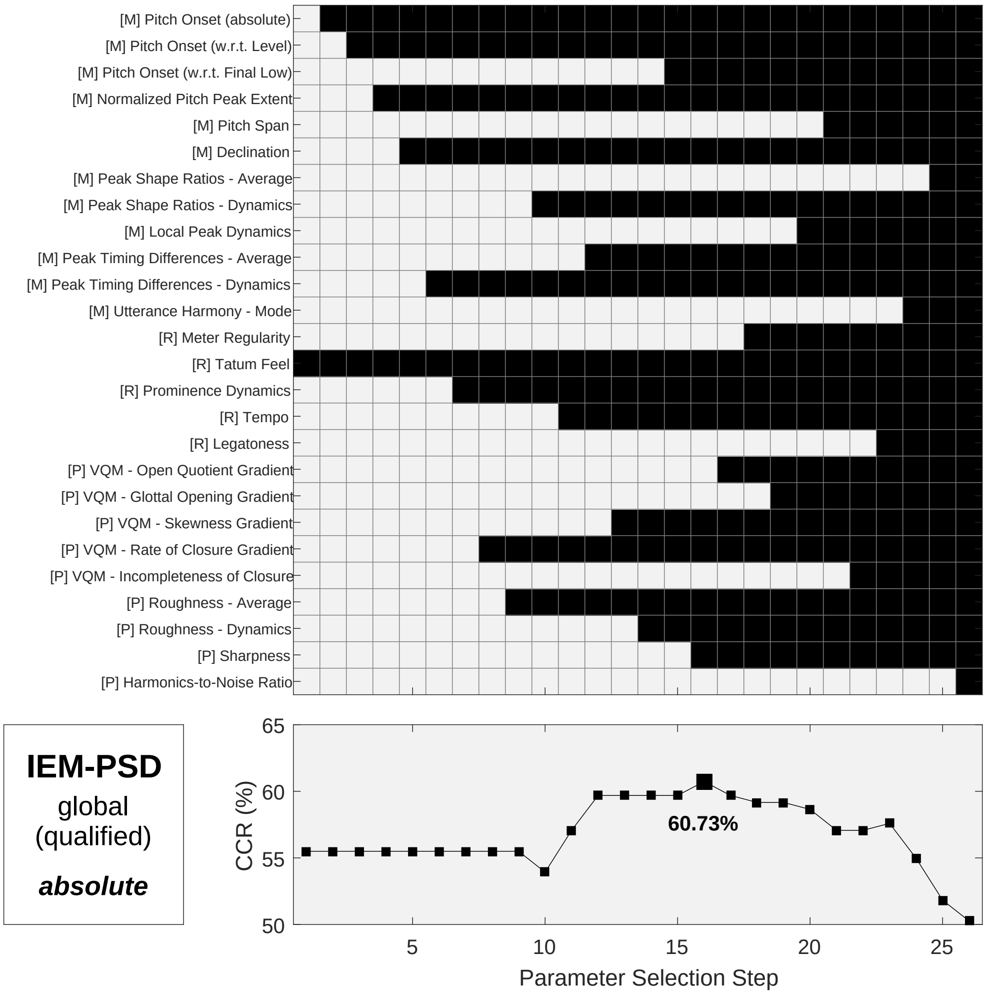


Figure A.21.: Sequential **forward** parameter selection for all **qualified** events from *IEM-PSD* using **absolute** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

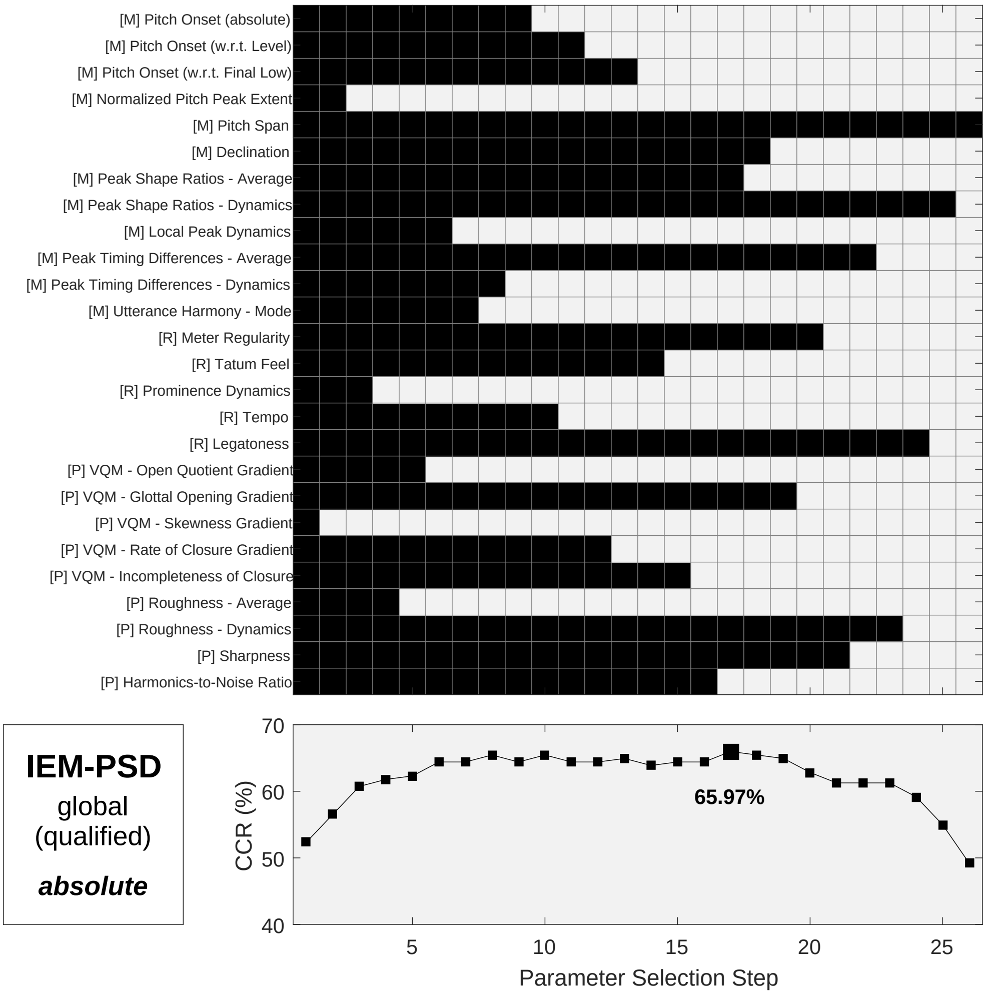


Figure A.22.: Sequential **backward** parameter selection for all **qualified** events from *IEM-PSD* using **absolute** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

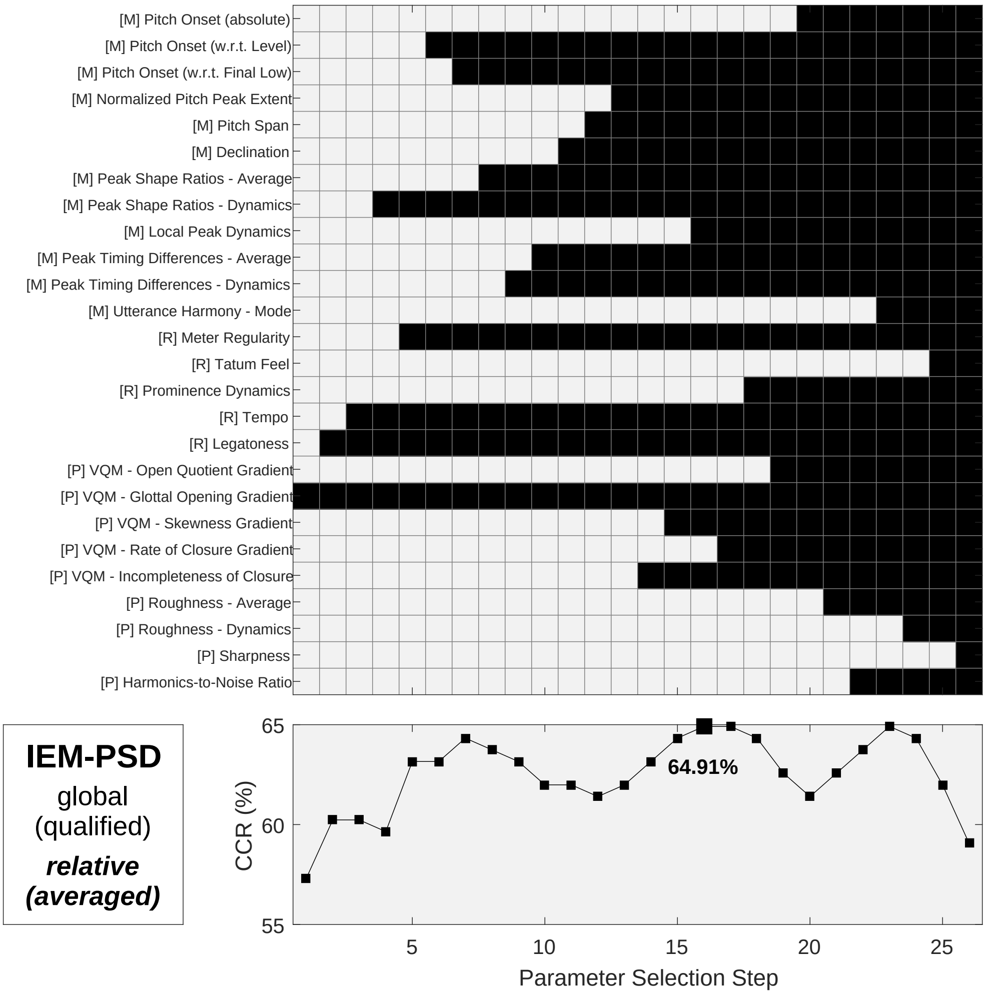


Figure A.23.: Sequential **forward** parameter selection for all **qualified** events from *IEM-PSD* using **relative** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

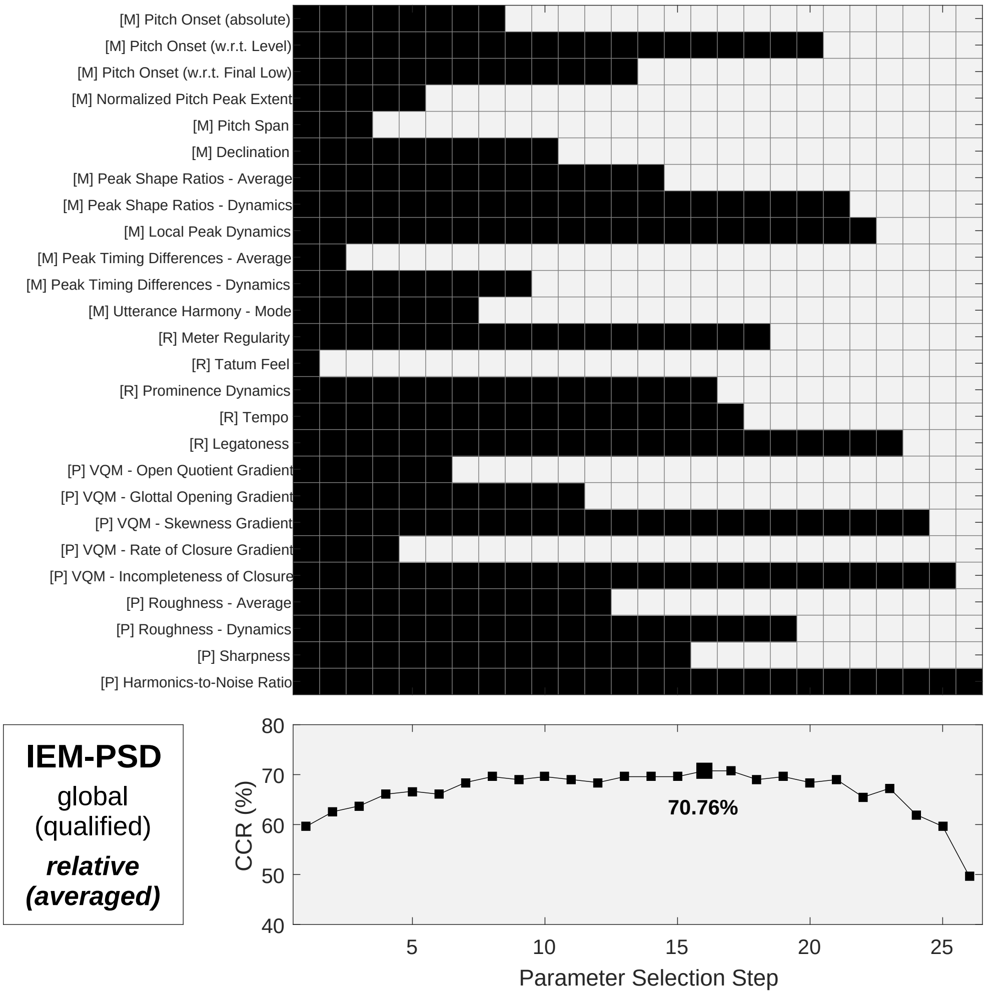


Figure A.24.: Sequential **backward** parameter selection for all **qualified** events from *IEM-PSD* using **relative** parameter values. Upper plot: parameters included in the evaluation set are marked black. Lower plot: correctly classified ratio (CCR) for SVM classification with the parameter set of the current selection step.

A.4. Parameter Variation Studies

Syllable Segmentation

Three parameters were evaluated in a full-factorial way with respect to the correct number of syllables detected by the algorithm:

Parameter	Values
Minimum Nucleus Length	[20, 30, 40, 50, 60, 70, 80] ms
Minimum Syllable Length	[40, 60, 80, 100, 120, 140] ms
MA Filter Span	[100, 150, 200, 250] ms

Table A.3.: Parameter values tested for syllable segmentation.

The following figure shows the average number of deviations between estimated and labeled syllables for each of the 168 possible combinations of these parameters:

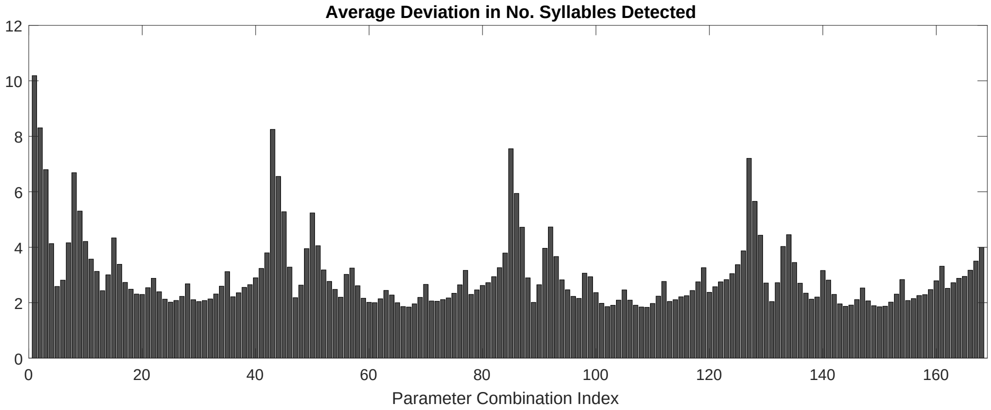


Figure A.25.: Average number of deviations between estimated and labeled syllables over all speakers for different parameter combinations.

The best parameter combination (min. nucleus length = 50ms, min. syllable length = 100ms, MA span = 200ms) results in an average error of 1.83 syllables per utterance.

Investigating the average error with the best parameter combination for each single speaker or each single emotion, respectively, reveals that there are no significant effects specific to the speakers or the emotions; the specific best result is an average deviation of 1.5 syllables per utterance, as shown in Fig. A.26 and Fig. A.27.

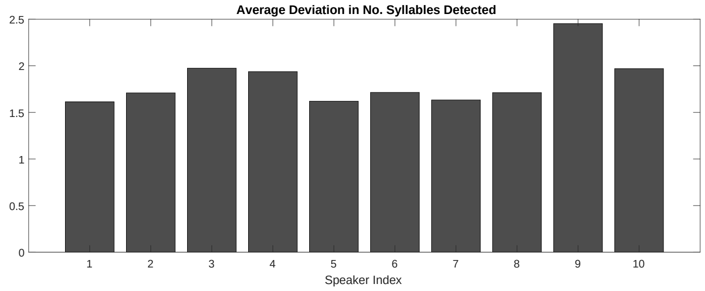


Figure A.26.: Average number of deviations between estimated and labeled syllables with optimum parameter combination for different speakers.

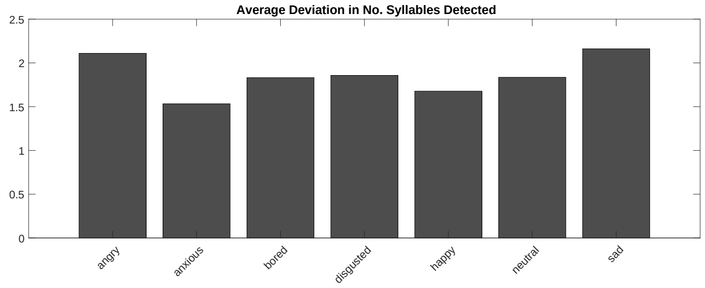


Figure A.27.: Average number of deviations between estimated and labeled syllables with optimum parameter combination for different emotions.

Bibliography

Abercrombie, D.
1967. Elements of general phonetics.

Al Moubayed, S., G. Ananthakrishnan, and L. Enflo
2010. Automatic prominence classification in swedish. In *Speech Prosody 2010 Proceedings*. Citeseer.

Allen, G. D.
1972. The location of rhythmic stress beats in english: An experimental study i. *Language and Speech*, 15(1):72–100.

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

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

Anagnostopoulos, C.-N., T. Iliou, and I. Giannoukos
2015. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177.

Ang, J., R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke
2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proceedings of Interspeech 2002*.

ANSI
1960. Usa standard acoustical terminology. *New York: American Standards Association*.

Auger, F., P. Flandrin, Y.-T. Lin, S. McLaughlin, S. Meignen, T. Oberlin, and H.-T. Wu
2013. Time-frequency reassignment and synchrosqueezing: An overview. *IEEE Signal Processing Magazine*, 30(6):32–41.

Aures, W.
1985. Berechnungsverfahren für den sensorischen wohlklang beliebiger schallsignale. *Acta Acustica united with Acustica*, 59(2):130–141.

Averty, P., S. Athenes, C. Collet, and A. Dittmar
2002. Evaluating a new index of mental workload in real atc situation using psychophysiological measures. *Digital Avionics Systems Conference, 2002. Proceedings. The 21st*, 2.

Baber, C., B. Mellor, R. Graham, J. Noyes, and C. Tunley
1996. Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication*, 20(1-2):37–53.

Baddeley, A.
2003. Working memory and language: An overview. *Journal of communication disorders*, 36(3):189–208.

Baddeley, A. D. and G. Hitch
1974. Working memory. *Psychology of learning and motivation*, 8:47–89.

Banse, R. and K. R. Scherer
1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.

Barra-Chicote, R., F. Fernandez, S. Lutfi, J. Lucas-Cuesta, J. Macias-Guarasa, J. Montero, R. San-Segundo, and J. Pardo
2009. Acoustic emotion recognition using dynamic bayesian networks and multi-space distributions. In *ISCA 2009, Brighton.*

Barrett, L. F.
1998. Discrete emotions or dimensions? the role of valence focus and arousal focus. *Cognition & Emotion*, 12(4):579–599.

Batliner, A., B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir
2011a. The automatic recognition of emotions in speech. *Emotion-Oriented Systems*, Pp. 71–99.

Batliner, A., S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, et al.
2011b. Whodunnit-searching for the most important feature types signalling emotion-related user states in speech. *Computer Speech & Language*, 25(1):4–28.

Beckman, M. E. and J. Edwards
1990. Of prosodic constituency. *Between the grammar and physics of speech*, P. 152.

Beckmann, M. and G. Ayers-Elam
1997. Guidelines for tobi labeling. *Unpublished ms. Version*, 3.

Bettermann, H., D. Amponsah, D. Cysarz, and P. Van Leeuwen
1999. Musical rhythms in heart period dynamics: a cross-cultural and interdisciplinary approach to cardiac rhythms. *American Journal of Physiology-Heart and Circulatory Physiology*, 277(5):H1762.

Bilmes, J. A.
1993. *Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm.* PhD thesis, Massachusetts Institute of Technology.

Boersma, P.
1993. 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.

Bolia, R. S. and R. E. Slyh
2003. Perception of stress and speaking style for selected elements of the susas database. *Speech Communication*, 40(4):493–501.

Boril, H., S. Sadjadi, and J. Hansen
2011. Utdrive: Emotion and cognitive load classification for in-vehicle scenarios. In *5th Biennial Workshop on DSP for In-Vehicle Systems, Kiel, Germany.*

Boser, B. E., I. M. Guyon, and V. N. Vapnik
1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, Pp. 144–152. ACM.

Bulut, M. and S. Narayanan
2008. On the robustness of overall f0-only modifications to the perception of emotions in speech. *The Journal of the Acoustical Society of America.*

Burges, C.
1998. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Burkhardt, F., A. Paeschke, M. Rolfes, and W. Sendlmeier
2005. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology (Interspeech '05).*

Cannon, W. B.
1932. The wisdom of the body.

CareerCast
2017. The most stressful jobs of 2017. (online: http://www.careercast.com/jobs-rated/most-stressful-jobs-2017, retrieved 27-Mar-2017).

Casale, S., A. Russo, G. Scebba, and S. Serrano
2008. Speech emotion classification using machine learning algorithms. In *Semantic Computing, IEEE Internation Converence on 26:33.*

Castaldo, R., P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia
2015. Acute mental stress assessment via short term hrv analysis in healthy adults: A systematic review with meta-analysis. *Biomedical Signal Processing and Control*, 18:370–377.

Clark, R. A.
1999. Using prosodic structure to improve pitch range variation in text to speech synthesis. International Congress of Phonetic Sciences.

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

Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor
2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80.

Crystal, D.
2011. *Dictionary of linguistics and phonetics*, volume 30. Wiley. com.

Cummins, F.
2002. Speech rhythm and rhythmic taxonomy. In *Speech Prosody 2002, International Conference*. Citeseer.

Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri
2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Daniel, P. and R. Weber
1993. Calculating psychoacoustical roughness. In *Proc. Internoise*, Pp. 251–265.

Daniel, P. and R. Weber
1997. Psychoacoustical roughness: Implementation of an optimized model. *Acta Acustica united with Acustica*, 83(1):113–123.

Darwin, C.
1872. The expression of the emotions in man and animals.

Dellaert, F., T. Polzin, and A. Waibel
1996. Recognizing emotion in speech. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, Pp. 1970–1973. IEEE.

Douglas-Cowie, E., N. Campbell, R. Cowie, and P. Roach
2003. Emotional speech: Towards a new generation of databases. *Speech Communication*.

Drach, E.
1926. *Die redenden Künste*, volume 221. Quelle & Meyer.

Duda, R. O., P. E. Hart, and D. G. Stork
2012. *Pattern classification*. John Wiley & Sons.

Ekman, P. and W. V. Friesen
1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124.

Ellis, D.
2006. Speech and audio processing and recognition, lecture 5: Speech modeling. online (retrieved 26-May-2010).

Fairbanks, G. and L. W. Hoaglin
1941. An experimental study of the durational characteristics of the voice during the expression of emotion. *Communications Monographs*, 8(1):85–90.

Fairbanks, G. and W. Pronovost
1939. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, 6(1):87–104.

Fant, G.
1968. Analysis and synthesis of speech processes. *Manual of phonetics*, 2:173–277.

Fant, G.
1970. *Acoustic Theory of Speech Production*, number 2. Walter de Gruyter.

Fant, G. and A. Kruckenberg
1989. Preliminaries to the study of swedish prose reading and reading style. *STL-QPSR*, 2(1989):1–83.

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

Fischer, S. R.
2001. *History of writing*. Reaktion Books.

Fletcher, H.
1940. Auditory patterns. *Reviews of modern physics*, 12(1):47.

Fontaine, J. R., K. R. Scherer, E. B. Roesch, and P. C. Ellsworth
2007. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

France, D. J., R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes
2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837.

Friberg, A. and J. Sundberg
1995. Time discrimination in a monotonic, isochronous sequence. *The Journal of the Acoustical Society of America*, 98(5):2524–2531.

Fry, D. B.
1958. Experiments in the perception of stress. *Language and speech*, 1(2):126–152.

Gobl, C. and A. Chasaide
2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*.

Godin, K. and J. Hansen
2008. Analysis and perception of speech under physical task stress. *Interspeech 2008*.

Gordon, J. W.
1987. The perceptual attack time of musical tones. *The Journal of the Acoustical Society of America*, 82(1):88–105.

Goslin, J. and U. H. Frauenfelder
1999. Syllable segmentation: are humans consistent? *Proceedings of Eurospeech 1999*.

Grabe, E. and E. Low
2002. Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7(515-546).

Griffin, G. and C. Williams
1987. The effects of different levels of task complexity on three vocal measures. *Aviation, space, and environmental medicine.*

Grimm, M., K. Kroschel, E. Mower, and S. Narayanan
2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800.

Hansen, J. and S. Bou-Ghazale
1997. Getting started with susas: A speech under simulated and actual stress database. *Fifth European Conference on Speech Communication and Technology (Eurospeech '97).*

Hansen, J. and S. Patil
2007. Speech under stress: Analysis, modeling and recognition. *Lecture Notes in Computer Science.*

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

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

Hansen, J., X. Zhang, M. Akbacak, U. Yapanel, B. Pellom, W. Ward, and P. Angkititrakul
2005. Cu-move: advanced in-vehicle speech systems for route navigation. *DSP for in-vehicle and mobile systems*, Pp. 19–45.

Harsin, C. A.
1997. Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception & psychophysics*, 59(2):243–251.

Hartmann, W. M.
2004. *Signals, sound, and sensation.* Springer Science & Business Media.

Hirst, D., A. Di Cristo, and R. Espesser
2000. Levels of representation and levels of analysis for the description of intonation systems. In *Prosody: Theory and experiment*, Pp. 51–87. Springer.

Hirst, D. and R. Espesser
1993. Automatic modelling of fundamental frequency using a quadratic spline function.

Hjortskov, N., D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard
2004. The effect of mental stress on heart rate variability and blood pressure during computer work. *European journal of applied physiology*, 92(1-2):84–89.

Höldrich, R. and M. Pflüger
1999. A parameterized model of psychoacoustical roughness for objective vehicle noise quality evaluation. In *Proc. 137th meeting of the Acoustical Society of America, Berlin*.

Honda, K.
2008. Physiological processes of speech production. In *Springer Handbook of Speech Processing*, Pp. 7–26. Springer.

Howell, P.
1988. Prediction of p-center location from the distribution of energy in the amplitude envelope: I. *Attention, Perception, & Psychophysics*, 43(1):90–93.

Huttunen, K., H. Keränen, E. Väyrynen, R. Pääkkönen, and T. Leino
2011. Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied Ergonomics*, 42(2):348–357.

Ikeno, A., V. Varadarajan, S. Patil, and J. Hansen
2007. Ut-scope: Speech under lombard effect and cognitive stress. In *IEEE Aerospace Conference*, Pp. 1–7.

Ishi, C. and N. Campbell
2002. Analysis of acoustic-prosodic features of spontaneous expressive speech. *First International Phonetics & Phonlogy*.

Jameson, A., J. Kiefer, C. M
"uller, B. Großmann-Hutter, F. Wittig, and R. Rummer
2006. Assessment of a user's time pressure and cognitive load on the basis of features of speech. *Journal of Computer Science and Technology*.

Johns-Lewis, C.
1986. Prosodic Differentiation Of Discourse Modes. *Intonation in Discourse*, Pp. 199–220.

Juslin, P. N. and P. Laukka
2003. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770.

Kahn, D.
2015. *Syllable-based generalizations in English phonology*, volume 15. Routledge.

Karam, Z. N., E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. G. Mcinnis
2014. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Pp. 4858–4862. IEEE.

Kaufmann, T., S. Sütterlin, S. Schulz, and C. Vögele
2011. Artiifact: a tool for heart rate artifact processing and heart rate variability analysis. *Behavior Research Methods*, Pp. 1–10.

Kehrein, R.
2002. *Prosodie und Emotionen*. Niemeyer Tübingen.

Kent, R. D.
2000. Research on speech motor control and its disorders: A review and prospective. *Journal of Communication disorders*, 33(5):391–428.

Kohler, K.
2005. Timing and communicative functions of pitch contours. *Phonetica*, 62(2-4):88–105.

Koolhaas, J., A. Bartolomucci, B. d. Buwalda, S. De Boer, G. Flügge, S. Korte, P. Meerlo, R. Murison, B. Olivier, P. Palanza, et al.
2011. Stress revisited: a critical evaluation of the stress concept. *Neuroscience & Biobehavioral Reviews*, 35(5):1291–1301.

Krajewski, J., S. Schnieder, D. Sommer, A. Batliner, and B. Schuller
2012. Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech. *Neurocomputing*, 84:65 – 75. From Neuron to Behaviour: Evidence from Behavioral Measurements.

Kuroda, I., O. Fujiwara, N. Okamura, and N. Utsuki
1976. Method for determining pilot stress through analysis of voice communication. *Aviation, Space, and Environmental Medicine*.

Lackner, H. K., I. Papousek, J. J. Batzel, A. Roessler, H. Scharfetter, and H. Hinghofer-Szalkay
2011. Phase synchronization of hemodynamic variables and respiration during mental challenge. *International journal of psychophysiology*, 79(3):401–409.

Ladd, D., K. Silverman, F. Tolkmitt, and G. Bergmann
1985. Evidence for the independent function of intonation contour type, voice quality, and f0 range in .... *The Journal of the Acoustical Society of America*.

Ladefoged, P.
1971. *Preliminaries to linguistic phonetics*. University of Chicago Press.

Laver, J.
1980. The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31:1–186.

Lazarus, R. S.
1998. From psychological stress to the emotions: A history of changing outlooks. *Fifty Years of the Research and Theory of RS Lazarus: An Analysis of Historical and Perennial Issues*, P. 349.

Lazarus, R. S.
2006. *Stress and emotion: A new synthesis*. Springer Publishing Company.

Lecluse, F., M. Brocaar, and J. Verschuure
1975. The electroglottography and its relation to glottal activity. *Folia Phoniatrica et Logopaedica*, 27(3):215–224.

Lee, C. M., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan
2004. Emotion recognition based on phoneme classes. In *Interspeech*, Pp. 205–211.

Levelt, W. J.
1999. Models of word production. *Trends in cognitive sciences*, 3(6):223–232.

Lindsay, P. and D. Norman
1977. *Human information processing: an introduction to psychology.* New York [etc.]: Academic Press.

Liscombe, J.
2007. *Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency*. PhD thesis, Spoken Language Processing Group, Columbia University.

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

Lugger, M. and B. Yang
2007. The relevance of voice quality features in speaker independent emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, Pp. IV–17. IEEE.

Lugger, M., B. Yang, and W. Wokurek
2006. Robust estimation of voice quality parameters under realworld disturbances. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, Pp. I–I. IEEE.

Luig, J.
2009. Investigations on A Robust Feature Set for Classification of Speech Under Stress. Master's thesis, Graz University of Technology.

Luig, J.
2011. IEM Report 45/11: The IEM Pilot Speech Database. Technical report, Institute of Electronic Music and Acoustics.

Maghbouleh, A.
1998. Tobi accent type recognition. *ISSUES*, 3:1.

Makhoul, J. and L. Cosell
1976. Lpcw: An lpc vocoder with linear predictive spectral warping. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'76.*, volume 1, Pp. 466–469. IEEE.

Malik, M., J. Bigger, A. Camm, R. Kleiger, A. Malliani, A. Moss, et al.
1996. Guidelines: Heart rate variability, standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17:354–381.

Marcus, S. M.
1981. Acoustic determinants of perceptual center (p-center) location. *Attention, Perception, & Psychophysics*, 30(3):247–256.

McCorry, L. K.
2007. Physiology of the autonomic nervous system. *American journal of pharmaceutical education*, 71(4):78.

Meinedo, H., J. P. Neto, L. B. Almeida, et al.
1999. Syllable onset detection applied to the portuguese language. In *in Proceedings EUROSPEECH 99*. Citeseer.

Melillo, P., M. Bracale, and L. Pecchia
2011. Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination. *Biomedical engineering online*, 10(1):96.

Mermelstein, P.
1975. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am*, 58(4):880–883.

Mixdorff, H.
2000. A novel approach to the fully automatic extraction of fujisaki model parameters. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3, Pp. 1281–1284. IEEE.

Mixdorff, H.
2002. Speech technology, tobi, and making sense of prosody. In *Speech Prosody 2002, International Conference*.

Morton, J., S. Marcus, and C. Frankish
1976. Perceptual centers (p-centers). *Psychological Review*, 83(5):405.

Mozziconacci, S. J. and D. J. Hermes
1997. A study of intonation patterns in speech expressing emotion or attitude: production and perception. *IPO Annual Progress Report*, 32:154–160.

Murray, I., C. Baber, and A. South
1996. Towards a definition and working model of stress and its effects on speech. *Speech Communication*, P. 10.

Murray, I. R. and J. L. Arnott
1993. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *The Journal of the Acoustical Society of America*, 93(2):1097–1108.

Nawka, T.
1987. Die auditive bewertung heiserer stimmen nach dem rbh-system. *Sprache Stimme Gehör*, 18:130–133.

Nicholson, J., K. Takahashi, and R. Nakatsu
1999. Emotion recognition in speech using neural networks. In *Neural Information Processing, 1999. Proceedings. ICONIP'99. 6th International Conference on*, volume 2, Pp. 495–501. IEEE.

Niebuhr, O.
2007. The signalling of german rising-falling intonation categories–the interplay of synchronization, shape, and height. *Phonetica*, 64(2-3):174–193.

Noetzel, A.
1991. Robust syllable segmentation of continuous speech using neural networks. In *Electro International, 1991*, Pp. 580–585. IEEE.

Nwe, T. L., S. W. Foo, and L. C. De Silva
2003. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

Oppenheim, A. V. and R. W. Schafer
2004. From frequency to quefrency: A history of the cepstrum. *IEEE signal processing Magazine*, 21(5):95–106.

Paeschke, A. and W. F. Sendlmeier
2000. Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Pal, P., A. N. Iyer, and R. E. Yantorno
2006. Emotion detection from infant facial expressions and cries. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 2, Pp. II–II. IEEE.

Papousek, I., K. Nauschnegg, M. Paechter, H. K. Lackner, N. Goswami, and G. Schulter
2010. Trait and state positive affect and cardiovascular recovery from experimental academic stress. *Biological Psychology*, 83(2):108–115.

Patterson, D. and D. R. Ladd
1999. Pitch range modelling: linguistic dimensions of variation. In *Proceedings of ICPhS*, volume 99, Pp. 1169–1172.

Paulus, E. and E. Zwicker
1972. Programme zur automatischen bestimmung der lautheit aus terzpegeln oder frequenzgruppenpegeln. *Acustica*, 27(253-266):17.

Petridis, S. and M. Pantic
2008. Audiovisual discrimination between laughter and speech. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, Pp. 5117–5120. IEEE.

Pike, K. L.
1945. The intonation of american english.

Plutchik, R.
1980. *Emotion: A psychoevolutionary synthesis.* Harper & Row New York.

Plutchik, R.
2001. The nature of emotions human emotions have deep evolutionary roots, a

fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Polzin, T. S. and A. Waibel
2000. Emotion-sensitive human-computer interfaces. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.*

Pompino-Marschall, B.
1989. On the psychoacoustic nature of the p-center phenomenon. *Journal of phonetics.*

Protopapas, A. and P. Lieberman
1997. Fundamental frequency of phonation and perceived emotional stress. *The Journal of the Acoustical Society of America*, 101(4):2267–2277.

Rajasekaran, P., G, and J. Picone
1986. Recognition of speech under stress and in noise. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '86)*, Pp. 1–4.

Rakowski, A.
1971. Pitch discrimination at the threshold of hearing. In *Proceedings of the Seventh International Congress on Acoustics*, volume 3.

Ramus, F., M. Nespor, and J. Mehler
1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292.

Rothkrantz, L., P. Wiggers, J. van Wees, and R. van Vark
2004. Voice stress analysis. *Lecture Notes in Computer Science*, Pp. 449–456.

Ruiz, R., P. P. de Hugues, and C. Legros
2010. Advanced voice analysis of pilots to detect fatigue and sleep inertia. *Acta Acustica united with Acustica*, 96:567–579(13).

Sabo, R., M. Rusko, A. Ridzik, and J. Rajčáni
2016. Stress, arousal, and stress detector trained on acted speech database. In *International Conference on Speech and Computer*, Pp. 675–682. Springer.

Scherer, K., D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger
2002. Acoustic correlates of task load and stress. *Seventh International Conference on Spoken Language Processing.*

Scherer, K. R.
1986. Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99(2):143.

Schötz, S.
2003. Prosody in relation to paralinguistic phonetics-earlier and recent definitions, distinctions and discussions. *Term paper for course in Prosody, Lund University, Dept. of Linguistics and Phonetics.*

Schuller, B., G. Rigoll, and M. Lang
2003. Hidden markov model-based speech emotion recognition. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 1, Pp. I–401. IEEE.

Schuller, B., J. Stadermann, and G. Rigoll
2006. Affect-robust speech recognition by dynamic emotional adaptation. In *In Proc. ISCA Speech Prosody 2006, Dresden.*

Schuller, B. W., A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, et al.
2007. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Interspeech*, Pp. 2253–2256.

Scott, S. K.
1993. *P-centres in speech an acoustic analysis.* PhD thesis, University College.

Scripture, E.
1921. A study of emotions by speech transcription. *Vox*, 31:179–183.

Selye, H.
1950. The physiology and pathology of exposure to stress.

Shastri, L., S. Chang, and S. Greenberg
1999. Syllable detection and segmentation using temporal flow neural networks. In *International Congress of Phonetic Sciences*, Pp. 1721–1724.

Sigmund, M.
2006. Introducing the database examstress for speech under stress. In *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, Pp. 290–293. IEEE.

Skinner, E. R.
1935. A calibrated recording and analysis of the pitch, force and quality of vocal tones expressing happiness and sadness; and a determination of the pitch and force of the subjective concepts of ordinary, soft, and loud tones. *Communications Monographs*, 2(1):81–137.

Sontacchi, A.
1998. Entwicklung eines modulkonzeptes für die psychoakustische geräuschanalyse unter matlab. Master's thesis, Masters Thesis, Technische Universität Graz.

Stevens, K. and C. Williams
1969. On determining the emotional state of pilots during flight - an exploratory study. *Aerospace Medicine*, 40:1369–1372.

Stevens, K. N. and H. M. Hanson
1995. Classification of glottal vibration from acoustic measurements. *Vocal fold physiology: Voice quality control*, Pp. 147–170.

Stevens, S. S.
1936. A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43(5):405.

Stevens, S. S., J. Volkmann, and E. B. Newman
1937. A scale for the measurement of the psychological magnitude: pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Streefkerk, B. M.
2002. *Prominence. Acoustic and lexical/syntactic correlates.* LOT.

Streeter, L. A., N. H. Macdonald, W. Apple, R. M. Krauss, and K. M. Galotti
1983. Acoustic and perceptual indicators of emotional stress. *The Journal of the Acoustical Society of America*, 73(4):1354–1360.

Stroop, J.
1935. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, 18:643–662.

t Hart, J.
1981. Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, 69(3):811–821.

Talkin, D.
1995. A robust algorithm for pitch tracking (rapt).

Tamburini, F. and P. Wagner
2007. On automatic prominence detection for german. In *Eighth Annual Conference of the International Speech Communication Association.*

Terken, J.
1991. Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America*, 89(4):1768–1776.

Tharion, E., S. Parthasarathy, and N. Neelakantan
2009. Short-term heart rate variability measures in students during examinations. *The National Medical Journal of India*, 22(2).

Traina, M., A. Cataldo, F. GAlullo, and G. Russo
2011. Effect. of anxiety due to mental stress on heart rate variability in healthy subjects. *Minerva Psichiatr*, 52:227–31.

Truong, K. P., A. Nieuwenhuys, P. Beek, and V. Evers
2015. A database for analysis of speech under physical stress: detection of exercise intensity while running and talking.

Van Bezooijen, R.
1984. *Characteristics and recognizability of vocal expressions of emotion*, volume 5. Walter de Gruyter.

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

Villing, R.
2010. *Hearing the Moment: Measures and Models of the Perceptual Centre.* PhD thesis, National University of Ireland Maynooth.

Villing, R., J. Timoney, T. Ward, and J. Costello
2004. Automatic blind syllable segmentation for continuous speech. In *Proceedings of ISSC*. Citeseer.

Villing, R., T. Ward, and J. Timoney
2006. Performance limits for envelope based automatic syllable segmentation. In *Irish Signals and Systems Conference, 2006. IET*, Pp. 521–526.

Visnovcova, Z., M. Mestanik, M. Javorka, D. Mokra, M. Gala, A. Jurko, A. Calkovska, and I. Tonhajzerova
2014. Complexity and time asymmetry of heart rate variability are altered in acute mental stress. *Physiological measurement*, 35(7):1319.

Vlasenko, B., B. Schuller, A. Wendemuth, and G. Rigoll
2007a. Combining frame and turn-level information for robust recognition of emotions within speech. *iesk.et.uni-magdeburg.de*.

Vlasenko, B., B. Schuller, A. Wendemuth, and G. Rigoll
2007b. Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. *LECTURE NOTES IN COMPUTER SCIENCE*.

Vuksanović, V. and V. Gal
2007. Heart rate variability in mental stress aloud. *Medical engineering & physics*, 29(3):344–349.

Wagner, P.
2005. Great expectations-introspective vs. perceptual prominence ratings and their acoustic correlates. In *Interspeech 2005*.

Wagner, P.
2008. *The rhythm of language and speech: Constraining factors, models, metrics and applications*. h, Rheinische Friedrich-Wilhelms-Universit¨at Bonn.

Wang, Y., S. Du, and Y. Zhan
2008. Adaptive and optimal classification of speech emotion recognition. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 5, Pp. 407–411. IEEE.

Ward Jr, J.
1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

WikimediaCommons
2005. Glottis positions (image). online (https://commons.wikimedia.org/wiki/File:Glottis_positions.png), retrieved 30-Mar-2017.

Williams, C. E. and K. N. Stevens
1972. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250.

Wu, S.-L., E. Kingsbury, N. Morgan, and S. Greenberg
1998. Incorporating information from syllable-length time scales into automatic

speech recognition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 2, Pp. 721–724. IEEE.

Yang, B. and M. Lugger
2010. Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90(5):1415–1423.

Yang, T., J. Yang, and F. Bi
2009. Emotion statuses recognition of speech signal using intuitionistic fuzzy set. In *Software Engineering, 2009. WCSE'09. WRI World Congress on*, volume 1, Pp. 204–207. IEEE.

Yao, K., K. Paliwal, and T. Lee
2005. Generative factor analyzed hmm for automatic speech recognition. *Speech Communication*, 45(4):435–454.

Yao, X., T. Jitsuhiro, C. Miyajima, N. Kitaoka, and K. Takeda
2015. Modeling of physical characteristics of speech under stress. *IEEE Signal Processing Letters*, 22(10):1801–1805.

Yoon, T.-J.
2010. Speaker consistency in the realization of prosodic prominence in the boston university radio speech corpus. In *Speech Prosody 2010 Proceedings*.

Zahorian, S. and H. Hu
2008. A spectral/temporal method for robust fundamental frequency tracking. *The Journal of the Acoustical Society of America*.

Zwicker, E.
1982. Psychoakustik.

Zwicker, E. and H. Fastl
1999. *Psychoacoustics: Facts and models*, volume 2. Springer Berlin.

List of Figures

List of Tables